



UNIVERSITÀ  
DEGLI STUDI  
DI BRESCIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DOTTORATO DI RICERCA  
IN  
TECHNOLOGY FOR HEALTH

---

SSD: ING-INF/05

CICLO  
XXXVIII

From Explainable and Robust Design to  
Scalable Foundation Models: A Technical  
Journey towards Trustworthy Medical AI

Dottorando:  
Ing. Edoardo Coppola

Supervisore: Prof. Alberto Signoroni, Università degli Studi di Brescia  
Co-Supervisore: Dr. Mattia Savardi, Università degli Studi di Brescia

# Abstract

Trustworthiness has become a central objective for the research and medical communities as learning systems move from controlled development settings to real-world clinical workflows. Although trustworthiness spans methodological, operational, and societal dimensions, this thesis adopts a focused technical scope and targets properties that materially affect clinical utility and can be formalised, optimised, and tested. Concretely, it investigates: (i) clinically realistic multimodal learning for integrating heterogeneous evidence, (ii) human-interpretable explanations to expose model drivers at both cohort and patient level, (iii) robustness and supervision-efficient adaptation under deployment-level distribution shifts, (iv) label-free test-time reliability estimation to anticipate failure when ground truth is unavailable, and (v) scalable self-supervised representation learning, culminating in medical foundation backbones. The thesis develops these themes across four application domains. It first presents an explainable multimodal framework for early COVID-19 prognosis from chest X-rays and triage variables, showing that controlled fusion and multi-level explanations can align predictive performance with clinical transparency. It then studies robustness under shift in multimodal 3D segmentation of multiple sclerosis lesions, proposing adaptation strategies that minimise target-domain supervision and introducing a mathematically grounded reliability estimator that flags likely test-time failures without labels. Next, it demonstrates large-scale self-supervised ECG representation learning through HuBERT-ECG, establishing transferable cardiac representations across heterogeneous datasets and tasks. Finally, it extends this paradigm to arbitrary lead configurations by introducing a lightweight ( $\sim 7\text{M}$  parameters) any-lead ECG foundation model based on sparse, physiology-aligned spatiotemporal graphs, masked node modelling, and stochastic lead sampling to learn lead-invariant representations transferable to reduced-lead and wearable settings.

Taken together, this thesis contributes a set of methods that supports a pragmatic path toward trustworthy foundation models in medicine. Since the prerequisites for fully general multimodal foundations in medicine remain hard to satisfy today, progress is driven by strong and carefully validated backbones, robustness paired with reliability to anticipate deployment drift, and interpretable evidence integration that clinicians can inspect and challenge. Our results show that deliberate alignment between model inductive bias, clinical prior, transparency by design and deployment conditions is essential for medical AI systems that are not only accurate, but dependable in practice.

# Abstract

Il passaggio dei sistemi di apprendimento da ambienti di sviluppo controllati a contesti clinici reali ha reso l'affidabilità un obiettivo centrale per le comunità di ricerca e mediche. Sebbene questo tema abbracci dimensioni metodologiche, operative e sociali, questa tesi adotta un perimetro tecnico mirato e si concentra su proprietà che influenzano in modo concreto l'utilità clinica e che possono essere formalizzate, ottimizzate e verificate sperimentalmente. In particolare, la tesi affronta: (i) l'apprendimento multimodale clinicamente realistico per integrare evidenze eterogenee, (ii) spiegazioni interpretabili dall'uomo per chiarire i fattori che guidano il modello sia a livello di popolazione sia di singolo paziente, (iii) robustezza e adattamento efficiente con supervisione limitata in presenza di cambiamenti nelle distribuzioni dei dati durante l'uso, (iv) stima dell'affidabilità delle predizioni in assenza di etichette o riscontri diagnostici al momento del test, e (v) apprendimento scalabile e auto-supervisionato di rappresentazioni, fino alla costruzione di modelli fondazionali in ambito medico. Questi temi vengono sviluppati in quattro domini applicativi. Il primo contributo consiste in un framework multimodale interpretabile per la stima della prognosi nei pazienti COVID-19 a partire da radiografie del torace e dati di triage, mostrando che una fusione controllata delle modalità e spiegazioni su più livelli possono conciliare prestazioni predittive e trasparenza clinica. La tesi studia poi la robustezza di modelli multimodali per la segmentazione 3D di lesioni da sclerosi multipla quando si verificano cambiamenti nella distribuzione dei dati, proponendo strategie di adattamento che richiedono una supervisione minima e introducendo uno stimatore di affidabilità capace di segnalare probabili errori durante l'utilizzo. Successivamente, viene mostrato come HuBERT-ECG, un modello che sfrutta l'apprendimento auto-supervisionato su larga scala, apprenda rappresentazioni ECG trasferibili su dataset e compiti clinici eterogenei. Infine, questo paradigma viene esteso a configurazioni arbitrarie di derivazioni introducendo un modello fondazionale ECG leggero (~7M parametri), capace di elaborare combinazioni variabili di derivazioni tramite grafi spaziotemporali sparsi e fisiologicamente coerenti, insieme a masked node modelling e campionamento stocastico delle derivazioni per ottenere rappresentazioni invarianti rispetto al numero e al tipo di derivazioni, trasferibili anche a contesti a poche derivazioni e a dispositivi indossabili. Nel complesso, la tesi propone un insieme di metodi che sostiene un percorso pragmatico verso modelli fondazionali affidabili in medicina. Poiché oggi i requisiti per modelli multimodali pienamente generali in ambito clinico sono ancora difficili da soddisfare, il progresso passa attraverso modelli di base solidi e accuratamente validati, robustezza accompagnata da

indicatori di affidabilità capaci di anticipare possibili derive durante l'impiego, e integrazione delle evidenze in forme che i clinici possano ispezionare e mettere criticamente alla prova. I risultati indicano che un allineamento consapevole tra bias induttivo del modello, conoscenza clinica, trasparenza progettuale e condizioni d'uso è essenziale per sistemi di intelligenza artificiale medica non solo accurati, ma realmente affidabili nella pratica.

# Contents

|                                                                                                                                                         |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Chapter 1 - Introduction .....                                                                                                                          | 1  |
| Chapter 2 - Explainable AI for COVID-19 prognosis from early chest x-ray and clinical data in the context of the COVID-CXR international hackathon..... | 5  |
| Abstract .....                                                                                                                                          | 5  |
| 2.1 Introduction .....                                                                                                                                  | 5  |
| 2.2 Related Works .....                                                                                                                                 | 6  |
| 2.3 Methods.....                                                                                                                                        | 7  |
| 2.3.1 Problem formulation.....                                                                                                                          | 7  |
| 2.3.2 Implementation .....                                                                                                                              | 8  |
| 2.4 Experimental Results.....                                                                                                                           | 10 |
| 2.4.1 The challenge dataset.....                                                                                                                        | 10 |
| 2.4.2 Model evaluation .....                                                                                                                            | 10 |
| 2.4.3 Explaining model predictions .....                                                                                                                | 11 |
| 2.5 Conclusions .....                                                                                                                                   | 13 |
| Chapter 3 - Towards Robust and Reliable Multimodal 3D Segmentation of Multiple Sclerosis Lesions .....                                                  | 14 |
| Abstract .....                                                                                                                                          | 14 |
| 3.1 Introduction .....                                                                                                                                  | 14 |
| 3.2 Related Work.....                                                                                                                                   | 15 |
| 3.3 Methods.....                                                                                                                                        | 16 |
| 3.3.1 Pre-training a segmentation model .....                                                                                                           | 16 |
| 3.3.2 Minimal-instance Adaptation .....                                                                                                                 | 17 |
| 3.3.3 Cross-domain Few-shot Adaptation .....                                                                                                            | 17 |
| 3.3.4 Latent Distance Analysis .....                                                                                                                    | 18 |
| 3.4 Experimental Results.....                                                                                                                           | 19 |
| 3.4.1 Datasets.....                                                                                                                                     | 20 |
| 3.4.2 Building a state-of-the-art MS segmentation baseline.....                                                                                         | 20 |
| 3.4.3 Minimal-instance Adaptation Yields Consistent Gains Across Datasets .....                                                                         | 21 |
| 3.4.4 Few-shot Cross-dataset Adaptation Significantly Enhances Multi-domain Generalisation .....                                                        | 22 |
| 3.4.5 LDA-quantified OOD Score Correlates with Errors .....                                                                                             | 24 |
| 3.4.6 Limitations .....                                                                                                                                 | 25 |
| 3.5 Conclusions and Future Works .....                                                                                                                  | 25 |

|                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------|----|
| Chapter 4 - HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications ..... | 26 |
| Abstract .....                                                                                                 | 26 |
| 4.1 Introduction .....                                                                                         | 26 |
| 4.2 Methods .....                                                                                              | 29 |
| 4.2.1 HuBERT-ECG Architecture and Theoretical Framework .....                                                  | 29 |
| 4.2.2 Implementation .....                                                                                     | 31 |
| 4.2.3 Unsupervised pseudo-label discovery .....                                                                | 32 |
| 4.2.4 Self-supervised pre-training.....                                                                        | 33 |
| 4.2.5 Supervised fine-tuning.....                                                                              | 33 |
| 4.3 Experimental Details .....                                                                                 | 34 |
| 4.3.1 Data and Pre-processing .....                                                                            | 34 |
| 4.3.2 Downstream evaluation via supervised fine-tuning .....                                                   | 35 |
| 4.3.3 Evaluation Metrics.....                                                                                  | 36 |
| 4.4 Results .....                                                                                              | 37 |
| 4.4.1 HuBERT-ECG is efficient on small-sampled datasets .....                                                  | 37 |
| 4.4.2 HuBERT-ECG remains efficient in increasingly difficult contexts .....                                    | 37 |
| 4.4.3 HuBERT-ECG enables accurate diagnostics in large-scale, complicated scenarios.....                       | 38 |
| 4.4.4 HuBERT-ECG can be a good predictor of future cardiovascular events .....                                 | 40 |
| 4.4.5 HuBERT-ECG navigates extremely complex scenarios .....                                                   | 40 |
| 4.4.6 Benchmarking HuBERT-ECG against reported human performance.....                                          | 40 |
| 4.4.7 Fine-tuning HuBERT-ECG requires less compute time than training from scratch .....                       | 42 |
| Chapter 5 - Toward a 7M-Parameter Any-Lead ECG Foundation Model: Methodology and Preliminary Results .....     | 47 |
| Abstract .....                                                                                                 | 47 |
| 5.1 Introduction .....                                                                                         | 47 |
| 5.2 Methods .....                                                                                              | 48 |
| 5.2.1 Learning a fine-grained codebook for ECG segments .....                                                  | 48 |
| 5.2.2 Any-lead ECG Representation Learning via Masked Node Modelling .....                                     | 49 |
| 5.2.3 The spatiotemporal topology .....                                                                        | 50 |
| 5.2.4 Transfer-guided Codebook Refinement .....                                                                | 52 |
| 5.2.5 Architecture .....                                                                                       | 52 |
| 5.3 Experimental details .....                                                                                 | 53 |
| 5.3.1 Data and preprocessing .....                                                                             | 53 |
| 5.3.2 Unsupervised codebook discovery .....                                                                    | 54 |
| 5.3.3 Self-supervised Pre-training .....                                                                       | 55 |

|                                                                       |    |
|-----------------------------------------------------------------------|----|
| 5.3.4 Supervised fine-tuning and downstream evaluation.....           | 55 |
| 5.4 Preliminary Results .....                                         | 56 |
| 5.4.1 Ablation on graph encoders and ECG topologies.....              | 56 |
| 5.4.2 Ablation on temporal encodings and lead embeddings .....        | 58 |
| 5.4.3 Ablation on the masking ratio.....                              | 58 |
| 5.4.4 Investigation on codebook size and multi-task pre-training..... | 59 |
| 5.4.5 Analysing cluster quality across encoding layers .....          | 60 |
| 5.5 Discussion and future steps .....                                 | 61 |
| Chapter 6 - Conclusions and Future Works .....                        | 63 |
| References.....                                                       | 65 |
| Supplementary Material.....                                           | 75 |
| Appendix A .....                                                      | 75 |
| Explainability maps creation algorithm.....                           | 75 |
| Appendix B.....                                                       | 76 |
| Appendix C.....                                                       | 80 |
| 1 Pre-training analysis .....                                         | 84 |
| 2 Ablation study.....                                                 | 90 |
| 3 Explaining fine-tuned diagnostic models .....                       | 92 |

# Chapter 1 - Introduction

Trustworthiness in medical AI has become a central objective for the research and medical communities as learning systems move from controlled development settings into clinical workflows and decision-support pathways [1]. The term “trustworthiness”, however, covers a vast landscape of challenges spanning methodological, operational, and societal dimensions. Recent consensus documents [1] and surveys [1], [2], [3], [4], [5] emphasise that trust in medical AI cannot be reduced to nor built on benchmark accuracy alone, but requires explicit attention to multiple system properties across the full lifecycle—from data collection and model development to deployment, monitoring, and human–AI interaction. At the same time, those same studies underline that no single work or project can address the whole spectrum of such properties and that progress comes from well-defined scopes. Accordingly, this thesis adopts a focused technical interpretation of trustworthiness, without seeking to exhaust its full breadth, concentrating on a subset of properties that materially affect clinical utility and that admit precise formalisation, optimisation, and testing. Concretely, the works collected here investigate (i) multimodal learning as a clinically realistic way to integrate heterogeneous evidence, (ii) human-interpretable explanations as a means to shed light on model’s internal functioning and predictions, (iii) robustness and efficient adaptation under deployment-level distribution shifts, (iv) reliability mechanisms that anticipate test-time failure when labels are unavailable, and (v) generalisable representation learning, culminating in the transition to self-supervised foundation models for broad clinical transfer.

We next motivate these five pillars in turn. Although discussed separately for clarity, they form a coherent technical arc: multimodal evidence raises new demands for modality-grounded explanations; explanations are meaningful only if models remain robust under data shifts; models can hardly guarantee robustness uniformly at deployment time, necessitating reliable indicators of uncertainty and out-of-distributions behaviour. In parallel, scalable, self-supervised foundational representations provide a complementary route to generalisation, especially when task-specific supervision is limited.

We begin with multimodality because evidence is rarely unimodal in real workflows, including medical decision-making, which is intrinsically multimodal. Clinicians consider images, physiological signals, laboratory values, structured variables, and narrative context to reach a judgement. Multimodal models aim to reflect this evidential structure and can improve performance by letting complementary sources constrain one another, especially in early-stage or resource-limited care pathways [6]. Yet multimodality also sharpens trustworthiness challenges. When data is heterogeneous, missing, or noisy, a single modality may silently dominate; cross-modal shortcuts can arise from dataset artefacts; and fusion mechanisms may become opaque as model complexity grows [7]. Therefore, multimodality must be treated not only as a route to higher accuracy, but as a setting where evidence integration, missing-data handling, and explainability become central engineering concerns and design choices.

Once multiple sources of evidence are fused, understanding how the model uses them becomes a trust-critical requirement. Explainability, in fact, is essential precisely because clinical AI systems operate in safety-critical environments and are used to support expert judgment. Clinicians must be able to interrogate outputs, detect failures, and relate predictions to familiar diagnostic cues [8], [9], [10], [11], [12], [13], [14]. However, the field has also shown that standard post-hoc explanations can be unstable, weakly connected to causal evidence, or difficult to interpret in practice, especially under distribution shifts [15]. This motivates a stronger stance: explainability should be a design objective rather than a cosmetic layer, operating at multiple levels: global feature contributions that reflect

cohort-level patterns, patient-specific rationales that support individualised reasoning, and modality-grounded visual or signal-based evidence that can be inspected for clinical plausibility [16].

That said, to make explanations dependable in practice, models must remain stable when confronted with the variability of real clinical environments [15]. This brings robustness and adaptation to the foreground. Robustness is the ability of a model to maintain acceptable performance under distribution shifts without requiring access to target-domain labels or retraining. Adaptation, by contrast, refers to procedures that update or recalibrate a model using information from the deployment domain in order to recover performance when robustness alone is insufficient [17]. In practice, these concepts are complementary: robustness aims to prevent degradation under shift, while adaptation aims to repair or attenuate it once shifts are inevitably encountered. A common adaptation strategy consists in using labelled data from the target distribution, but the practical cost of annotation can be prohibitive in many domains, motivating research into efficient adaptation protocols under extremely limited supervision [18]. From a trustworthiness standpoint, the aim is not only to build models that generalise in idealised cross-validation, but to develop systems that remain stable under realistic clinical drift and can be repaired or tuned with minimal additional labelling burden.

Even with robustness-oriented design and careful adaptation, clinical variability can still push models into unfamiliar regimes. This motivates mechanisms to flag when predictions are likely to be unreliable at test time, which closely relate to the ability of a model to recognise when it may be wrong. In fact, in safety-critical settings, it is insufficient for a system to be accurate on average; it must also signal when its predictions should be treated cautiously or deferred to human expertise. The medical AI community has highlighted uncertainty calibration, out-of-distribution detection, and reliability estimation as core mechanisms for safe use [19]. Reliability is thus framed here as an operational property, tied to deployment: trustworthy systems should not only produce predictions, but also expose credible indicators of confidence and shift, especially when ground truth is unavailable at test and deployment time [20].

Finally, these challenges unfold against a broader scalability constraint: the need of general-purpose representations that transfer widely across datasets and tasks, especially when task-specific supervision is scarce. These considerations motivate the move toward generalisable representation learning and foundation models. Much of medical AI has been built in a task-specialist paradigm: models trained with supervision for a single condition, a single modality, or a narrowly defined clinical endpoint. This approach has produced valuable tools, but it exposes a structural limitation: specialist models often do not adapt gracefully to new domains, new patient populations, or even modest shifts within the same task, because their representations are tuned to the idiosyncrasies of the training distribution rather than to reusable physiological or structural patterns [21] [22]. By contrast, foundation models aim to learn broad, transferable representations through large-scale self-supervised pretraining on diverse, mostly unlabelled data. In doing so, they shift the burden from expensive per-task annotation to scalable pretraining, and enable data and compute-efficient fine-tuning whenever a specific clinical objective, domain, or distribution must be addressed [23] [24] [23]. Outside medicine, unimodal foundation models have already transformed natural language processing [25], [26], [27], [28], computer vision [29], [30], and audio processing [31], [32] by showing that self-supervised pre-training can yield representations that generalise across tasks with minimal supervision [21]. Methodological advances and unprecedented data availability have even enabled the rise of large-scale multi-modal learning (e.g. Multi-modal Large Language Models such as the Gemini Series or today’s GPTs), which integrates heterogeneous streams (e.g., text–image–audio) within a shared representational space and has consequently expanded our contemporary conception of what a “foundation model” can be: not only a transferable backbone, but increasingly a single model capable of addressing many qualitatively different tasks via prompting or lightweight adaptation. In medicine, progress toward this modern, multi-modal ideal has been slower, largely

because the prerequisites that made such models possible elsewhere are harder to satisfy in clinical settings. Medical data are distributed across institutions, subject to privacy and regulatory constraints, collected with changing protocols, and frequently accompanied by labels that are sparse, noisy, or tied to local clinical practice rather than reusable ontologies [21], [33], [34]. As a result, many medical AI systems remain task specialists, trained under narrowly defined supervision and evaluated on limited distributions, with representations that do not necessarily transfer when tasks, cohorts, or acquisition conditions change [20]. Recent efforts have begun to shift this landscape by introducing unimodal and multimodal medical foundation models in radiology, pathology, segmentation, and biomedical reasoning [35], [36], [37], [38], [39], [40], [41], [42], but substantial gaps remain across several domains and modalities. In such under-served areas, aiming directly for fully general multi-modal foundation models is often technically unrealistic. The data requirements are steep: large-scale corpora, cross-institution diversity, consistent multi-modal alignment, and task-level supervision for evaluation and instruction all need to coexist. When these conditions are absent, a pragmatic and scientifically grounded path is to first establish general-purpose self-supervised backbones within a given modality or domain. This aligns with the original foundation-model paradigm: representations learned from large, diverse unlabelled data that can be efficiently fine-tuned to many downstream tasks, datasets, and distributions with minimal additional supervision [24]. The objective is not to lower ambition, but to build the representational substrate that makes later multi-modal, prompt-capable medical models feasible and reliable. This backbone-first conception also reframes what “foundational” should mean in medicine today: a model is foundational to the extent that it supports broad applicability and efficient transfer, not merely because it is large or multimodal. Accordingly, the central technical requirement becomes comprehensive evaluation across heterogeneous tasks and sources—especially in low-resource settings, for rare conditions, and under distribution shifts. If a self-supervised backbone consistently transfers across diverse clinical scenarios with lightweight adaptation, it constitutes a genuine foundation for that domain, and a necessary step toward richer, multi-modal medical foundation models in the future.

Against this backdrop, the thesis develops a coherent technical progression from multi-modal, explainable decision support, to robustness and reliability under shift, and final to scalable self-supervised foundational representations.

Chapter 2 presents an explainable multi-modal framework for prognosis estimation in COVID-19 patients using early chest X-rays and triage clinical data [43]. The chapter emphasizes clinically realistic evidence integration and transparency: modalities are fused in a controlled and interpretable way, and the resulting predictions are accompanied by explanations at both model level (to describe global decision drivers) and patient level (to support case-by-case human reasoning).

Chapter 3 focuses on robustness, efficient adaptation, and reliability in multimodal 3D segmentation of multiple sclerosis lesions [44]. It proposes two complementary adaptation strategies aimed at quantifying and minimising the supervision required to adapt a model to unseen datasets—first adapting to each target source independently, then extending adaptation to multiple unseen sources jointly while explicitly mitigating forgetting on the original training domain. Building on this, the chapter introduces a label-free test-time reliability estimator, named Latent Distance Analysis, that operates in latent space and is mathematically grounded. This mechanism enables models to anticipate and signal performance degradation under distribution shift, and is compatible with any form of test-time adaptation.

Chapter 4 presents HuBERT-ECG as a self-supervised foundation backbone for scalable cardiac applications, contributing to a clinical domain where robust unimodal foundations are still absent [45]. The chapter develops large-scale self-supervised pretraining to learn transferable representations from heterogeneous ECG data, and demonstrates broad applicability through lightweight fine-tuning across a wide range of tasks, datasets, and clinical objectives. In line with the thesis scope, emphasis is placed on validating representation generalisability through systematic cross-scenario evaluation rather than assuming it from scale alone.

Chapter 5 introduces a lightweight any-lead ECG foundation model based on sparse spatiotemporal ECG graphs and large-scale pre-training via masked node modelling and stochastic lead sampling. Although in progress, this work extends foundation ECG models towards scenarios that are far from the clinical standards (e.g. wearables) and broadens applicability of these solutions maintaining efficiency.

Chapter 6 concludes by synthesising methodological insights across these works, re-situating them within the broader landscape of trustworthy medical AI, and highlighting open technical challenges and future directions suggested by the results of the thesis.

# Chapter 2 - Explainable AI for COVID-19 prognosis from early chest x-ray and clinical data in the context of the COVID-CXR international hackathon

---

This chapter is based on the article authored by Coppola et al., published after peer review in the proceedings of SPIE Medical Imaging 2023 [43]. Within the narrative arc on the technical dimensions of trustworthiness introduced in Chapter 1, this chapter represents the starting point of the thesis and focuses on interpretability by design in multimodal medical AI. Here, trustworthiness is primarily framed as the ability of a model to expose intelligible decision pathways to clinicians: from transparent and controlled modality fusion at the architectural level, to modality-grounded explanations at both model and patient level. This chapter therefore establishes interpretability as a foundational trust mechanism in settings characterised by limited data, heterogeneous and low-quality inputs, and safety-critical clinical decisions.

---

## Abstract

Most AI studies on chest X-rays (CXRs) during the COVID-19 pandemic focused on diagnosis, while prognosis and clinical interpretability remained underexplored. In this work, developed within the COVID-CXR international hackathon, we present a multimodal machine learning framework that predicts patient prognosis by combining CXR analysis with clinical data collected at triage. CXRs undergo automatic pre-processing and are enriched with a severity index (Brixia-score) derived from a validated deep learning model, while missing clinical values are handled through clinically informed imputation and feature selection. Among several classifiers tested, a Random Forest achieves 75% sensitivity, 73.9% specificity, and 74.4% balanced accuracy, ranking second overall in predictive performance. Our explainability framework—based on model-level feature importance and patient-specific SHAP values and lung explainability maps—is rated first by health professionals for clinical interpretability. Our multimodal system demonstrates that integrating imaging and clinical data with transparent explanations provides accurate and trustworthy support for physicians in managing COVID-19 patients, helping improve allocation of healthcare resources more effectively.

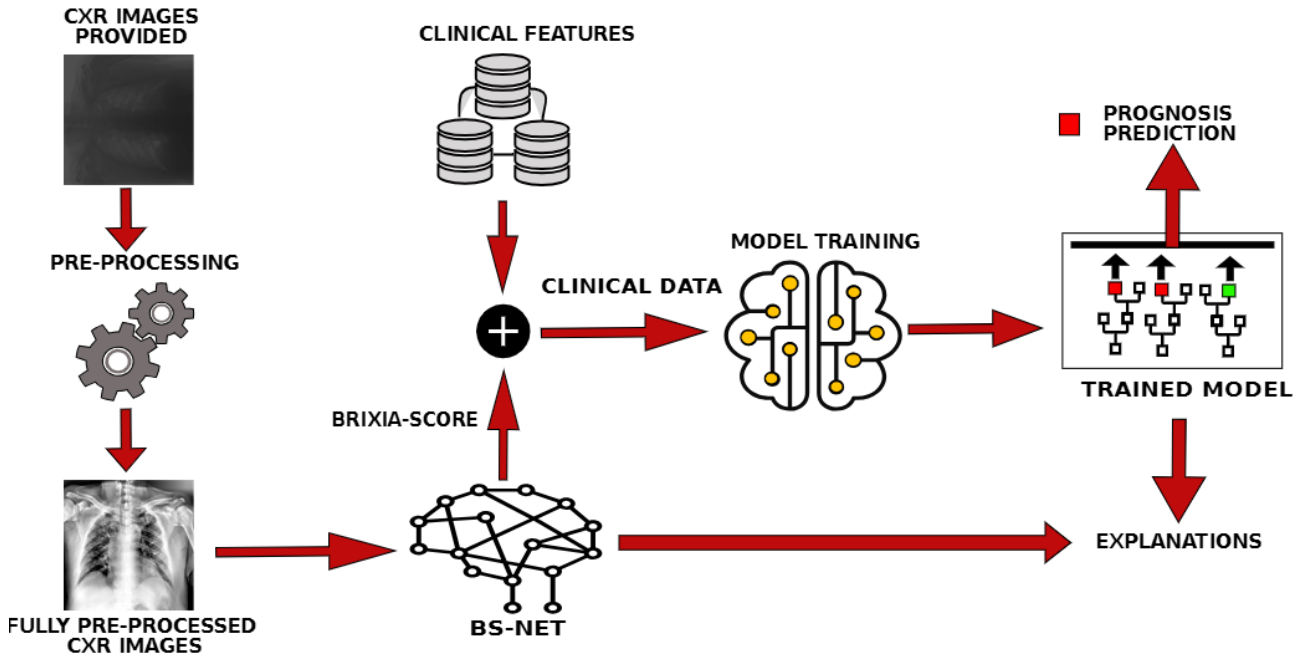
## 2.1 Introduction

The COVID-19 pandemic placed unprecedented strain on healthcare systems worldwide, exposing the urgent need for tools that could support clinical decision-making and resource allocation. Early research in medical imaging largely concentrated on chest X-ray (CXR)-based diagnosis of COVID-19, despite well-recognized limitations of this approach [46]. In contrast, tasks such as severity assessment [47], deterioration prediction [48], and patient prognosis [49], [50], [51] received far less attention. Moreover, few studies addressed the critical issue of interpretability, leaving clinicians with limited insight into how machine learning (ML) models reached their conclusions.

This study originates from our participation in the COVID-CXR international hackathon, where researchers were challenged to design ML solutions that combined imaging and clinical data to predict patient prognosis. Participants were required not only to achieve accurate predictions but also to provide explanations that could be evaluated by health professionals, radiologists, and computer scientists.

Building on our participation to this challenge, we present a multimodal ML framework that integrates CXR analysis with clinical features collected at triage. CXRs undergo automated pre-processing and are enriched with a severity index (Brixia-score [52]) derived from a validated deep learning model. Clinical features are curated through informed imputation and feature selection to ensure robustness. We then train and compare multiple classifiers, ultimately selecting a Random Forest as the most effective. To enhance trust and usability, we embed a suite of interpretability tools, including model-specific feature importance, patient-level SHAP values, and patient-specific lung explainability maps.

The remainder of this chapter is organized as follows: Section 2.2 reviews related work; Section 2.3 describes our methodology and implementation; Section 2.4 presents the experimental results; ultimately, Section 2.5 concludes the chapter by summarizing our findings, their clinical relevance and the importance of explainability in AI-driven prognosis systems. Figure 2.1 depicts the overall pipeline followed in this study.



**Figure 2.1:** Overall pipeline of the study. Low-quality CXR images are first pre-processed and enhanced. They are then fed into BSNet for the computation of the Brixia-score, later added to the clinical features collected at triage. This new set of patient-specific features are used to train multiple models, the best of which is eventually used for testing and prognosis estimation. Model and patient-specific explanations are produced by both BSNet and the final prognosis estimator to justify the predictions.

## 2.2 Related Works

Since the onset of the COVID-19 pandemic, most AI-based studies on chest X-ray (CXR) interpretation have concentrated on diagnosis [53]. Among the first, Oh et al. [54] introduced convolutional neural networks (CNNs) to analyse CXRs of patients suspected of pneumonia for diagnostic purposes. However, many subsequent studies suffered from limited datasets, which often resulted in poor generalizability of diagnostic models [55].

In contrast to this dominant diagnostic focus, significantly fewer studies explored CXR interpretation beyond diagnosis. Notably, Signoroni et al. proposed BSNet [47], a multi-module deep neural network trained on the large BrixIA dataset (~4,700 CXRs), which estimates pneumonia severity

through a multi-regional scoring system. BSNet not only outperformed experienced radiologists in severity assessment but also provided super-pixel-based explainability maps that highlighted relevant lung regions. Shamout et al. [48] investigated short-term deterioration prediction in the Emergency Department, integrating CXRs with clinical variables. Their hybrid system, combining CNN-based imaging analysis with machine learning on clinical data, achieved performance comparable to radiologists.

Other studies focused directly on prognosis prediction. Soda et al. analysed more than 800 patients by fusing CXRs and clinical data, and supported their outcomes with activation maps [49]. Cheng et al. combined a dataset comparable in size to BrixIA with a relatively small set of clinical parameters, reporting feature importance for mortality prediction but limited interpretability [51]. Jiao et al. included over 2,000 patients in their study, yet offered no explainability mechanisms alongside their prognostic model [51]. Collectively, these works demonstrate that multimodal approaches—integrating imaging and clinical data—consistently improve prognostic accuracy, but they also reveal a gap in providing interpretable outputs that clinicians can trust.

In this context, we present a multimodal machine learning framework that leverages CXR images and nearly 40 clinical variables from more than 1,500 COVID-19 patient records collected across multiple hospitals. Our approach not only predicts patient prognosis but also delivers interpretable justifications through feature importance analysis, patient-level SHAP values [56], and subject-specific lung-region explainability maps.

## 2.3 Methods

To design a multimodal learning framework that integrates CXR images and clinical features collected at triage to estimate prognosis for COVID-19 patients, we start by mathematically formulating the problem and proceed by outlining our implementation.

### 2.3.1 Problem formulation

Let each patient record be represented as a pair  $(I, F)$ , where  $I \in \mathbb{R}^{H \times W}$  denotes the CXR image and  $F \in \mathbb{R}^d$  is a  $d$ -dimensional vectorial representation of the clinical features. More specifically,  $F$  is a concatenation of heterogeneous attributes of varying nature encoded numerically:  $F = [f_1^c, \dots, f_{d_c}^c \parallel f_1^g, \dots, f_{d_g}^g \parallel f_1^r, \dots, f_{d_r}^r]$ , where  $f_i^c \in \mathbb{R}$  are continuous features (e.g. laboratory values such as lactate dehydrogenase),  $f_i^c \in \mathbb{Z}$  are discrete or integer-valued features (e.g. days of fever), and  $f_i^c \in \{0, 1, \dots, K\}$  are categorical features (e.g. hospital ID, presence of comorbidities). Thus  $d = d_c + d_g + d_r$ . Such multimodal dataset formulation is however ideal and does not take into account real-world challenges such as image corruption and low quality, and frequent missing values. Conceptually, a collected image can be seen as  $\tilde{I} = I + \mathcal{E}(I)$ , with  $\mathcal{E}(I)$  being a corruption function, and missing values can be modelled through the use of a binary missingness mask  $M \in \{1, ?\}^d$ , where  $M_i = 1$  if  $f_i$  is observed, and  $M_i = ?$  if feature  $f_i$  is missing. Here  $?$  denotes a NotANumber (NaN) token, which differs from any numeric value but propagates under arithmetic (e.g.  $? \cdot x = ? \forall x \in \mathbb{R}$ ). This avoid conflating missingness with valid numerical values, such as 0, and allows us to preserve a simple formulation. The observed clinical vector can therefore be written as  $F_{obs} = M \odot F$ , where  $\odot$  denotes the element-wise product. Under the Missing At Random assumption, each mask entry  $M_i$  is modelled as a conditional Bernoulli random variable:  $M_i \mid Z \sim \text{Bernoulli}(p_i(Z))$ , where  $Z$  denotes observed covariates (e.g. hospital site, basic vitals), and  $p_i(Z) = \sigma(\alpha_i + \beta_i^T Z)$ , with  $\sigma(\cdot)$  being the logistic function. This formulation captures systematic patterns of missingness

(e.g. certain laboratories tests omitted on hospital workflow or patient condition or, realistically, due to resource saturation) while avoiding the unrealistic assumption of Missing Completely At Random.

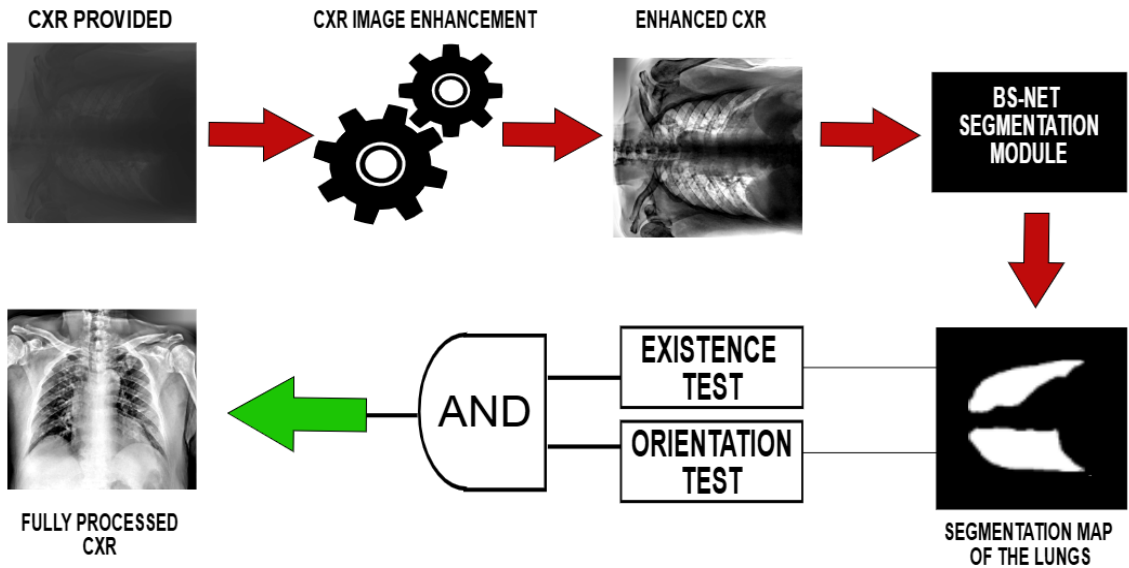
Let  $y$  be the categorical target variable representing all possible prognostic outcomes, then the next step is to learn a prognosis estimator  $\phi : (I, F) \rightarrow \hat{y}$ ,  $\hat{y} \in Im(y)$  that approximates the mapping between CXRs and clinical features, and the prognosis. However, most off-the-shelf estimator are not robust to corrupted inputs, meaning that missing values in  $F$  and degraded CXRs may prevent direct application of  $\phi$ . This issue requires the definition of two recovery operators,  $R_I$  and  $R_F$ , that map the corrupted CXR  $\tilde{I}$  to a recovered  $I_{rec}$  that resembles  $I$  and the partially observed features to a complete representation  $F_{CPLT}$ , respectively. The recovered patient record can therefore be defined as  $x = (I_{rec}, F_{CPLT})$  and effectively used to learn the prognosis estimator  $\phi$ .

### 2.3.2 Implementation

Our framework relies on two recovery operators that prepare heterogeneous inputs for the prognosis estimator  $\phi$ :

- CXR recovery operator  $R_I$  — responsible for image pre-processing, standardization and enhancement;
- Clinical recovery operator  $R_F$  — responsible for clinically informed imputation of missing values.

**Image recovery and severity encoding.** Learning or inferring from low-quality CXRs poses a significant challenge as the risk to rely on spurious artefacts rather genuine pathology-related signals is non-negligible. In these settings, carefully designing the recovery operator  $R_I$  becomes crucial to ensure clinical viability and trustworthiness under extreme real-world settings. For these reasons, the CXR recovery operator, a conceptual representation of which is shown in Figure 2.2, is designed as a composition of multiple operations: resizing to a common shape of  $512 \times 512$ , normalization to the  $[0, 1]$  range, contrast-limited adaptive histogram equalization (CLAHE, clip=0.01), median filtering (kernel size = 3), and percentile clipping between the 2<sup>nd</sup> and 98<sup>th</sup> percentiles. Residual corruptions such as misorientation and inversion are corrected using the lung segmentation module



**Figure 2.2:** CXR image recovery pipeline. Corrupted and low-quality CXRs are first enhanced through several technique, then two tests on the existence and visibility of the lungs and the correct orientation of the image are to be met before output the fully recovered CXR. If the tests are not passed, a feedback message is returned to the user.

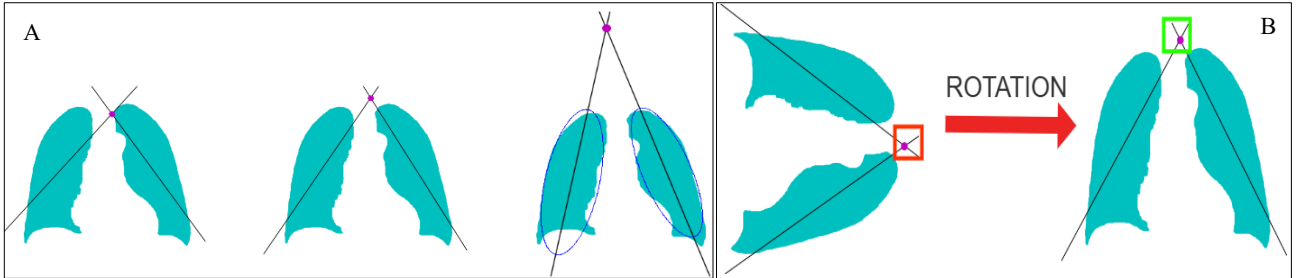
of BSNet. Two tests are designed: (i) a lung existence test, verifying the presence of two sufficiently large connected components in the segmentation map, and (ii) an orientation test, fitting geometric primitives (e.g. least-squares lines, RANSAC, or ellipse fitting) to lung boundaries and checking the intersection point. If the intersection lays outside the upper image region, the CXR is rotated until it matches that standard orientation for medical interpretation (see Figure 2.3). Images failing either test are discarded with a feedback message, enabling manual recovery on a restricted set of CXRs.

Once recovered, CXRs are processed by BSNet to compute a six-region severity score, namely Brixia-score. Each element in this sequence ranges in  $[0, 3]$ , where 0 means “no lung abnormalities” and 3 means “interstitial and dominant alveolar infiltrates”. Together with these 6 regional scores, BSNet also outputs a confidence value. We sum the six regional scores and average the corresponding confidence values to obtain two additional clinical features to enrich the representation  $F_{CPLT}$ .

**Clinical recovery and feature selection.** The clinical feature recovery operator processes the clinical features after the addition of the aggregated Brixia-score and its associated confidence. This operator leverages a clinically informed imputation of the pervasive missing values. Specifically, our strategy consists of:

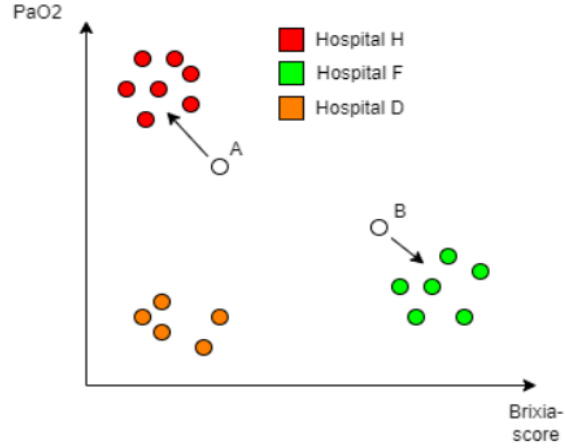
- imputing categorical variables with intermediate numeric codes for binary attributes (e.g. 0.5) representing disease presence/absence and with the most probable category inferred from correlated features;
- imputing numerical features using k-nearest neighbour regression, which was found effective in recovering missing genes [57] (see Figure 2.4).

After imputation, with explainability in mind, we reduce dimensionality by applying a model-based feature selection with varying predictive importance thresholds.



**Figure 2.3:** (A) Detection of the intersection point of the lung lines fitted by least-squares method (left), RANSAC (centre), and ellipse fitting (right). (B) Once the intersection point is found, its position within the image borders is used to estimate the rotation to apply to fully recover the correct CXR orientation.

**Prognosis estimation and explanation generation.** After feature selection, we train multiple classifiers including decision trees, Support Vector Machines (SVMs), gradient boosting (XG Boost), extra trees, and Random Forest. We evaluate performance in terms of sensitivity, specificity, and balanced accuracy. Eventually, to produce explanations for model predictions, we consider both model and patient-specific explanations that can justify model outputs from both a model and subject perspective. Specifically, we opt for a (i) model-specific feature importance, which measures the overall impact that selected features have on prediction changes, (ii) patient-specific SHAP values [56], which measure how much specific subject characteristics contribute to a certain outcome, (iii) lung explainability maps, as presented in the paper by Signoroni et al. [47], which depict how multiple regions in the lungs contribute to the computed six-region Brixia-score [52].



**Figure 2.4:** Pictorial representation of clinically informed value imputation of missing features. Unknown attributes are imputed via KNN regression.

## 2.4 Experimental Results

### 2.4.1 The challenge dataset

In the context of the COVID-CXR hackathon, the provided dataset consists of 1103 training records and a blind test set of 486 instances. Each record is characterized by 38 clinical features, including categorical, continuous and discrete attributes. A description of these features is provided by Soda et al. [49]. Given the real-world and extreme settings in which this data was collected, missingness represents a hard challenge, as some attributes are unknown in percentages that range from 20% to 80%. Other features are even completely absent (e.g. presence of many cardiovascular diseases). The target variable takes two possible values: mild and severe.

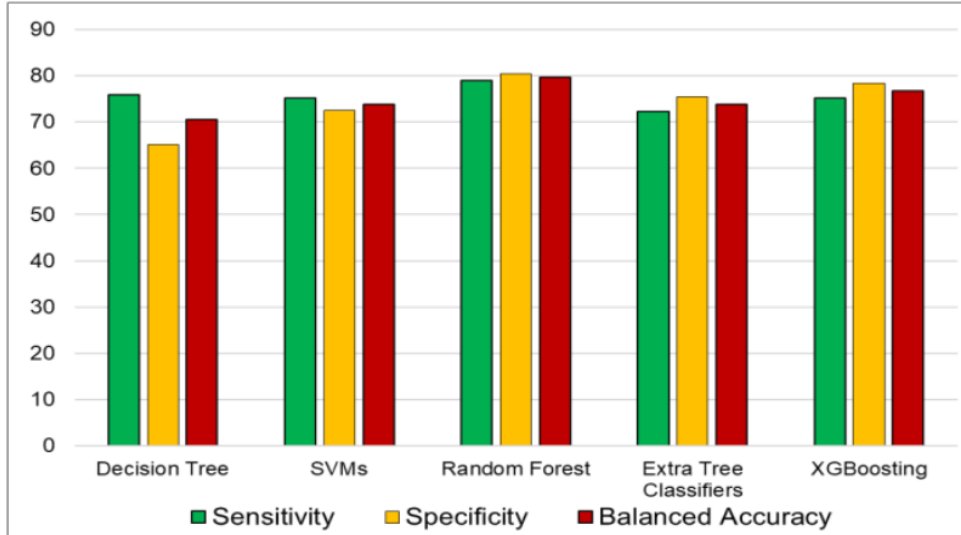
Concerning the CXR images in the dataset, their presentation is highly heterogeneous as they differ in aspect ratio, shape, orientation and colour inversion. In addition, these CXR images are not provided in the traditional DICOM format but in JPEG, indicating that they have been likely taken by a smartphone or a camera. This acquisition results in blurriness, darkness and low sharpness. An example of provided CXR is given in Figure 2.1 and in Figure 2.2.

### 2.4.2 Model evaluation

After enriching clinical features with the aggregated Brixia-score and the corresponding confidence, we train and evaluate multiple prognosis estimator on a set of 19 features selected for their predictiveness. These features include Sex, DaysFever, Hospital, PaCO<sub>2</sub>, Temp\_C, pH, Brixia-score confidence, DifficultyInBreathing, RBC, D\_dimer, Glucose, PCT, Age, CRP, SaO<sub>2</sub>, WBC, PaO<sub>2</sub>, Brixia-score, LDH (see Supplementary Material – Appendix A – Table 1 for the meaning of the selected features). Noteworthy, the presence of the Brixia-score and the associated confidence among the most predictive features underscore the high quality and viability of our recovery operator  $R_I$ , since feeding BSNet with images that deviated too far from the DICOM standard used during its training would have resulted in unreliable and unproductive scores. The available records are then split into a training set containing 75% of the instances and a validation set with the remaining 25%. We first use the internal validation set to tune the models' hyper-parameters. Next, we conduct a 5-

fold cross-validation of the tuned models to identify the best model candidate for evaluation on the blind test set. We report the results of this model search in Figure 2.5.

After selecting the Random Forest as best prognosis estimator, we evaluate its performance on the blind test set and report a sensitivity of 75%, a specificity of 73.9% and a balanced accuracy of 74.4%. Thanks to this scoring, our solution ranked second on the challenge leaderboard, aligning with previous works [49], [50], [51].

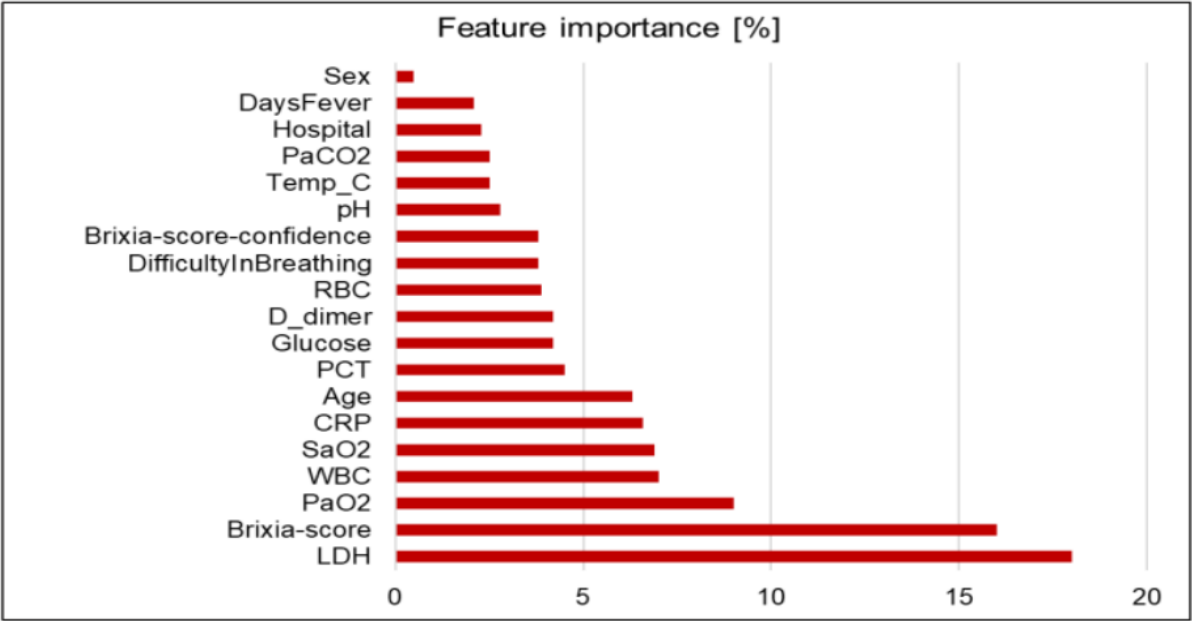


**Figure 2.5:** Performance of multiple classifier on the internal validation set. The Random Forest classifier show the highest performance according to sensitivity, specificity and balanced accuracy and is thus selected for the final evaluation on the challenge blind test set.

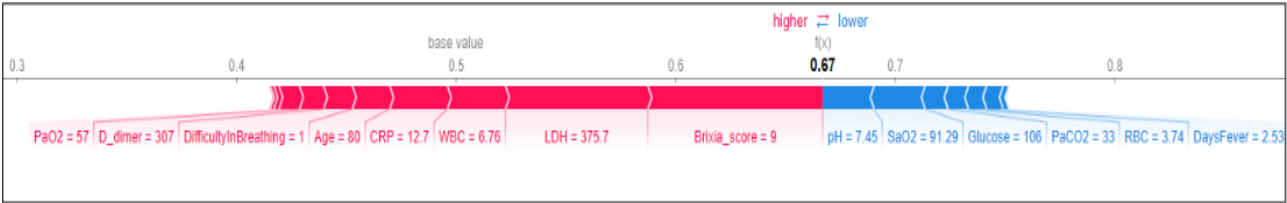
### 2.4.3 Explaining model predictions

The first type of explanation produced by our multimodal framework is the feedback message returned to the user when the image recovery fails. This feedback in fact allows the user to manually validate and, if necessary, re-acquire the input image to satisfy the tests presented in the section 2.3.2. The second type of explanation is a model-level feature importance that measures how much a trained classifier relies on the input features to formulate the final prediction (see Figure 2.6). These importance scores can provide physicians with insights into the model internal functioning. The third factor of interpretability of our framework comes from SHAP values [56] and is particularly useful to measure the impact of specific patient’s features in the computation of their prognosis. These estimates can indicate which features pushed the model towards a mild prognosis or a severe one and the force of such push (see Figure 2.7). Eventually, the last explanation comes from the lung explainability maps (see Figure 2.8). These are produced following the algorithm in Supplementary Materials – Appendix A. These explainability maps present segmented lungs subdivided into super-pixels that are coloured accordingly to the severity scores assigned to the region containing those super-pixels. The super-pixel colour saturation then gives an insight into the contribution of that super-pixel to the Brixia-score assigned to the corresponding lung region: the higher the saturation, the greater the contribution of the super-pixel to the corresponding severity regional score predicted

by BSNet. Taken together, these explanations placed our solution on the 1<sup>st</sup> position on the leaderboard for the best clinical interpretability according to a panel of health specialists.



**Figure 2.6:** Feature importance scores from the perspective of the Random Forest classifier. See Supplementary Material – Appendix A – Table 1 for the meaning of the reported features.



**Figure 2.7:** Example of explanation based on SHAP values. Red features push the model to predict a severe prognosis for this subject. Blue one push the model to the opposite prediction. Features with longer segments impact more on corresponding prediction. In the example, LDH and Brixia-score contribute more than WBC to pushing the model towards a prediction of severe prognosis. In contrast, pH and SaO2 are the two most contributing features for predicting a mild outcome. See Supplementary Material – Appendix A – Table 1 for the meaning of the reported features.



**Figure 2.8:** Example of lung explainability map. A fully processed CXR (left), its six-region Brixia-score (centre), and the super-pixel explainability map of the lungs (right). The higher the super-pixel saturation, the greater its contribution to the regional severity of the region to which the super-pixel belongs. The regional scores range from 0 (no lung abnormalities) to 3 (interstitial and dominant alveolar infiltrates).

## 2.5 Conclusions

The COVID-19 pandemic exposed the critical challenge of healthcare resource exploitation, motivating the development of AI solutions to support clinical decision-making and patient management. While most prior work concentrated on CXR-based diagnosis, far fewer studies addressed prognosis estimation, and almost none provided interpretable outputs that clinicians could trust.

In this study, we present an AI system that integrates CXR analysis with clinical data collected at triage to predict patient prognosis. Beyond prediction, our framework enriches outcomes with multiple layers of explanation: global feature importance, patient-specific SHAP values, and lung-region maps derived from BSNet. In the context of the COVID-CXR international hackathon, this approach demonstrated both competitive accuracy and superior interpretability, ranking second in predictive performance and first in clinical explainability according to a panel of health professionals, radiologists, and computer scientists.

Notably, even the team achieving the highest accuracy (76.4% balanced accuracy versus our 74.3%) relied on BSNet to generate the Brixia-score, underscoring the strong predictive value of this severity index and the ability of BSNet to generalize across centres without retraining. Our findings align with authoritative clinical evidence showing that severity scores extracted from CXRs improve the prediction of adverse outcomes[58]. Taken together, these results highlight that AI-driven systems for prognosis must not only achieve high accuracy but also deliver transparent, clinically meaningful explanations to foster trust and adoption in real-world healthcare settings.

# Chapter 3 - Towards Robust and Reliable Multimodal 3D Segmentation of Multiple Sclerosis Lesions

---

This chapter is based on the article authored by Coppola et al., published after peer review in the special issue “Open the Brain: Horizons and Challenges in Brain Image Analysis” of Pattern Recognition Letters (2025) [44]. Within the narrative arc on the technical dimensions of trustworthiness introduced in Chapter 1, this chapter marks a deliberate shift with respect to Chapter 2 by introducing robustness, efficient and realistic adaptation, and test-time reliability as primary and independent axes of trustworthiness. These dimensions are orthogonal to interpretability, yet operationally prior in realistic deployment scenarios, since explanations lose practical value if a model’s predictions are degraded under distribution shifts or if the system cannot signal when it is operating on unfamiliar data. Accordingly, this chapter enriches trustworthiness with the ability of a system to maintain acceptable performance under shift, to be repaired with minimal supervision when robustness alone is insufficient, and to expose reliability indicators that allow clinicians to contextualise or defer model outputs. In this setting, robustness, adaptation, and reliability emerge as necessary preconditions for meaningful interpretability in real-world clinical use.

---

## Abstract

Accurate 3D segmentation of multiple sclerosis lesions is critical for clinical practice, yet existing approaches face key limitations: many models rely on 2D architectures or partial modality combinations, while others struggle to generalise across scanners and protocols. Although large-scale, multi-site training can improve robustness, its data demands are often prohibitive. To address these challenges, we propose a 3D multi-modal network that simultaneously processes T1-weighted, T2-weighted, and FLAIR scans, leveraging full cross-modal interactions and volumetric context to achieve state-of-the-art performance across four diverse public datasets. To tackle data scarcity, we quantify the minimal fine-tuning effort needed to adapt to individual unseen datasets and reformulate the few-shot learning paradigm at an “instance-per-dataset” level (rather than traditional “instance-per-class”), enabling the quantification of the minimal fine-tuning effort to adapt to multiple unseen sources simultaneously. Finally, we introduce Latent Distance Analysis, a novel label-free reliability estimation technique that anticipates potential distribution shifts and supports any form of test-time adaptation, thereby strengthening efficient robustness and physicians’ trust.

## 3.1 Introduction

Multiple Sclerosis (MS) is a chronic autoimmune disease that damages the myelin sheath in the central nervous system, leading to a range of neurological symptoms that often worsen over time and significantly impact quality of life [59]. Magnetic Resonance Imaging (MRI) is the gold standard for the diagnosis and monitoring of MS, as it enables neurologists to visualize inflammatory lesions that reflect the disease activity and the response to treatment [60]. While other methods, such as Optical Coherence Tomography (OCT), are increasingly integrated into the clinical management of MS [61], they do not provide direct access to these lesions. For this purpose, multiple MRI sequences are used: T1-weighted (T1w), T2-weighted (T2w) and Fluid-Attenuated Inversion Recovery (FLAIR). FLAIR scans, in particular, are the most sensitive to MS lesions, which appear as hyperintense regions whose interpretation requires careful scrutiny. Accurate lesion delineation on MRI volumes is invaluable for both monitoring and treatment planning. However, manual delineation, still common in clinical workflows, is time-consuming and inconsistent between raters [62]. This inconsistency arises from the natural variability of MS lesions - ambiguous borders, diverse sizes/shapes, similarity to healthy

structures - underscoring the critical need for robust computer-aided segmentation. Research in this area spans from semi-automated tools [63], [64], limited by scalability and generalisability [65], to deep learning approaches facing data-centric limitations: scarcity of large, multi-modal MRI datasets [66], inconsistencies from differing annotation protocols across these often modest collections [67], and significant inter-site variability [59]. Consequently, many approaches are restricted to single-modality inputs or 2D architectures, sacrificing vital cross-modality and volumetric information. This limitation ultimately leads to suboptimal performance when handling the heterogeneity of MS lesions and data.

To improve robustness, some have trained models on multicentre datasets [66], while others adapted models to new distributions via fine-tuning [68], [69], domain adaptation [70], or fully-test-time training [59]. However, these strategies present fundamental downsides. Multicentre approaches assume access to relatively large amounts of labelled data in the target domain, which is often infeasible in clinical practice. Conversely, adaptation methods risk sacrificing global generalisability by specialising the model to a single new site. This dilemma is exacerbated in the context of MS, where datasets are limited in both size and availability.

The MRI heterogeneity issues are an instance of the broader problem of distribution shift [71], often framed by the distinction between in-distribution and out-of-distribution (OOD) samples [72], [73], [74]. However, to our knowledge, no prior work has addressed how to estimate the reliability of segmentation predictions before or after adaptation. This is a key need to build clinical trust.

To address these challenges, we present a study of cross-domain MS lesion segmentation under minimal supervision. Unlike many previous works, our approach leverages a 3D full-volume architecture with complete multi-modal input (T1w, T2w, FLAIR). We focus on three main goals: (i) quantifying how much labelled data is needed to adapt a model to multiple, unseen sites, (ii) exploring cross-domain generalisation via a revised few-shot learning paradigm, where  $t$  instances are drawn from each target dataset, and (iii) proposing a novel, online, differentiable method for test-time reliability estimation based on latent space analysis.

Our main contributions are:

- Minimal fine-tuning quantification: we empirically quantify the minimal fine-tuning effort to improve performance on multiple, diverse MS datasets, offering insights into adaptation feasibility under real-world annotation constraints;
- Few-shot cross-dataset adaptation under multiple shifts: we propose and evaluate a realistic few-shot adaptation paradigm where a minimal number of labelled instances is sampled from each target dataset, enabling joint adaptation across multiple domains with significant distributional shifts;
- Latent Distance Analysis (LDA): we introduce an online, differentiable, test-time reliability estimation technique based on latent feature statistics, enabling uncertainty-aware deployment without requiring additional annotations.

The remainder of this chapter is organised as follows. Section 3.2 reviews related work in MS lesion segmentation, model transfer under distribution shift, and out-of-distribution detection. Section 3.3 details our proposed methodology. Section 3.4 presents the experimental results and discusses the limitations of our approach. Finally, Section 3.5 summarises our findings and outlines the directions for future work.

## 3.2 Related Work

**Segmentation of Multiple Sclerosis Lesions.** Early MS lesion segmentation methods relied on atlas-based [63] or semi-automated pipelines [64] with limited generalisability and scalability [65]. Deep learning approaches introduced more scalable solutions, starting with 2D single-modality models

[68], [75], followed by multi-view CNNs approximating 3D context [76]. Bi- and trimodal setups have been explored [77], [78], [79], though often constrained to 2D data processing. Patch-based 3D architectures improved spatial coherence [80], [81], and models such as LST-AI [60] and dual path 3D CNNs [82] [83] added bi-modal capacity. Still, most works either process sequences of small 3D patches [75], [84], [85], instead of full volumes, or leverage partial modality combinations. Our work combines full-volume modelling with tri-modal input (T1w, T2w, FLAIR) for maximal lesion visibility and structural context exploitation.

**Transferring and Adapting Models under Distribution Shifts.** Domain shifts caused by scanner and site variability often degrade performance on unseen domains [67] [71]. Adaptation is further hampered by the inherent data scarcity of MS scans. Therefore, building intrinsically robust and easily adaptable models is crucial. Hindsholm et al. [66] aimed at robustness via multi-site training, while Kamraoui et al. [83] and Brosch et al. [80] opted for data augmentation and hierarchical modelling. Regarding adaptation to unseen sources, test-time optimisation [59] and few-shot fine-tuning have been proposed. Valverde et al. [86] showed that single-instance fine-tuning is effective. Similarly, Gaj et al. [69] included a few volumes from longitudinal series to enhance segmentations. Differently, Zhang et al. [70] generated target-style inputs trying to bridge domain gaps under annotation constraints. We extend these efforts by quantifying the minimal number of target instances necessary to adapt across multiple datasets, simulating practical scenarios with only a few labelled instances per domain available. We also reformulate few-shot learning as instance-per-dataset rather than instance-per-class, enabling scalable multi-domain adaptation under constrained supervision and multiple simultaneous shifts.

**Detecting OOD samples before adaptation.** Prior to adaptation, detecting whether new samples are out-of-distribution is crucial. Post hoc methods like Maximum Softmax Probability [73], [74] estimate OOD scores from softmax outputs, although often relying on erroneous, overconfident classifiers. To counter this, others leverage latent features [87], [88]. Alternatively, Outlier Exposure methods use auxiliary OOD data to penalise confident predictions on unseen distributions [89], but assume access to known and representative OOD samples, which is often unrealistic in medical settings. In contrast, we introduce Latent Distance Analysis, a test-time, differentiable, label-free reliability estimator that scores new samples based on their distance to training distribution clusters in latent space.

### 3.3 Methods

To experiment with adaptation across datasets and setups, it is first necessary to pre-train a segmentation model that could serve as both experimental baseline and starting point for future adaptations. We detail pre-training in Section 3.3.1, minimal-instance adaptation on individual target datasets in Section 3.3.2, and cross-dataset few-shot adaptation in Section 3.3.3. We eventually describe our LDA technique in Section 3.3.4.

#### 3.3.1 Pre-training a segmentation model

To build a baseline MS lesion segmentation model, we pre-train in a supervised manner a 3D, multi-modal version of the nnUNetV2 [90], a description of which is provided in Table 3.1. When pre-training on an arbitrary labelled dataset D0, we optimise the sum of the following loss functions:

$$\text{Eq. 1.} \quad \mathcal{L}_{Dice}(\hat{y}, y) = \lambda_1 \cdot \frac{2 \sum(y \cdot \hat{y}) + \varepsilon}{\sum y^2 + \sum \hat{y}^2 + \varepsilon}$$

$$\text{Eq. 2.} \quad \mathcal{L}_{BCE}(\hat{y}, y) = \lambda_2 [y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

where the first function is a soft Dice loss, the second one is a standard binary cross-entropy,  $\lambda_1, \lambda_2$  are learnable parameters and  $\varepsilon = 10^{-8}$  is used for numerical stability. Model weights are updated with SGD ( $\eta = 0.01, \rho = 0.99, \omega = 3 \cdot 10^{-5}$ ), using a batch size of 2 volumes for 1000 epochs. To enhance generalisation, we apply a set of independent neurologist-validated data augmentations during training, including vertical flipping and  $\pm 10^\circ$  rotation, both performed with a probability of 0.2.

| Stage                             | Feature maps | Kernel size           | Stride    | # Convs |
|-----------------------------------|--------------|-----------------------|-----------|---------|
| 1                                 | 32           | 3 x 3 x 3             | 1 x 1 x 1 | 2       |
| 2                                 | 64           | 3 x 3 x 3             | 2 x 2 x 2 | 2       |
| 3                                 | 128          | 3 x 3 x 3             | 2 x 2 x 2 | 2       |
| 4                                 | 256          | 3 x 3 x 3             | 2 x 2 x 2 | 2       |
| 5                                 | 320          | 3 x 3 x 3             | 2 x 2 x 2 | 2       |
| 6                                 | 320          | 3 x 3 x 3             | 1 x 2 x 2 | 2       |
| <b>Tri-modal input size</b>       |              | 3 x (128 x 128 x 128) |           |         |
| <b>Total trainable parameters</b> |              | ~ 30.4 million        |           |         |

**Table 3.1:** Our nnUNetV2 3D full-volume architecture for tri-modal (T1w, T2w, FLAIR) input. All Conv3d layers have bias enabled and are followed by InstanceNorm3d ( $\varepsilon = 10^{-5}$ , affine=True) and LeakyReLU activations. The Decoder mirrors the encoder. Heterogeneous inputs are resized to a standard shape of  $128^3$  and then remapped to the original shape.

### 3.3.2 Minimal-instance Adaptation

To assess the feasibility and effectiveness of model adaptation under severe data and annotation constraints on new target datasets  $D_1, \dots, D_M$ , we fine-tune the pre-trained model on each dataset individually. For each target source, fine-tuning is performed using 1, 3, or 5 randomly sampled training instances until training loss (Eqs. 1, 2) or metrics cease to improve by a minimum of  $\Delta = 1 \cdot 10^{-3}$ . We do not use any validation set, skip any hyper-parameter tuning and utilise the high  $\Delta$  to avoid overfitting on fine-tuning instances. The adapted model is then evaluated on the precomputed testing partition of the corresponding dataset. As a baseline, we also evaluate the pre-trained model without any adaptation, representing direct transfer across domains. This series of experiments quantifies the *minimal* amount of supervision required to achieve meaningful performance gains under diverse domain shifts in realistic downstream settings.

### 3.3.3 Cross-domain Few-shot Adaptation

In contrast to isolated domain adaptations, we simulate a realistic deployment scenario in which a model must generalise across multiple unseen sites simultaneously with limited supervision from each. To this end, following the protocol of the previous paragraph, we fine-tune the pre-trained nnUNet on pooled sets created by sampling  $t = 1, 3, 5, 10$ , or 15 instances from each target dataset ( $D_0, \dots, D_M$ ) and merging them into a single adaptation training set of size  $t \times M$ . This setting revises classical few-shot learning by focusing on distributional diversity rather than class diversity, enabling us to investigate simultaneous cross-domain adaptation with minimal annotation effort. This few-shot adaptation technique is also extensible to an arbitrary number of new datasets, i.e.  $M$  can be arbitrary large, by simply sampling  $t$  (with  $t \in [1, \min_{0 \leq i \leq M} \text{size}(D_i)]$ ) labelled instances from each, merging them, and fine-tuning jointly. Noteworthy, the inclusion of  $D_0$  in the adaptation pool serves to prevent catastrophic forgetting of the original training domain — a factor often overlooked in domain adaptation works, but crucial for practical deployment. After each of these five adaptation

experiments, the adapted model is evaluated on the test sets of all participating datasets to measure robustness and generalisation across heterogenous domains.

### 3.3.4 Latent Distance Analysis

**Defining the out-of-distribution score.** To estimate test-time prediction reliability under potential distribution shifts, we introduce Latent Distance Analysis, an online, differentiable, geometry-based method for measuring the deviation of individual test samples from the training distribution in latent space. Our approach operates entirely at inference time and does not require target-domain labels nor training examples, allowing fully-test-time adaptation [91]. Moreover, working in latent space enables lightweight adaptation of only the model components (or part of them) that lead to the target latent features. We first extract latent representations from the pre-trained encoder for each training instance. Let  $\mathbf{z}_i \in \mathbb{R}^d$  denote the latent feature vector for the  $i$ -th training volume. We aggregate all such vectors  $\mathbf{z}_i, i = 1, \dots, N$  and cluster them into  $K$  groups using either: (i) density-based approaches (e.g. DBSCAN), suitable when no a-priori knowledge of acquisition protocols is available, (ii) centroid-based methods (e.g. K-Means), where  $K$  is set based on known factors such as the number of sites or scanners, allowing domain-informed partitioning. This results in  $K$  clusters, each with centroid  $\boldsymbol{\mu}_k \in \mathbb{R}^d$  and covariance matrix  $\Sigma_k \in \mathbb{R}^{N \times N}$  estimated from the cluster members. When a new individual test instance  $x_t$  is acquired, we compute its latent representation  $\mathbf{z}_t$  and the Mahalanobis distance to each cluster centroid:

$$\text{Eq. 3.} \quad d_{t,k} = \sqrt{(\mathbf{z}_t - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{z}_t - \boldsymbol{\mu}_k)}.$$

To aggregate the relative importance of each cluster, we divide each distance by the cluster cardinality  $w_k$ :  $\widetilde{d}_{t,k} = \frac{d_{t,k}}{w_k}$ . We then aggregate the distances using a softmax-inspired function to take into account each cluster's contribution:

$$\text{Eq. 4.} \quad s_t = -\log\left(\frac{1}{K} \sum_{k=1}^K \exp(-\widetilde{d}_{t,k})\right)$$

This yields an OOD score  $s_t \in \mathbb{R}_0^+$  for the test instance  $x_t$ . In particular, scores near 0 indicate strong alignment with the training distribution while scores that deviate from it indicate a completely unfamiliar instance. To better identify strong deviations from cluster typicality, we also present training volume scores. Because real in-distribution samples can vary around these centroids, we do not expect zero OOD scores for training examples. This aligns with the intuition that in-distribution variability should be preserved in uncertainty estimation.

**LDA Operational Properties and Sensitivity.** Here we present some notable properties of our LDA framework. First, it is fully differentiable, allowing  $s_t$  to be directly optimised as a loss function. This opens the door to any kind of test-time adaptation, especially fully-test-time adaptation [91], where only access to test instances and pre-computed cluster statistics is required. Second, in contrast to several OOD detection methods that require batch access, LDA can operate in a fully streaming mode on individual samples. Third, since  $\mathbf{z}_t$  is computed by a model encoder, adaptation can be restricted to the encoder weights only, which facilitates optimisation.

Along with operational properties, it is important to acknowledge LDA's sensitivity to two key factors: the number of clusters  $K$  and the number of training instances  $N$ . To analyse how

$K$  impacts the OOD score  $s_t$ , we begin by defining the minimum normalised distance  $m_t = \min_k(\widetilde{d}_{t,k})$  between the test representation  $\mathbf{z}_t$  and the centroids  $\boldsymbol{\mu}_k$ . We can refactor the exponential sum:  $\sum_{k=1}^K e^{-\widetilde{d}_{t,k}} = \sum_{k=1}^K e^{-\widetilde{d}_{t,k}} \cdot \frac{e^{m_t}}{e^{m_t}} = e^{-m_t} \sum_{k=1}^K e^{-(\widetilde{d}_{t,k} - m_t)} := e^{-m_t} S$  where  $S \in [1, K]$  since all

exponential terms lie in  $[0,1]$ . Then the OOD score becomes  $s_t = -\log\left(\frac{1}{K}e^{-m_t s}\right) = m_t - \log\left(\frac{s}{K}\right)$ . Given that  $S/K \in [1/K, 1]$ , we can derive tight bounds:

$$\text{Eq. 5.} \quad m_t \leq s_t \leq m_t + \log(K)$$

Thus, the score is always at least  $m_t$  and at most  $\log(K)$  higher. Increasing  $K$  cannot reduce  $s_t$  below  $m_t$ , but can increase its spread. In practice, over-partitioning may inflate scores unnecessarily (small clusters imply small  $m_t$ , large  $\widetilde{d}_{t,k}$ ). Conversely, if the data is truly multi-cluster, more clusters help keep  $m_t$  low by ensuring at least one centroid remains close to  $\mathbf{z}_t$ . To set  $K$ , prior knowledge is helpful, although density-based clustering remains a valid, agnostic, and outlier-robust alternative. Next, we examine the effect of dataset size  $N = \sum_{k=1}^K w_k$ . We introduce cluster proportions  $p_k = \frac{w_k}{N}$  and assume they remain fixed as  $N$  grows (e.g. the scanners workload does not change). We also reasonably assume there is a stable generative process behind the training observations. Then, after training and fixing the encoder weights  $\Theta_{\text{enc}}$  we can claim that the centroids are substantially stable as  $N$  grows. To relate  $s_t$ , collecting large amounts of data (i.e. large values of  $N$  that push  $\widetilde{d}_{t,k}$  toward 0) enables the following Taylor expansion:  $e^{-\widetilde{d}_{t,k}} \approx 1 - \widetilde{d}_{t,k} + O(N^{-2})$ . Averaging over  $k$ , we get:  $\frac{1}{K} \sum_{k=1}^K e^{-\widetilde{d}_{t,k}} \approx 1 - \frac{1}{KN} \sum_{k=1}^K \frac{d_{t,k}}{p_k} + O(N^{-2})$ . Taking the logarithm and recalling that  $\log(1 - \epsilon) \approx -\epsilon$  for small  $\epsilon$ , we obtain:

$$\text{Eq. 6.} \quad s_t \approx \frac{1}{KN} \sum_{k=1}^K \frac{d_{t,k}}{p_k} + O(N^{-2})$$

Therefore, as  $N$  increases, the OOD score  $s_t$  decays predictably as  $O(1/N)$ . This reflects the intuition that with more data, centroids stabilise, distances normalise by larger  $w_k$ , and the model becomes more confident about in-distribution membership. In summary, LDA emerges as a robust estimator: its OOD scores are tightly bounded between interpretable limits, governed by key parameters ( $K, N$ ), and converge reliably with increasing data.

### 3.4 Experimental Results

In this section, we describe the datasets used in our experiments (Section 3.4.1) and present the results obtained by the pre-trained segmentation model (Section 3.4.2). We then evaluate its performance when minimally adapted to new target datasets, both individually (Section 3.4.3) and simultaneously (Section 3.4.4). Supplementary Material - Appendix B - Figure 3 and Supplementary Material - Appendix B - Figure 4 show this performance for visual comfort. Next, we assess the ability of the LDA-derived OOD scores to rank-correlate with error metrics (Section 3.4.5). Finally, we outline the main limitations of our approach (Section 3.4.6). The experiments are conducted on a node equipped with NVIDIA A100 GPUs (forward pass speed  $\sim 1$  sec/scan). Training losses and durations, expressed in epochs, are reported in Supplementary Material - Appendix B - Figure 1. Performance uncertainty is estimated using 95% confidence intervals (CI) derived from test set bootstrapping ( $N = 1000$ ). During every evaluation, to remain as dataset-agnostic as possible, we use a segmentation threshold of 0.5. In addition, to probe threshold effects, we conduct sensitivity analyses with multiple thresholds and find that performance remain overall stable, underscoring the robustness of our approach. (see Supplementary Material - Appendix B - Figures 5 and 6).

### 3.4.1 Datasets

For both pre-training and adaptation, we use four publicly available, real-world MS datasets. Each source comprises multi-modal MRI scans (T1w, T2w, and FLAIR) with corresponding expert lesion annotations. These datasets vary in acquisition protocols, scanner types, and cohort characteristics, providing a diverse foundation for studying domain variability and model generalisability under distribution shifts.

- ICPR2024 [92], available for the MS-LesSeg competition<sup>1</sup>, includes 92 training and 22 test scans from about 90 patients, taken at multiple time points and with different scanners. Although collected in a single centre, it exhibits substantial inter- and intra-subject variability.
- ISBI2015 [93] provides 21 annotated longitudinal MRI scans from five patients. All scans follow a uniform imaging protocol and are acquired with the same scanner, offering a highly consistent setup.
- MICCAI2016 [94] is a multicentre dataset comprising 15 training scans (one per patient across three centres) and 38 test scans from four centres. Three centres overlap between training and test sets, which introduces inter-site variability.
- 3DMRMS [95] consists of 30 MRI series acquired at a single site using consistent scanner hardware. While homogeneous in protocol, it adds anatomical and scanner-specific variability to the dataset pool.

We exclude the MICCAI2021 dataset [96] because it targets lesion progression via differential annotations, which deviates from our focus on full lesion segmentation at each time point. To ensure reproducibility and fair comparisons across experiments, we define fixed test sets for each dataset. Specifically, we use the official test sets for ICPR2024 and MICCAI2016, and extracted test sets for ISBI2015 (5 scans stratified across patients and time points) and 3DMRMS (15 scans stratified by acquisition machine). This extraction is necessary because ISBI2015 has no available labels for the official test partition and 3DMRMS does not provide a testing split. For pre-training nnUNet, we use the entire training set of ICPR2024, as it offers a diverse and relatively large collection of multi-modal MS volumes. For minimal-instance adaptation, we sample instances from the training partitions of ISBI2015, MICCAI2016 and 3DMRMS. For cross-domain few-shot adaptation, we instead sample from all sources.

### 3.4.2 Building a state-of-the-art MS segmentation baseline

After a 9-hour pre-training on the entire ICPR2024’s training set, our baseline model achieves state-of-the-art performance on this dataset and improves upon our 2<sup>nd</sup> place in the MS-LesSeg competition [92]. This competition is considered one of the most challenging benchmark for MS lesion segmentation. Also, when directly transferred to external datasets including ISBI2015, MICCAI2016 and 3DMRMS (0-instance fine-tuning configuration), we surpass or remain competitive with previous works according to several metrics (Table 3.2). This solid performance provides ideal foundation for the subsequent adaptation analysis.

|                     | ICPR2024 |     |      |    | ISBI2015 |       |       |       | MICCAI2016 |      |      |      | 3DMRMS |     |      |    |
|---------------------|----------|-----|------|----|----------|-------|-------|-------|------------|------|------|------|--------|-----|------|----|
|                     | Dice     | Rec | Prec | F1 | Dice     | Rec   | Prec  | F1    | Dice       | Rec  | Prec | F1   | Dice   | Rec | Prec | F1 |
| [70]                | -        | -   | -    | -  | ~0.60    | ~0.70 | ~0.55 | ~0.58 | -          | -    | -    | -    | -      | -   | -    | -  |
| [81]                | -        | -   | -    | -  | 0.58     | 0.45  | 0.92  | 0.61  | 0.70       | 0.78 | 0.53 | 0.63 | -      | -   | -    | -  |
| [80]                | -        | -   | -    | -  | 0.68     | -     | -     | -     | -          | -    | -    | -    | -      | -   | -    | -  |
| [83]                | -        | -   | -    | -  | 0.48     | 0.35  | 0.88  | 0.50  | 0.72       | 0.74 | 0.73 | 0.74 | -      | -   | -    | -  |
| [86]                | -        | -   | -    | -  | 0.58     | 0.48  | 0.84  | 0.61  | -          | -    | -    | -    | -      | -   | -    | -  |
| MadSeg <sup>1</sup> | 0.71     | -   | -    | -  | -        | -     | -     | -     | -          | -    | -    | -    | -      | -   | -    | -  |

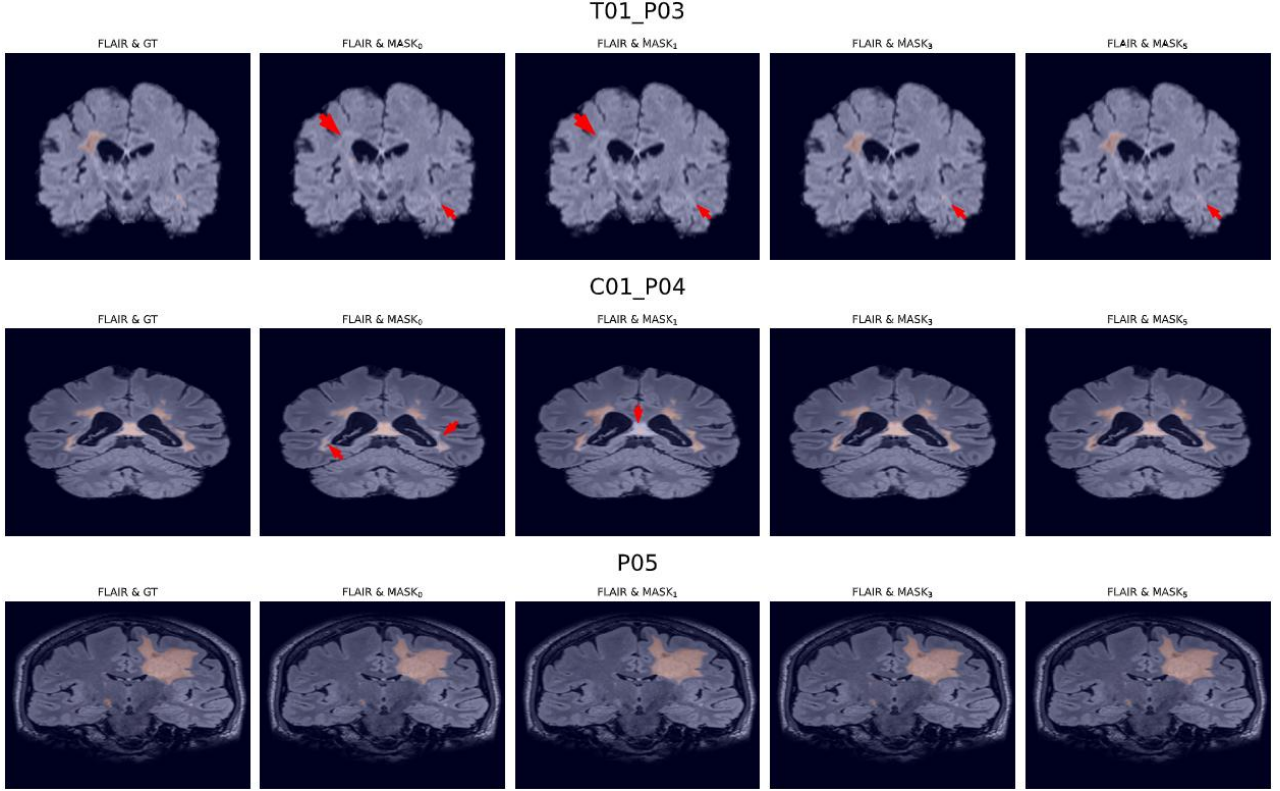
<sup>1</sup> <https://iplab.dmi.unict.it/mfs/ms-les-seg/>

|                                 |               |               |               |               |               |               |               |               |               |               |               |               |               |               |               |               |
|---------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| M3S:Meet<br>MIALMS <sup>†</sup> | 0.71          | -             | -             | -             | -             | -             | -             | -             | -             | -             | -             | -             | -             | -             | -             | -             |
| nnUNet FT<br>0 (*)              | 0.71±<br>0.04 | 0.70±<br>0.05 | 0.79±<br>0.06 | 0.74±<br>0.04 | 0.71±<br>0.07 | 0.70±<br>0.11 | 0.82±<br>0.06 | 0.76±<br>0.07 | 0.62±<br>0.06 | 0.62±<br>0.06 | 0.92±<br>0.08 | 0.74±<br>0.06 | 0.53±<br>0.10 | 0.47±<br>0.09 | 0.95±<br>0.08 | 0.63±<br>0.09 |
| nnUNet FT<br>1 (*)              | /             | /             | /             | /             | 0.63±<br>0.15 | 0.34±<br>0.19 | 0.92±<br>0.05 | 0.5±0<br>.15  | 0.62±<br>0.06 | 0.63±<br>0.06 | 0.81±<br>0.07 | 0.71±<br>0.06 | 0.60±<br>0.08 | 0.54±<br>0.07 | 0.91±<br>0.10 | 0.67±<br>0.08 |
| nnUNet FT<br>3 (*)              | /             | /             | /             | /             | 0.82±<br>0.02 | 0.73±<br>0.05 | 0.94±<br>0.02 | 0.82±<br>0.02 | 0.68±<br>0.06 | 0.67±<br>0.06 | 0.89±<br>0.08 | 0.77±<br>0.08 | 0.67±<br>0.10 | 0.69±<br>0.09 | 0.89±<br>0.12 | 0.78±<br>0.10 |
| nnUNet FT<br>5 (*)              | /             | /             | /             | /             | 0.76±<br>0.07 | 0.60±<br>0.11 | 0.94±<br>0.04 | 0.73±<br>0.07 | 0.70±<br>0.05 | 0.72±<br>0.06 | 0.89±<br>0.08 | 0.80±<br>0.06 | 0.70±<br>0.09 | 0.75±<br>0.07 | 0.85±<br>0.12 | 0.80±<br>0.09 |
| nnUNet J-<br>FT 1 (*)           | 0.62±<br>0.06 | 0.55±<br>0.07 | 0.71±<br>0.09 | 0.62±<br>0.06 | 0.72±<br>0.08 | 0.52±<br>0.12 | 0.94±<br>0.04 | 0.67±<br>0.08 | 0.62±<br>0.05 | 0.47±<br>0.06 | 0.9±0<br>.08  | 0.62±<br>0.06 | 0.59±<br>0.08 | 0.50±<br>0.07 | 0.93±<br>0.12 | 0.65±<br>0.09 |
| nnUNet J-<br>FT 3 (*)           | 0.64±<br>0.04 | 0.68±<br>0.06 | 0.66±<br>0.04 | 0.67±<br>0.05 | 0.80±<br>0.05 | 0.81±<br>0.06 | 0.88±<br>0.06 | 0.84±<br>0.05 | 0.67±<br>0.06 | 0.71±<br>0.06 | 0.89±<br>0.08 | 0.79±<br>0.07 | 0.70±<br>0.09 | 0.72±<br>0.07 | 0.88±<br>0.11 | 0.79±<br>0.09 |
| nnUNet J-<br>FT 5 (*)           | 0.64±<br>0.04 | 0.57±<br>0.05 | 0.85±<br>0.05 | 0.68±<br>0.04 | 0.83±<br>0.03 | 0.77±<br>0.04 | 0.94±<br>0.02 | 0.85±<br>0.03 | 0.65±<br>0.06 | 0.68±<br>0.07 | 0.92±<br>0.08 | 0.78±<br>0.06 | 0.72±<br>0.10 | 0.73±<br>0.10 | 0.92±<br>0.10 | 0.82±<br>0.10 |
| nnUNet J-<br>FT 10 (*)          | 0.67±<br>0.06 | 0.68±<br>0.07 | 0.78±<br>0.07 | 0.72±<br>0.06 | 0.84±<br>0.03 | 0.82±<br>0.05 | 0.92±<br>0.01 | 0.87±<br>0.03 | 0.68±<br>0.05 | 0.70±<br>0.06 | 0.92±<br>0.07 | 0.80±<br>0.06 | 0.74±<br>0.08 | 0.70±<br>0.08 | 0.88±<br>0.08 | 0.82±<br>0.08 |
| nnUNet J-<br>FT 15 (*)          | 0.62±<br>0.05 | 0.65±<br>0.06 | 0.71±<br>0.07 | 0.68±<br>0.05 | 0.82±<br>0.04 | 0.87±<br>0.02 | 0.84±<br>0.06 | 0.85±<br>0.04 | 0.64±<br>0.06 | 0.71±<br>0.05 | 0.84±<br>0.08 | 0.77±<br>0.07 | 0.66±<br>0.10 | 0.71±<br>0.09 | 0.84±<br>0.10 | 0.77±<br>0.10 |

**Table 3.2:** Results and comparisons with the literature across datasets and adaptation setups. “Rec” stands for Recall, “Prec” for Precision. Results preceded by ~ symbol indicate estimates taken by looking at graphs. - symbol denotes experiments not performed by previous studies; / symbol indicates an experiment not included in a given adaptation setup. “FT t” denotes fine-tuning the pre-trained nnUNet with t labelled samples from a single target dataset. “J-FT t” denotes joint fine-tuning with t labelled samples per dataset, pooled across all target datasets. \* denote our solutions. Values are reported with 95% confidence intervals.

### 3.4.3 Minimal-instance Adaptation Yields Consistent Gains Across Datasets

Adaptation to a specific dataset is considered effective when the adapted model outperforms the non-adapted counterparts on that source. Table 3.2 and Supplementary Material – Appendix B - Figure 3 show the performance of the pre-trained nnUNet without adaptation (0-instance fine-tuning), as well as results after fine-tuning on 1, 3, and 5 labelled samples. This allows us to quantify the minimum annotation effort required to achieve meaningful gains under domain shift. On MICCAI2016 and 3DMRMS, we observe consistent improvements in Dice score, Recall, and F1 with increasing numbers of fine-tuning samples, although with diminishing returns. Precision, in contrast, fluctuates slightly on MICCAI2016 and gradually decreases on 3DMRMS—likely reflecting shifts in the model’s sensitivity-specificity balance during adaptation. Notably, fine-tuning with three volumes yields significant gains over both the baseline and 1-instance adaptation across all metrics—except Precision, which follows a distinct trend. These results confirm that fine-tuning with just 1-3 labelled examples consistently improves performance, demonstrating practical low-resource adaptation for clinical settings. For benchmarking purposes, we compile in Table 3.2 the aforementioned results (FT) and 3D segmentation performance from the literature. We highlight that, with minimal supervision, we outperform previous solutions on ISBI2015 across all metrics and surpass state-of-the-art works on MICCAI2016 in terms of Recall and F1. However, for 3DMRMS, comparisons remain challenging, as—to our knowledge—no existing work has addressed full 3D segmentation on this dataset. Examples of segmentations across datasets and various levels of adaptation are depicted in Figure 3.1. The segmentation masks for the ISBI2015 patient reflect overall performance trends: adapting with 1 volume worsens results, suggesting a significant shift, while 3 samples improve alignment with the ground truth. On MICCAI2016, the baseline model detects major lesions but misses inflamed regions near the lateral ventricles. Fine-tuning with 1 volume improves these at the cost of corpus callosum segmentation, which is fully recovered with 3 instances. On 3DMRMS, both baseline and adapted models effectively segment major lesions and localize minor ones, especially with 35 volumes. However, tiny inflamed regions remain undetected on this slice. This overall trend is also discernible in Figure 3.2, which depicts segmentation masks on the same patient volumes.

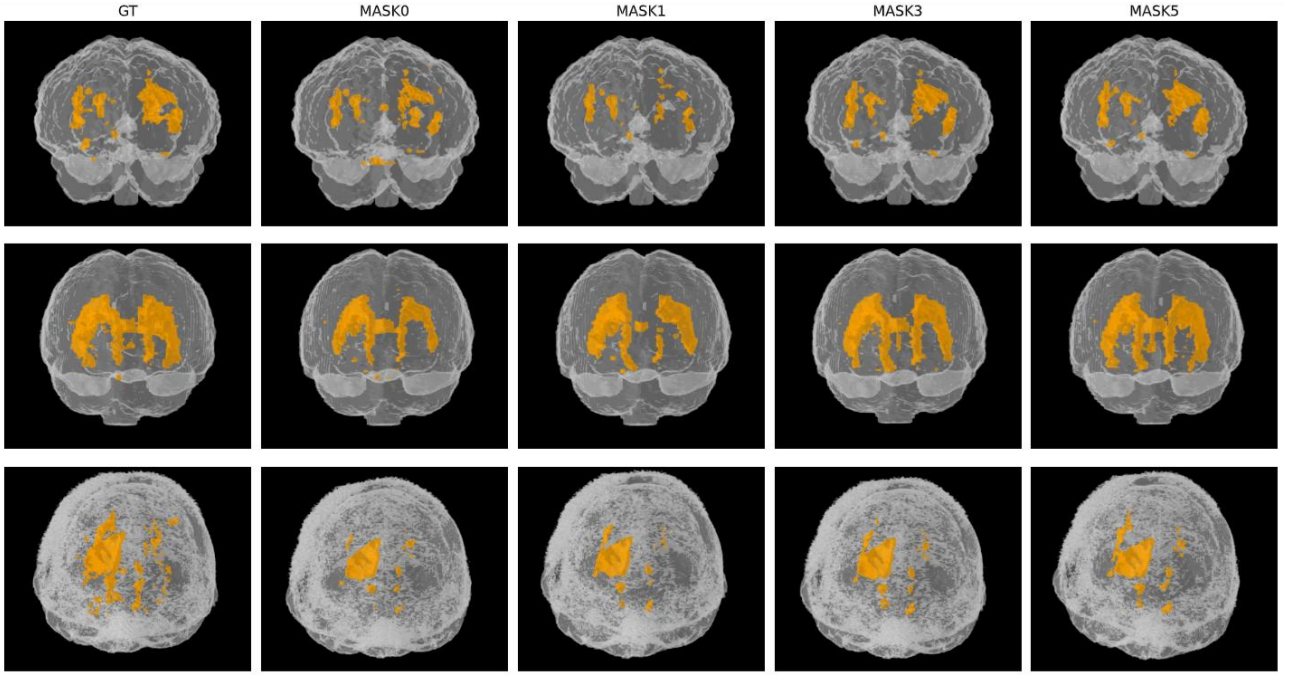


**Figure 3.1:** Examples of segmentation by increasingly adapted models on instances from ISBI2015 (top), MICCAI2016 (centre), and 3DMRMS (bottom). Patient identifiers are reported as row titles. The number of instances used for fine-tuning are reported as subscript of MASK. ‘GT’ stands for ground-truth. Arrows point at major unsegmented/partially segmented lesions. Mid-volume slices are considered. (**Zoom in for better view**).

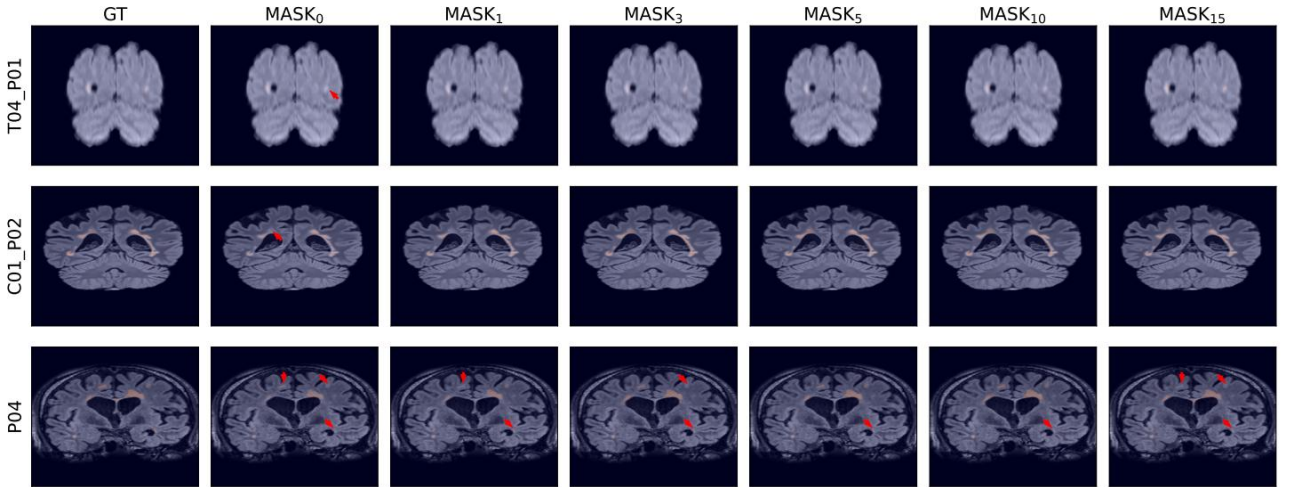
#### 3.4.4 Few-shot Cross-dataset Adaptation Significantly Enhances Multi-domain Generalisation

Adapting a model to a specific distribution can improve performance on that particular domain but may compromise generalisation to others. Table 3.2 and Supplementary Material – Appendix B - Figure 4 show the results of our cross-dataset few-shot adaptation, where pooled sets of 1, 3, 5, 10, and 15 labelled volumes per target dataset, including ICPR2024, are used jointly for multi-dataset fine-tuning. This setup enables the quantification of the minimal cross-domain fine-tuning effort necessary for adaptation. It also allows to evaluate the adaptation’s effect on generalisation across diverse sources simultaneously. The results demonstrate clear and consistent improvements across all unseen datasets and metrics. These gains are evident when compared to both the non-adapted baseline and previous adaptation settings. We also assess how performance changes on the original dataset and find a non-negligible decrease. This reduction, in contrast to previous trends, requires several per dataset for compensation and improvement. Overall, these findings highlight that jointly conditioning adaptation on multiple heterogeneous distributions can drastically improve cross-domain generalisation, even with minimal supervision. This observation significantly strengthens the findings by Hindsholm et al. [66]. Notably, this strategy is especially practical in realistic low-resource scenarios where only a few annotated instances per dataset are available. In these cases, pooling limited supervision across datasets proves more effective and scalable than exhaustive per-dataset adaptations. For both comparative clarity and benchmarking purposes, we compare our results (J-FT) against those obtained in the previous adaptation setups (FT) as well as prior literature (see Table 3.2). On ISBI2015, we further improve all metrics, surpassing the performance of both our previous adaptations and prior works. We also strengthen our Precision and F1 scores on MICCAI2016,

remaining competitive in terms of Dice and Recall. On 3DMRMS, few-shot cross-dataset adaptation delivers significant improvements over previous results in terms of Dice and F1, while only



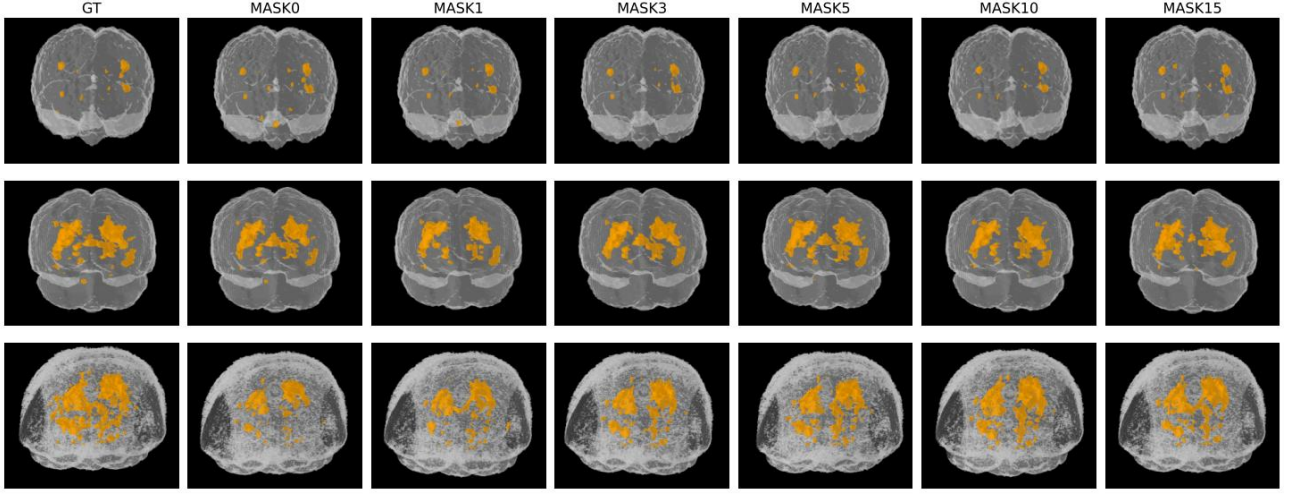
**Figure 3.2:** Examples of 3D segmentations on instances from ISBI2015 (top), MICCAI2016 (centre) and 3DMRMS (bottom) after adaptation with 0, 1, 3, 5 volumes from each individual target dataset. (**Zoom in for better view**).



**Figure 3.3:** Segmentation examples from ISBI2015 (top), MICCAI2016 (centre), and 3DMRMS (bottom) by increasingly adapted models on pooled sets of instances from each dataset. Patient identifiers are reported as column titles. The cardinalities of the pools used for fine-tuning are reported as subscript of MASK. ‘GT’ stands for ground-truth. Arrows point at major unsegmented/partially segmented lesions. Mid-volume slices are considered. (**Zoom in for better view**).

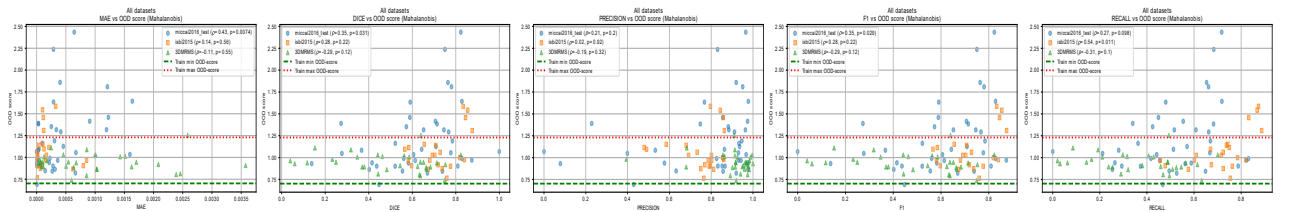
approaching prior Recall and Precision. We hypothesise that conditioning on heterogeneous sources acts as regularization, reducing dataset-specific biases and driving the observed gains in generalisation. Examples of segmentations after cross-dataset adaptation are presented in Figure 3.3. For the ISBI2015 patient, cross-source fine-tuning helps identify very tiny lesions with just a single additional volume; for MICCAI2016 subject, pooled sets of samples help reduce false negatives in already detected lesions; for the 3DMRMS subject, increasingly larger pools enable the localisation of missed regions. This overall trend is also visible in Figure 3.4, which depicts segmentation masks on the same subject volumes.

### 3.4.5 LDA-quantified OOD Score Correlates with Errors



**Figure 3.4:** Examples of 3D segmentations on instances from ISBI2015 (top), MICCAI2016 (centre) and 3DMRMS (bottom) after adaptation with pooled sets 0, 1, 3, 5, 10, 15 volumes from each target dataset. **(Zoom in for better view).**

Before adapting a model to a new distribution, it is crucial to assess whether incoming samples truly lie out-of-distribution. Conversely, when a model has already been adapted, it becomes important to determine whether the adaptation has successfully reduced the degree of OOD-ness, especially without ground truth labels at test time. In both these contexts, our LDA framework offers a practical, online, and label-free mechanism for estimating sample typicality. Knowing that ICPR2024 training volumes were acquired using three scanners—as supported by a t-SNE visualisation presented in Supplementary Material – Appendix B - Figure 2—we leverage this prior knowledge for clustering latent features into  $K = 3$  clusters using k-means. As shown in Figure 3.5, LDA assigns higher OOD scores to samples from unseen sites, effectively flagging them as atypical relative to the training distribution. These scores deviate from zero (representing centroid-based typicality) and capture how far a test instance lies from the densest regions of the latent space. Moreover, OOD scores exhibit rank-correlations with various error metrics across datasets, which strengthens the credibility of LDA as a meaningful unsupervised reliability estimator. However, we also observe that not all rank-correlations are statistically significant or consistently aligned in sign across metrics and datasets. For example, there exists a positive statistically significant rank-correlation between MICCAI2016 OOD scores and corresponding Mean Absolute Errors (MAEs), but the same relationship becomes insignificant on ISBI2015 and changes also sign on 3DMRMS. Conversely, there is a desirable negative rank-correlation between Dice, Recall, Precision and F1 metrics, and OOD scores of 3DMRMS samples. This rank correlation is always statistically except for Precision.



**Figure 3.5:** Scatter plots showing the relationship between OOD scores and various prediction error metrics across all datasets. Positive rank-correlation between OOD scores and MAE is desirable, while a negative rank-correlation is expected for other metrics. Green and red dashed lines indicate the minimum and maximum OOD scores found in the training distribution, respectively. **(Zoom in for better view).**

### 3.4.6 Limitations

Despite strong performance across domains and encouraging LDA results, some limitations remain. The modest size of the public datasets, while broader than in prior works, may still constrain the model’s generalisability. Single-instance adaptation is also effective but not universally so (e.g., ISBI2015 in Figure 3.1), suggesting that minimal collection efforts may still be required. Furthermore, the model occasionally produces incomplete segmentations of small, challenging lesions (e.g., Figure 3.3, instance P04), highlighting an intrinsic task complexity and a clear area for future refinement. Finally, clinical applicability of both our methodology and LDA require validation on larger and more diverse datasets before real-world deployment.

### 3.5 Conclusions and Future Works

In this work, we propose a 3D multi-modal model for MS lesion segmentation that, through minimal adaptation, achieves state-of-the-art performance across four diverse datasets. Using just 1-3 labelled samples, we demonstrate effective fine-tuning both per-domain and jointly across sources, enabling practical deployment in low-resource clinical settings. We also introduce Latent Distance Analysis, an online, differentiable, and label-free reliability estimator that anticipates distribution shifts at test-time. It rank-correlates with segmentation errors, potentially paving the way to lightweight, trust-aware adaptation.

Future work will include designing new loss functions to improve small lesion segmentation, scaling LDA to larger, more heterogenous cohorts to assess its stability across acquisition conditions and, by leveraging its reliability signal, running encoder-only test-time adaptation to evaluate the quality of this lightweight fine-tuning. Both these steps are essential to enable a prospective clinical validation and ensure transferability to routine hospital data. This translation is immediately enabled by our core technical contributions. The demonstrated efficacy of minimal-instance adaptation directly addresses the critical bottleneck of data scarcity, showing that a prospective study is feasible even with a realistically small annotated cohort. Efficient adaptation, combined with LDA reliability scores designed to foster physician’s trust, lays the groundwork for safe clinical translation in collaboration with our neurology and neuroradiology departments. A concrete path forward includes retrospective validation on multicentre clinical archives, followed by small-scale prospective studies embedded into neuroradiology workflows to monitor real-time performance. In this setting, we plan to evaluate the non-adapted model alongside LDA outputs to assess whether predictions and reliability signals support or hinder clinical interpretation. We plan to repeat this process after encoder-only adaptation to examine potential shifts in trust and performance.

# Chapter 4 - HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications

---

This chapter is based on the article authored by Coppola et al., currently under revision at Nature Communications and available as a preprint on medRxiv [45]. Within the narrative arc on the technical dimensions of trustworthiness introduced in Chapter 1, this chapter transitions from transparency-driven model design and domain-specific adaptation strategies to representation learning as a scalable and wide-coverage mechanism for trustworthiness. Here, robustness and adaptability are not abandoned but pursued through more general means: the learning of reusable and transferable representations via large-scale self-supervised pre-training. The central question is therefore not whether a single model instance is interpretable or optimally adapted in isolation, but whether the representations it learns consistently transfer across datasets, tasks, and clinical scenarios under lightweight fine-tuning. Accordingly, this chapter places strong emphasis on extensive and heterogeneous validation as a prerequisite for trust, treating generalisation, efficient adaptation, and stability across settings as empirically verifiable properties. While this new paradigm inevitably reduces direct interpretability, it preserves adaptability, enhances robustness and scalability, and remains compatible with model-agnostic reliability estimators.

---

## Abstract

Deep learning models have shown remarkable performance in electrocardiogram (ECG) analysis, but the limited availability and size of ECG datasets have constrained their success, resulting in systems that are more task specialists than versatile generalists. To counter this, we introduce HuBERT-ECG, a novel self-supervised foundation ECG model pre-trained on a large and diverse dataset of 9.1 million 12-lead ECGs encompassing 164 cardiovascular conditions. By simply adding a proper output layer, HuBERT-ECG can be fine-tuned for a wide array of downstream tasks, from diagnosing diseases to predicting future cardiovascular events. Across 17 diverse real-world scenarios, HuBERT-ECG achieves AUROCs from 0.843 on small datasets to 0.99 on larger sources. When fine-tuned to detect 164 overlapping conditions simultaneously, our model delivers AUROCs above 0.9 and 0.95 for up to 140 and 97 diseases, respectively. HuBERT-ECG can also predict death events within a 2-year follow-up with AUROCs up to 0.91. We release pre-trained models and code as building baselines.

## 4.1 Introduction

The electrocardiogram (ECG) has long been a cornerstone of cardiovascular diagnostics, serving as a non-invasive, cost-effective and widely available tool for assessing cardiac and noncardiac diseases [97]. Through its recurring waveforms, the ECG captures a unique “language of the heart”, encoding both physiological and pathological vital information through distinctive electrophysiological “fingerprints” [98]. This language, although seemingly simple in its structure, reveals complex and meaningful patterns that allow the identification of deviations from expected cardiac functions, providing valuable insights into heart health.

Integrating deep learning (DL) into ECG analysis represents a transformational opportunity to significantly improve its clinical utility [99]. In fact, DL algorithms can transform raw ECG traces into powerful digital biomarkers, enabling early detection, risk stratification, and intervention across a wide range of cardiovascular conditions. Previous studies have demonstrated substantial impact of DL in specific areas of ECG interpretation. For example, DL models have consistently shown solid

performance in detecting conditions where the ECG is the gold standard for diagnosis (e.g. atrial fibrillation, tachycardia, and bradycardia), leading to validated algorithms [100]. Conversely, DL applications in the diagnosis of morphological conditions, such as heart failure, pulmonary thromboembolism and aortic stenosis, are less widespread but growing [101]. Although these conditions rely primarily on imaging modalities, previous works demonstrate that DL can successfully use ECG data to predict cardiovascular conditions beyond those traditionally associated with ECG patterns, including non-cardiac conditions such as liver disease or diabetes [102], providing a valuable complement to conventional imaging approaches [103]. Finally, the use of DL to predict future cardiovascular events (CVEs), such as myocardial infarction and stroke, represents the most ambitious application of this technology. While early results are promising, this area presents greater challenges due to the complexity of predicting outcomes from ECG patterns [104].

From a machine learning perspective, current limitations in ECG dataset size, availability, range of conditions and patient demographics have led to developing ECG models that are more task specialists [105] rather than competent and versatile generalists, traits that typically characterise foundation models [23]. In fact, specialised models often struggle to adapt to new domains or even different data distributions within the same task [106]. For example, an ECG model trained solely to detect atrial fibrillation on a dataset where each recording is labelled as either positive or negative for this condition will be unable to identify other cardiac abnormalities unless retrained on a new dataset. Similarly, if a model is trained exclusively on ECGs from individuals over 65, where atrial fibrillation is prevalent, it might fail to perform well on younger patients. In stark contrast, by pre-training in a self-supervised manner on large and diverse unlabelled datasets, foundation models can learn robust and generalised data representations that are transferable to a wide variety of downstream tasks, requiring only lightweight fine-tuning for specific domains or distributions [21].

Unimodal foundation models have achieved significant success across traditional domains—including natural language processing (e.g., the GPT series [25], [26], [27], [28], BERT [107], T5 [108]), computer vision (e.g., DINO [29], MAE [30]), and audio processing (e.g., Wav2Vec2 [31], HuBERT [32])—by leveraging large-scale self-supervised pre-training on extensive unannotated datasets. More recently, these advancements have also facilitated multimodal learning, where models integrate multiple data modalities to achieve a more holistic understanding of information within data (e.g., Google Gemini<sup>2</sup>, GPT-4o<sup>3</sup>). In the medical domain, progress has been slower due to the limited availability of large medical datasets [21], [33]. However, this trend has begun to shift, leading to the emergence of foundation medical models such as CONCH [35], Prov-GigaPath [36] and Virchow [37] for computational pathology; KAD [38] and CheXzero [39] for radiology; MedSAM [40], for medical image segmentation; EchoCLIP [41] for echocardiography; and BiomedGPT [42] for various biomedical tasks. While recent efforts in electrocardiography attempted to develop foundation ECG models, they invariably lack one or more critical characteristics: large-scale and diverse pre-training data; self-supervised learning to learn generalisable representations; comprehensive evaluation across multiple datasets and tasks (particularly with scarce data); efficient fine-tuning for broad applicability. For instance, several works [109], [110], [111], [112], [113], [114] leverage contrastive-learning methods, but miss the combination of large-scale, diverse pre-training and extensive evaluation. HeartBEiT [115] exploits self-supervision during its large-scale pre-training, but suffers from limited source diversity and narrowly scoped evaluation, testing performance on only three tasks. KED [116] demonstrates impressive zero-shot capabilities via text supervision, but lacks large-scale, diverse pre-training data, omits self-supervision (reaching less general representations [29]) and is evaluated only on form and rhythm conditions, thereby limiting broad diagnostic utility. Most recently, ECGFounder

---

<sup>2</sup> <https://gemini.google.com/>

<sup>3</sup> <https://openai.com/index/hello-gpt-4o/>

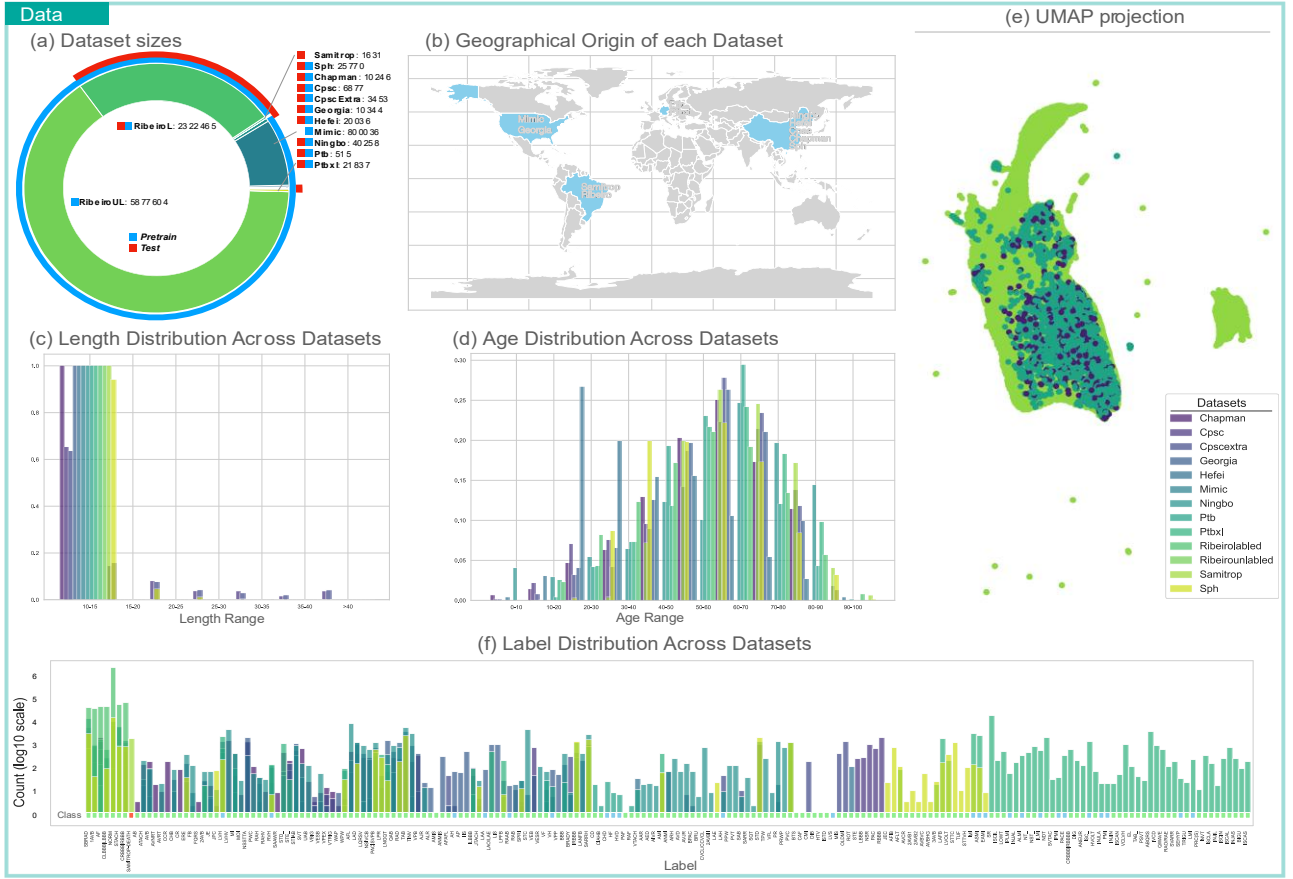
[117] offers the ability to work for the identification of 150 conditions through a supervised training of a model on a large single-site database (Harvard Emory) of 10 million ECGs.

In this work, inspired by HuBERT’s [32] architecture, we present HuBERT-ECG, a new ECG foundation model specifically designed for broad transferability and generalisation across a wide range of datasets and tasks, from diagnosing diseases to predicting future CVEs. HuBERT-ECG is pre-trained through self-supervision on an extensive dataset of 9.1 million ECGs from a large and diverse population across four countries. Noteworthy, over 2.4 million ECGs in our dataset are annotated with one or more conditions from a comprehensive set of 164 diagnoses, allowing HuBERT-ECG to learn and generalise over a wide spectrum of pathologies. These diagnostic labels are derived exclusively from the ECG signal—without integration of clinical, biomarker, or imaging data—and accurately reflect how ECGs are interpreted in current practice by automated systems or ECG-reading physicians. When fine-tuned and evaluated on the unparalleled number of 17 independent datasets, simulating numerous real-world scenarios of varying complexity, size and diversity, HuBERT-ECG demonstrates both efficiency and accuracy, showcasing the potential to serve as a strong foundation for addressing three crucial types of tasks: (1) detecting conditions where the ECG is the primary diagnostic tool (e.g., atrial fibrillation, bundle branch blocks); (2) supporting the identification of conditions where the ECG plays a non-primary but potentially informative role (e.g., heart failure, myocardial infarction); and (3) predicting future cardiovascular events (e.g., death, hospitalization). We stress, however, that the intent of this work is not to introduce a diagnostic model or a universal downstream cardiovascular predictor, but rather to establish and validate HuBERT-ECG as a self-supervised backbone that yields robust, generalizable ECG representations. Downstream evaluations—including those on challenging benchmarks—are used solely to demonstrate the versatility of these representations.

Finally, to support ongoing research on cardiovascular diseases and accelerate innovation, we release HuBERT-ECG models and code to the community<sup>4</sup>. We make available HuBERT-ECG in three model sizes (or configurations)—SMALL (30 million parameters), BASE (93 million parameters) and LARGE (188 million parameters)—to both meet the computational constraints of future users and address a broad range of applications where some model sizes may perform better than others. By sharing insights from comprehensive, benchmarked evaluations on many diverse downstream datasets, we aim to facilitate HuBERT-ECG successful adoption and customisation for research purposes.

---

<sup>4</sup> [Edoardo-BS \(Edoardo Coppola\)](#)



**Figure 4.1:** General information on the 9.1 million ECG dataset. **(a)** Used dataset and their sizes. The blue bar highlights the dataset used for the pre-training while the red one is the testing one. **(b)** Geographical origin of the considered datasets. **(c)** ECG length in seconds. **(d)** Age distribution. **(e)** UMAP projection (uniform sample of 30% of all the dataset for the sake of visualisation). **(f)** Label distribution and grouping according to 3 classes (see the coloured bar below the distribution): Class 0: Green colour for labels for which the ECG the primary diagnosis tool; Class 1: Blue colour for labels for which the ECG is not the primary diagnostic tool but a potentially informative support; Class 2: Red colour for labels for which the ECG is used to predict future CVEs. Label abbreviations and corresponding diagnosis are reported in Supplementary Table 1. **(Zoom in for better view).**

## 4.2 Methods

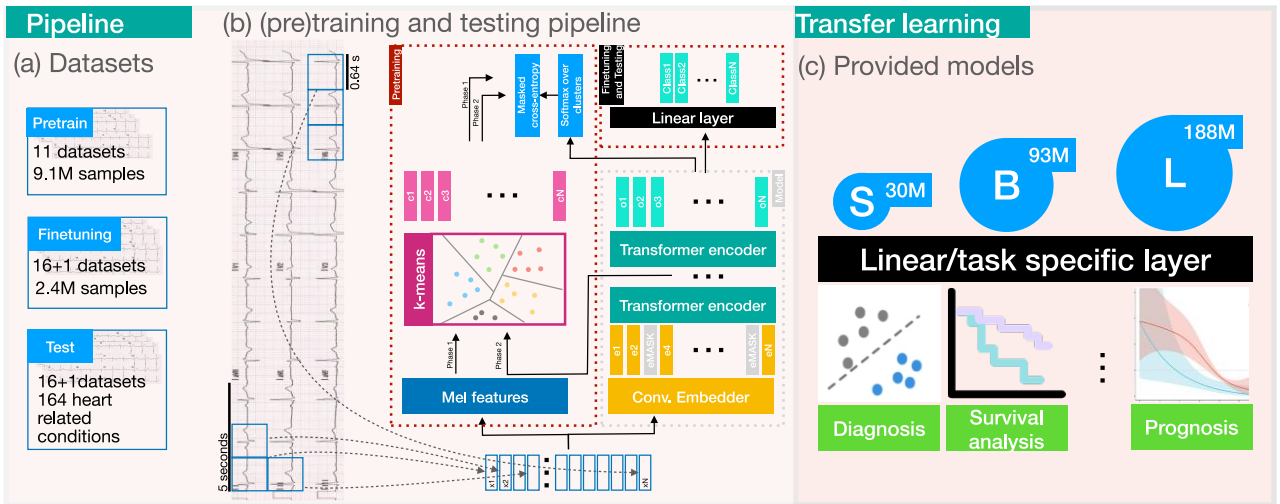
### 4.2.1 HuBERT-ECG Architecture and Theoretical Framework

A schematic representation of HuBERT-ECG architecture, its pre-training phases and its fine-tuning are shown in Figure 4.2b.

**Discovering fine-grained pseudo-labels for ECG fragments.** In order to produce ground-truth pseudo-labels for the self-supervised pre-training, standard 12-lead ECGs are first flattened into one-dimensional signals and then split into non-overlapping fragments. This fragmentation is necessary to frame short portions of the ECG signal from which feature descriptors can be extracted to fit a clustering model. Its purpose is to discover and provide the fragments with cluster assignments (i.e., pseudo-labels) that are finer than the coarse and noisy ECG feature-based ones used as vocabulary for ECGBERT [118]. Conceptually, let  $X = [x_1, x_2, \dots, x_N]$  be a flattened 12-lead ECG composed of  $N$  non-overlapping fragments  $x_i$ , where  $i = 1, 2, \dots, N$ . Let also  $F = [F_1, F_2, \dots, F_N] = [[f_1, f_2, \dots, f_d]_1, [f_1, f_2, \dots, f_d]_2, \dots, [f_1, f_2, \dots, f_d]_N]$  be the sequence of  $d$ -dimensional features  $F_i$

extracted from each fragment, and  $K$  a clustering model fitted on  $F$  that finds  $C$  different clusters. Then, for each index  $i = 1, 2, \dots, N$ , under the hypothesis that  $F_i$  is a good descriptor of  $x_i$ ,  $K(F_i) = z_i \in \{1, 2, \dots, C\}$  is the pseudo-label assigned to  $x_i$  and  $Z = [z_1, z_2, \dots, z_N]$  is the sequence of these assignments for each ECG fragment. For the sake of both terminological flexibility and clarity, the terms *pseudo-labels* and *cluster assignments* will be used interchangeably.

**Representation learning by predicting pseudo-labels for masked embeddings.** HuBERT-ECG operates as a masked prediction model, henceforth denoted as  $h$ . When the flattened 12-lead ECG is fed into  $h$ , the signal is initially processed by a convolutional waveform embedder. This component captures local contextual information and generates  $E = [e_1, e_2, \dots, e_N]$ , a sequence of continuous ECG embeddings. Positional encodings are then integrated into the embeddings to provide the model with information about the original sequential position of each lead segment within the flattened ECG signal. This mechanism allows the model to account for inter-lead relationships that might be less evident after the 12-lead ECG is transformed into a one-dimensional sequence. Subsequently, a set of random indices  $M = \{j \mid j \in \{1, 2, \dots, N\}\}$ , where  $|M| < N$ , is generated to determine which embeddings are to be masked, i.e., replaced by a special learnable embedding  $e_{MASK}$ . After masking, a Transformer encoder [119], which learns global contextual information, consumes the resulting masked sequence of embeddings  $E'$  and produces the output sequence of encodings  $O = [o_1, o_2, \dots, o_N]$ . Eventually, a probability distribution over the  $C$  possible pseudo-labels is constructed for all the  $N$  indices:  $p_h(\cdot \mid E', i) \forall i \in \{1, 2, \dots, N\}$ . This distribution is then used in a standard cross-entropy loss that, however, considers only the indices in  $M$ :  $L(h; E, M, Z) = \sum_{j \in M} \log(p_h(z_j \mid E', j))$ <sup>5</sup>. By training the model  $h$  to predict the cluster assignments of the masked embeddings  $E'_j, \forall j \in M$ , which correspond to fragments of the input ECG not seen by the encoder, the model is forced to learn the most informative representations from the visible segments.



**Figure 4.2:** An overview of the proposed deep learning pipeline. (a) **Datasets:** The model is pre-trained on 11 datasets comprising over 9.1 million samples. Then, the model is fine-tuned and tested on 16 datasets and their aggregation covering 164 heart-related conditions. (b) **Pre-training and Testing Pipeline:** The pipeline begins with Mel feature extraction, followed by k-means clustering, and includes two training phases with masked cross-entropy loss and softmax classification. A convolutional embedder and a Transformer encoder process the data, culminating in a linear output layer for classification. (c) **Transfer Learning and Provided Models:** Three model variants (SMALL, BASE, LARGE) with increasing parameter counts (30M, 93M, and 188M) are pre-trained, tested and, eventually, provided to enable transfer learning for many downstream tasks such as diagnosis, survival analysis, and prognosis. Fine-tuning can be performed by simply adding a task-specific linear layer.

<sup>5</sup> We express cross-entropy loss as a simple sum of log probabilities, instead of the traditional sum over  $p(\cdot) \log(p(\cdot))$ , because, in masked cross-entropy,  $p(\cdot)$  terms are implicitly equal to 1, given the one-hot encoding of labels.

**Multi-task learning via cluster ensembles.** To increase the granularity of the representations being learned during the pre-training, one can use pseudo-labels generated by an ensemble of clustering models characterised by an increasing number of clusters. This approach constitutes a multi-task learning framework, where a new task is introduced with each additional clustering model. Within this framework, each embedding is associated with multiple pseudo-labels, one from each clustering model in the ensemble. The rationale behind this design is that, while a single clustering model may yield imprecise or coarse cluster assignments, an ensemble of models with an increasing number of clusters can mitigate the introduction of noisy targets and provide valuable complementary information to the model. Denoting the number of clustering models composing the ensemble as  $\Gamma$ , which corresponds to the number of tasks being addressed, the loss function can be reformulated as  $L(h; E, M, \{Z^\gamma\}) = \sum_{\gamma \in \Gamma} \sum_{j \in M} \log(p_h^\gamma(z_j^\gamma | E', j))$ , where  $Z^\gamma$  represents the sequence of cluster assignments generated by the  $\gamma$ -th clustering model and  $\{Z^\gamma\}$  is the set of cardinality  $\Gamma$  comprising all such sequences.

**Refining cluster assignments.** Analogous to the work of Hsu et al. [32], the quality of the learned representations can be further enhanced by generating more refined cluster assignments for the ECG fragments that are to be used in subsequent pre-training iterations. Importantly, this “finer” mapping does not refer to a temporal refinement, but rather to a more nuanced clustering of the features extracted from the ECG fragments. To achieve it, latent representations are extracted from intermediate layers of the partially pre-trained model  $h$  and then clustered. Consequently, while the cluster assignments remain aligned with the original fragments in the input  $X$ , their improved quality following this refinement can exert a positive impact on downstream task performance.

#### 4.2.2 Implementation

While the design of HuBERT-ECG mostly aligns with that proposed by Hsu et al. [32], we implement two modifications related to the initial convolutional embedder and the masking strategy. First, since ECGs sampling rate is lower than that of audio signals, there is no need of high-stride convolutions with long filters. A shallower convolutional block with shorter filters proves equally effective. In our model, the convolutional embedder, architecturally tuned to balance embedding information granularity and computational efficiency, generates embeddings at a temporal resolution 0.64 seconds from a flattened 5-second 12-lead ECG sampled at 100 Hz, resulting in 93 embeddings. Second, instead of randomly selecting  $p\%$  of the embeddings as starting points for masked spans, we opt to mask only the selected embeddings without spanning over adjacent ones. This approach avoids the imprecision introduced by overlapping spans, which complicates determining the exact number of masked embeddings. Our method is equivalent to constructing singleton spans and provides more accurate control over the masking process, enabling us to tune the optimal value of  $p$  to use during the self-supervised pre-training (Supplementary Material – Appendix C – Section 2.1). Additionally, masking individual embeddings rather than spans has also demonstrated efficacy in the work of Hu et al. [120].

Following the masking operation, the Transformer encoder consumes the masked sequence of embeddings and produces  $O$ , the output sequence of encodings. The probability distribution over the cluster assignments from a generic clustering model of the ensemble is parameterised by a look-up embedding matrix  $A^\gamma$  with shape  $C^\gamma \times W$  according to the following formula:

$$p_h^\gamma(k | E', i) = \frac{\exp(\cos\_sim(B^\gamma \cdot o_i, A_k^\gamma)/\tau)}{\sum_{c'=1}^{C^\gamma} \exp(\cos\_sim(B^\gamma \cdot o_{c'}, A_{c'}^\gamma)/\tau)}$$

where  $C^\gamma$  denotes the number of clusters found by the  $\gamma$ -th clustering model,  $B^\gamma$  is a projection matrix used to align the dimensionality of  $O$  with the embedding dimension  $W$ ,  $A_k^\gamma$  is the  $k$ -th row of the look-up embedding matrix  $A^\gamma$  that corresponds to the  $k$ -th cluster assignment,  $\cos\_sim(\cdot, \cdot)$  is the cosine similarity between two vectors, and  $\tau$  is the temperature that scales the logits. The superscript  $\gamma$  denotes the  $\gamma$ -th task being addressed when using an ensemble of  $T$  clustering models. Specifically, there are as many projection and look-up embedding matrices as there are tasks in the ensemble.

Given the extensive and heterogenous nature of our dataset, as well as the diverse needs of future fine-tuning users, we propose HuBERT-ECG in three model sizes: SMALL, BASE and LARGE. As we scale the size, we maintain the same convolutional embedder but increase the encoder depth and width, as well as the label embedding dimension  $W$ . Table 4.1 provides a detailed summary of the architectural specifications for the three model configurations. Following pre-training, to fine-tune HuBERT-ECG for specific downstream tasks, we delete the look-up embedding and projection matrices and attach a randomly initialised linear layer atop the encoder to map the pooled output sequence into logits. Specifically, we exploit the Pytorch implementation provided by Hugging Face and adapt its source code to our specific requirements. All models are trained on a computational node equipped with NVIDIA A100 GPUs.

|                         |                        | SMALL          | BASE | LARGE |
|-------------------------|------------------------|----------------|------|-------|
| Convolutional embedder  | strides                | 4, 2, 2, 2, 2  |      |       |
|                         | kernels width          | 10, 3, 3, 2, 2 |      |       |
|                         | channels               | 512            |      |       |
| Transformer encoder     | layers                 | 8              | 12   | 16    |
|                         | internal dimension     | 512            | 768  | 960   |
|                         | feed-forward dimension | 2048           | 3072 | 3840  |
|                         | attention heads        | 8              | 8    | 12    |
| Embedding dimension $W$ |                        | 256            | 256  | 512   |
| Number of parameters    |                        | 30M            | 93M  | 188M  |

**Table 4.1:** HuBERT-ECG architecture summary.

#### 4.2.3 Unsupervised pseudo-label discovery

HuBERT-ECG BASE undergoes two consecutive pre-training iterations. To generate target pseudo-labels for the first one, that do not include any diagnostic or clinical information, we perform a series of k-means clustering experiments (see Supplementary Material – Appendix C - Section 1.1) and find that 100 clusters of feature descriptors consisting of 39-dimensional vectors of Mel Frequency Cepstral Coefficients [121] (MFCCs) (13 base coefficients augmented with their first and second-order derivatives) to be a good and efficient starting point for pre-training. Furthermore, although we obtain only marginal benefit from using a cluster ensemble (see Supplementary Material – Appendix C –Section 2.2), when using multiple clusterings for pseudo-label discovery, the additional pseudo-labels are generated by two MFCC-based k-means models with 200 and 300 clusters, respectively.

For the second iteration, to produce enhanced and more refined pseudo-labels, we run the k-means algorithm again, this time increasing the number of clusters. Specifically, we use 500 clusters of latent representations extracted from the 8<sup>th</sup> encoding layer of HuBERT-ECG BASE after the first iteration. Upon completion of the second iteration, we pre-train the SMALL and LARGE model configurations for one iteration using pseudo-labels generated by clustering representations extracted from 9<sup>th</sup> encoding layer into 500 and 1000 clusters, respectively (Supplementary Material – Appendix C – Section 1.4). Due to memory constraints, we cannot load the entire dataset into memory and, therefore, we opt for a batched implementation of k-means wherein batches of MFCC descriptors, or latent representations, are yielded and the centroid positions are updated incrementally. We set a batch

size of 9300, derived from yielding 93 descriptors from 100 ECGs, and use k-means++ [122] with 20 random restarts for a better centroid initialisation.

#### 4.2.4 Self-supervised pre-training

For each pre-training iteration we reserve 90% of the unlabelled dataset for training and the remaining 10% for internal validation. We set a batch size of 448 instances and find empirically the optimal masking percentage  $p = 33\%$  (see Supplementary Material – Appendix C - Section 2.1) The first iteration consists of 80k training steps, while the second iteration counts 770k steps. The third iteration, instead, consists of 362.5k and 422.5k steps for the SMALL and LARGE model sizes, respectively. We use Adam optimiser with  $\beta = (0.9, 0.98)$ , an initial weight decay of 0.01 and a dropout probability of 0.1 along with a learning rate scheduler that ramps up for the first 8% of the training steps and then decays linearly to zero. Peak learning rates are set to  $5e-5/5e-5/2.5e-5$  for the BASE, SMALL and LARGE model configurations, respectively. In addition, we find benefits from exploiting a *dynamic regularisation* during pre-training (Supplementary Material – Appendix C – Section 1.3). This technique “penalises” the model by incrementally increasing its dropout probability and weight decay if performance does not improve on the internal validation set for *penalty-count* consecutive times. Otherwise, if performance improves on the internal validation set, with respect to the best validation loss or accuracy, the model is “rewarded” by reversing the effects of the last penalty, i.e., decreasing its dropout probability and weight decay. It is important to note that the initial weight decay and dropout probability are the minimum achievable values, while *penalty-count* emerges as a significant hyperparameter to be tuned according to the frequency with which validations are performed. During the first and third iterations, with randomly initialised models, we set the *penalty-count* to 4 and perform an internal validation every 2500 steps. For the second iteration, the internal validation is performed every 5000 steps, while maintaining the same value of *penalty-count*.

#### 4.2.5 Supervised fine-tuning

To fine-tune HuBERT-ECG we follow a straightforward protocol: we attach atop the pre-trained model a randomly initialised linear layer and fine-tune all the weights of the resulting architecture, except those of the convolutional embedder. In contrast to Hsu et al. [32], we do not use the *freezing-steps* hyperparameter, as our experiments do not demonstrate any efficacy in maintaining the encoder’s parameters fixed while training only the final linear layer. Drawing inspiration from Devlin et al. [107], we reduce the batch size to 64 instances and the learning rate to  $1e-5$ . Also, to gain finer control over the identification of a good model candidate, we validate our models every 50 or 500 steps, contingent on the dataset size, hence more frequently than during pre-training. Due to the high number of trainable parameters and the limited size of most datasets, HuBERT-ECG shows a tendency to overfitting but we observe no benefits from using either a strong dropout (up to 0.5) or a high weight decay (up to 0.1), nor from freezing some encoding layers. However, experiments with LayerDrop [123] ( $[0.0, 0.1, 0.15, 0.2]$ ) indicate its potential to mitigate validation loss divergence and metrics degradation. In addition, we note improvements when using a time-aligned random crop as data augmentation technique, a strategy that we also replicate at test time when we take the most confident prediction among those made on multiple crops of the same instance. In summary, our findings suggest that the best performance is obtained when fine-tuning the entire Transformer encoder, zeroing the dropout probability, maintaining the weight decay at 0.01, and sweeping over the LayerDrop probability. We track all experimental runs using Weights and Biases<sup>6</sup>.

---

<sup>6</sup> <https://wandb.ai/site>

## 4.3 Experimental Details

### 4.3.1 Data and Pre-processing

While many ECG-focused studies addressing general clinical questions primarily rely on datasets from the Physionet Challenges [124], [125], overlooking other substantial and valid sources, HuBERT-ECG self-supervised pre-training leverages both public and access-on-demand 12-lead ECG datasets. These include the labelled and unlabelled partitions from Ribeiro [126] (except the predefined, publicly available test-set), CPSC and CPSC-Extra [127], PTB [128] and PTB-XL [129], the publicly available partition from Georgia [130], Chapman-Shaoxing [131] and Ningbo First Hospital [132], the partition from the Tianchi Arrhythmia Competition<sup>7</sup>, Shandong Provincial Hospital [133], and MIMIC-IV ECG [134]. For supervised fine-tuning and evaluation we reuse all the aforementioned sources and include the new SaMi-Trop [135] dataset. We instead exclude the unlabelled partition of Ribeiro and MIMIC-IV for this phase. Although some datasets appear during both self-supervised pre-training and supervised fine-tuning, this is a widely accepted practice in developing foundation models that undergo a self-supervised pre-training. Notable examples include those by Caron et al. [29], [136] in computer vision, Hsu et al. [32] and Baevski et al. [31] in speech representation learning, Azizi et al. [137], Pai et al. [138] and Haghighi et al. [139] in various types of medical image analyses, Ritcher et al. [140] in cell genomics, and ultimately Mehary & Strodthoff [113] and Tang et al. [141] in the ECG domain. In addition, a recent survey [142] on the impact of self-supervised pre-training for diagnostic tasks involving many different types of medical data, points out that pre-training with self-supervision and fine-tuning a model on the same data brings no tangible benefit. This is due to the lack of any information leakage, as no clinical/downstream labels or task-specific signal is exposed during the self-supervised pre-training. In our case, pre-training target variables are also pseudo-labels, discovered in an unsupervised manner using feature descriptors (MFCCs and latent representations) that do not carry any disease-specific, clinical, or dataset-specific information. The test sets used for evaluating the fine-tuned HuBERT-ECG are held out and not used for pre-training or fine-tuning the model. For self-supervised pre-training, all labels and clinical annotations are removed from the pre-training datasets, resulting in an unlabelled dataset of 9.1 million ECGs (Figure 4.2a–Pretrain). The effectiveness of the data selection in capturing the diversity of ECG signals is illustrated through a UMAP [143] projection of the ECG embeddings (Figure 4.1e). To assess HuBERT-ECG’s downstream utility on clinical tasks and datasets of varying complexity, we fine-tune it on each collected dataset<sup>8</sup> and test it on hold-out partitions of such data sources. Additionally, we create a challenging new dataset, which we name *Cardio-Learning*, by merging all the above sources into a new one comprising over 2.4 million ECGs with 164 potentially co-occurring conditions. These conditions can be grouped into three categories based on the diagnostic role of the ECG, as shown in Figure 4.1f. During the creation of Cardio-Learning, to ensure label consistency across different sources, diagnostic labels are harmonised independently by three expert cardiologists. First, diagnostic labels that appear under slightly varying nomenclature are consolidated using unique terms. For example, all instances of atrial fibrillation are uniformly labelled as “AF” instead of “AFIB” or “AF\_” (see Supplementary Material – Appendix C - Table 1). Second, diagnostic concepts that appear under multiple synonyms are standardised to a single diagnostic category. For instance, following established practices [124], [125], ECGs labelled as premature atrial contraction (PAC) or supraventricular premature beats (SVPB) are standardised to PAC|SVPB, as both conditions refer to the same concept. Third, hierarchically broader diagnostic categories are assigned to ECGs already labelled with more specific ones (e.g., Left Bundle Branch Block ECGs also receives the Bundle Branch Block label). However, we avoid two actions: (1) we do not infer more specific conditions from more general labels unless explicitly confirmed in the original

<sup>7</sup> <https://tianchi.aliyun.com/competition/entrance/231754/introduction>

<sup>8</sup> PTB-XL generates 6 different datasets that differ from each other in the presented conditions.

annotation (e.g., Bundle Branch Block ECGs do not receive the Left Bundle Branch Block label); (2) if a constituent dataset includes both a more specific and a more general condition, but explicitly does not assign the general condition to ECGs labelled with the specific one, we respect this distinction and do not add it.

Remarkably, as shown in Figures 4.1b and 4.1d, the ECGs used in this study were obtained from patients exhibiting a broad age distribution and diverse geographical origins spanning four countries. To the best of our knowledge, this constitutes one of the largest and most heterogeneous dataset ever assembled to date in terms of the number of conditions, demographics, and geographic origin of the individuals.

For pre-processing, consistent with the approach of Natarajan et al. [144], a finite impulse response bandpass filter is applied to exclude frequencies outside the [0.05, 47] Hz range, which is reported to encompass the dominant components of P waves, T waves and QRS complexes [145]. Subsequently, we investigate the impact of the sampling rate, which has no standard value and regulates the degree of dilution of the information content, on both the upstream and downstream performance. Our experimental findings indicate that resampling the ECGs at 100 Hz provides a suitable trade-off between downstream performance and training time, while preserving all the relevant physiological information content according to the Nyquist-Shannon theorem (Supplementary Material – Appendix C – Section 1.2). Following this, the ECG signal is rescaled to the [-1, 1] range, mirroring the procedure employed by Natarajan et al. [144]. Finally, unlike other works, we use 5-second 12-lead ECGs instead of 10-second recordings, hence halving memory occupation and accelerating both training and inference. In addition, the selected temporal and spectral parameters align with those derived in Mehari & Strodthoff [146].

#### 4.3.2 Downstream evaluation via supervised fine-tuning

In order to assess the robustness of HuBERT-ECG’s self-supervised representations on clinically relevant datasets and tasks simulating real-world scenarios, we fine-tune the model on each labelled dataset considered in this study. These datasets differ greatly from each other in the number of instances (Figure 4.1a), patients’ age distribution (Figure 4.1d), and spectrum of possible conditions, each belonging to one of the three classes outlined in Figure 4.1f.

When training-validation-test splits are predefined, or when at least the test set is known and defined a-priori, as in the case of PTB-XL, SPH or Ribeiro, we adhere to the original split to allow fair comparisons with previous works. However, for datasets where such a split is not available or applicable (i.e., Hefei, Ningbo, Chapman, CPSC, Georgia), we first extract a fixed hold-out test set, containing 10% of the dataset instances, using a stratified sampling to ensure that all the dataset diagnostic classes are represented. We then perform a 4-fold cross-validation on the remaining instances to tune the hyperparameters of the four models, selecting the best candidate from each fold for the evaluation on the hold-out test set. Finally, we average performance metrics that the four best candidates obtained during inference. Instead, for datasets with an extremely low cardinality compared to the model size, such as PTB, CPSC-Extra and SaMi-Trop, we split the dataset four times in a stratified manner into four *<training, test>* folds. Then, for each fold, lacking a validation set, we skip any hyperparameter tuning and train four models until either a near-zero training error is reached or a minimum improvement delta is not achieved. We then run inference directly on the corresponding hold-out test set with the last model checkpoint. When fine-tuning on PTB, CPSC-Extra and SaMi-Trop, to prevent overfitting on these small datasets, we set a minimal early stopping’s patience of 2 to allow for transitory fluctuations in the training loss descent and set a the minimum improvement delta to  $1 \times 10^{-3}$  for both the loss function and evaluation metrics. Lastly, to construct the Cardio-Learning dataset, we merge the training, validation, and test sets from each constituent dataset to form the corresponding overall training, validation, and test partitions, ensuring the

prevention of any data leakage between these partitions. We perform a 4-fold cross-validation even on Cardio-Learning, which plays simply the role of additional, challenging evaluation benchmark without constituting an intermediate learning step.

Prior to any fine-tuning, we analyse the label distribution of each dataset and exclude the instances labelled with conditions that occur only once. We do so because these conditions are either unlearnable or untestable, as their solitary instances cannot be included in both the training and test sets. However, if a condition occurs twice in the dataset, we assign one instance to the training set and one to the test set, allowing the model to attempt to learn the condition while providing a way to assess its generalisability within the limits of this setup. Eventually, when a condition occurs three times, we place one example per split. During each fine-tuning procedure, we optimise a binary cross-entropy loss function but refer to the macro-averaged multi-label AUROC on the validation set for checkpointing and model selection. The model checkpoint that yields the highest AUROC is used for inference.

### 4.3.3 Evaluation Metrics

To evaluate the downstream performance of HuBERT-ECG across the diverse datasets considered in this study, we primarily rely on the multi-label AUROC metric, which is widely used in the literature to address the challenges of multi-label classification and class imbalance [113], [120], [147], [148]. The multi-label AUROC generalizes the concept of AUROC, originally developed for binary and multi-class problems, to multi-label scenarios where individual instances can be labelled with more than one class. In this context, the AUROC is computed independently for each condition, treating the prediction of each label as a separate binary classification task. The resulting AUROC scores for individual labels are then averaged to obtain the macro-averaged AUROC, providing a holistic perspective on the model’s performance across all diagnostic labels. The AUROC represents the area under the Receiver Operating Characteristic (ROC) curve, which graphically plots the true positive rate (TPR, or sensitivity) against the false positive rate (FPR, or  $1 - \text{specificity}$ ) at varying classification thresholds. Mathematically, TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = 1 - \frac{TN}{TN + FP}$$

where TP, FN, TN, and FP denote true positives, false negatives, true negatives, and false positives, respectively. The AUROC value quantifies the model’s ability to discriminate between instances with or without a specific condition. A ideal classifier achieves an AUROC of 1, with a corresponding ROC curve that reaches the top-left corner of the FPR/TPR plane (coordinate (0,1)). AUROC values closer to 1.0 indicate superior model performance, while a value of 0.5 suggests no predictive capabilities no better than random chance.

On Ribeiro, the largest dataset used in our study, we are able to compute AUPRC as additional metric. This metric is based on the concept of Precision, which quantifies the proportion of correctly predicted positive instances among all instances predicted as positive for a specific diagnostic category and classification threshold. Precision is mathematically defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

High precision values indicate that the model minimises overpredicting positive conditions, thereby reducing the occurrence of false positive alarms. The AUPRC is the area under the precision recall (i.e., sensitivity) curve and evaluates the trade-off between precision and sensitivity across varying classification thresholds. This metric reflects the model’s ability to identify true positive samples without being tricked by confounding factors into predicting false positive instances.

## 4.4 Results

### 4.4.1 HuBERT-ECG is efficient on small-sampled datasets

In diagnostic scenarios with small datasets, such as PTB (515 instances, 17 conditions), CPSC (6,878 ECGs, 9 conditions) and CPSC-Extra (3,453 ECGs, 52 conditions), HuBERT-ECG consistently shows solid performance and exhibits distinct macro-trends. On PTB (Figure 4.3a), the SMALL and BASE configurations perform comparably well when fine-tuned, achieving AUROCs of  $0.843 \pm 0.029$  and  $0.848 \pm 0.03$ , respectively.

On CPSC (Figure 4.3b), the SMALL model size achieves an AUROC of  $0.945 \pm 0.002$  and is slightly surpassed by the BASE ( $0.949 \pm 0.002$ ) and LARGE ( $0.949 \pm 0.006$ ) models. On this dataset we outperform contrastive learning works (e.g., Liu et al. [111], [112], Tian et al. [116] and Zhang et al. [114] achieved AUROCs of 0.932, 0.915, 0.910, and 0.920, respectively) while remaining competitive with models specifically pre-trained and optimised for arrhythmia detection only (e.g., Na et al. [148] with 0.980).

On CPSC-Extra (Figure 4.3c), a more challenging dataset, HuBERT-ECG SMALL performs best, reaching an AUROC of  $0.894 \pm 0.076$ . This result illustrates the advantages of smaller architectures in such intricate, low-data contexts, where larger model configurations might overfit and perform less effectively (BASE  $0.756 \pm 0.02$ ; LARGE  $0.774 \pm 0.023$ ). Noteworthy, fine-tuned models also provide 0.03-0.058 AUROC improvements over their randomly initialised counterparts, with the less relevant exception of the BASE configuration on CPSC-Extra, demonstrating the clear benefit of pre-training even on smaller datasets.

### 4.4.2 HuBERT-ECG remains efficient in increasingly difficult contexts

Diagnostic complexity extends beyond simple metrics like dataset size or the number of classes. It is significantly influenced by the inherent subtlety and difficulty of identifying certain conditions, low per-class sample sizes, and highly skewed label distributions. These factors are well-exemplified by the Georgia and Chapman datasets, each featuring over 10,000 ECGs labelled with 65 and 52 conditions unevenly distributed. In these complex diagnostic scenarios, fine-tuning HuBERT-ECG consistently leads to better performance than training from scratch, yielding improvements of 0.065-0.078 on Georgia and 0.039-0.061 on Chapman.

On Georgia, HuBERT-ECG BASE reaches an AUROC of  $0.832 \pm 0.007$  (Figure 4.3d). For the sake of comparison, Tian et al. [116] and Zhang et al. [114] facilitated the learning problem by limiting the diagnostic scope to 20 and 46 classes, reaching AUROCs of 0.900 and 0.866. When we fine-tune our BASE configuration on these exact subsets, we substantially match the reported AUROCs:  $0.895 \pm 0.079$  on the 20-class problem and  $0.858 \pm 0.093$  on the 46-condition one. In contrast, when directly evaluated on the same condition subsets without retraining, despite being fine-tuned to detect all 65 diagnoses simultaneously, HuBERT-ECG BASE reaches competitive ( $0.853 \pm 0.003$ ) and matching ( $0.855 \pm 0.093$ ) AUROC scores.

On Chapman (Figure 4.3e), all model configurations perform similarly, with the SMALL one topping ( $0.859 \pm 0.008$ ). To address the label distribution skewness, Chapman’s authors suggested grouping conditions into four macro-categories, highly simplifying the diagnostic task. Following this approach, Liu et al. [111] and Kiyasseh et al. [109] reported AUROCs of 0.997 and 0.932,

respectively. When our BASE configuration is fine-tuned under the same grouping, it matches the state-of-the-art with an AUROC of  $0.9945 \pm 0.0009$ . Notably, when we directly test all model configurations on these four broader classes—despite being fine-tuned to recognise the original 52 co-occurring conditions—they achieve highly competitive AUROC values (SMALL  $0.975 \pm 0.002$ ; BASE  $0.975 \pm 0.009$ ; LARGE  $0.984 \pm 0.001$ ) without marked performance loss.

This strong performance underscores HUBERT-ECG’s robustness across varying levels of diagnostic granularity, even without simplifications of the underlying problem during training.

#### 4.4.3 HuBERT-ECG enables accurate diagnostics in large-scale, complicated scenarios.

Trained on extensive datasets, diagnostic foundation models are expected to perform robustly in large-scale, complicated scenarios, where they should identify a disease among many others, even accounting for significant inter-patient variations that can substantially affect diagnostic accuracy. To explore HuBERT-ECG’s transferability to these scenarios, we leverage large datasets with few to tens of possible conditions including Ribeiro, PTB-XL, Hefei, SPH and Ningbo.

On Ribeiro, a large-scale dataset (2.3 million ECGs, 6 conditions), every HuBERT-ECG configuration performs excellently regardless of weights initialization. The BASE model performs best when fine-tuned (0.9989) and surpasses the trained-from-scratch counterpart (0.9984) whereas the SMALL and LARGE configurations perform similarly (Figure 4.3f). Table 4.2, which includes only the best model configuration in terms of AUROC alongside existing models for conciseness, shows that HuBERT-ECG BASE outperforms the model from Ribeiro et al. [126] in terms of sensitivity, specificity, AUROC and area under precision and recall curve (AUPRC). We also surpass ECGFounder [117] on all these metrics. The full comparison including all model sizes, which provide comparable or slightly higher performance on other metrics, is provided in Supplementary Material – Appendix C - Table 7.

The PTB-XL [129] dataset comprises 21,837 ECGs annotated with 71 possible labels covering diagnostic, form and rhythm statements. The authors state that 44 diagnostic conditions can be aggregated into 23 diagnostic subclasses and 5 broader diagnostic superclasses, while there are 12 possible rhythm statements and 19 different form abnormalities. Therefore, one can evaluate a model on 6 different datasets, referred to as PTB-XL *All*, *Form*, *Rhythm*, *Diagnostic (Diag.)*, *Diag. Subclass*, *Diag. Superclass*, that differ from each other in the reported labels. On these sources, fine-tuned models surpass randomly initialised counterparts for almost each model configuration (Figure 4.3g-l). Particularly, HuBERT-ECG BASE achieves an AUROC greater than 0.900 on 5 out of 6 datasets. An extensive comparison with works from the literature is presented in Supplementary Table 8. Many of these studies focus only on some PTB-XL class subsets (e.g., *All*, *Superclass*), limiting possible comparisons, and most of them leverage models specifically optimised for this ECG benchmark. In this context, the dataset-agnostically pre-trained HuBERT-ECG proves competitive with state-of-the-art works tailored on PTB-XL [147] and surpasses other solutions [111], [112], [114]. Unlike prior approaches such as KED [116], ECGFounder [117] and ECGFM [114]—which restricted their scope to simplified subsets of PTB-XL *All* (46 diagnostic classes for KED and 30 for the others) and reported AUROC scores of 0.858, 0.924 and 0.917 respectively—our models demonstrate markedly superior performance, even under more challenging conditions. For fair benchmarking, we fine-tune our BASE model following the same configurations of these methods. Under identical conditions, our model reaches superior AUROCs:  $0.911 \pm 0.003$  on the 46-condition problem and  $0.921 \pm 0.009$  on the 30-class one. Differently, when directly evaluated on these class subsets, despite being fine-tuned for the full 71-class learning problem, HuBERT-ECG models achieve the following AUROCs: 0.900 (SMALL), 0.904 (BASE), and 0.901 (LARGE) on KED’s 46-class subset; 0.918 (SMALL), 0.926 (BASE), and 0.923 (LARGE) on ECGFounder and ECGFM’s 30-class subset. This collection of robust and superior performance, even under more challenging learning conditions, highlights the broader generalisability of HuBERT-ECG’s self-supervised representations compared to the those

learnt by existing supervised or contrastive methods. On Hefei (more than 20k samples, 29 conditions), all fine-tuned models surpass the respective randomly initialised counterparts, with the BASE configuration performing best with an AUROC of  $0.9661 \pm 0.0025$  (Figure 4.3m). On SPH [133], a collection of 25,770 ECGs annotated with 44 diagnostic statements, all our models achieve comparable results, with the LARGE model size reaching the highest AUROC (0.943) (Figure 4.3n). To the best of our knowledge, we are the first to address this dataset in its full complexity. Finally, on Ningbo [132] (34,905 ECGs, 76 co-existing conditions), fine-tuned models generally outperform the trained-from-scratch counterparts, with the BASE size achieving the highest AUROC of  $0.9439 \pm 0.0053$  (Figure 4.3o).

| Condition      | Models          | Sensitivity                  | Specificity                  | AUROC                               | AUPRC                               |
|----------------|-----------------|------------------------------|------------------------------|-------------------------------------|-------------------------------------|
| AF             | Ribeiro et al.  | 0.769                        | 1.000                        | 0.885                               | 0.773                               |
|                | Li et al.       | 1.000                        | 0.996                        | 0.999                               | N.A.                                |
|                | HuBERT-ECG BASE | 1.0000<br>(CI:0.9999–1.0000) | 1.0000<br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) |
| IAVB           | Ribeiro et al.  | 0.929                        | 0.995                        | 0.962                               | 0.807                               |
|                | Li et al.       | 0.864                        | 0.957                        | 0.949                               | N.A.                                |
|                | HuBERT-ECG BASE | 0.7500<br>(CI:0.6071-0.8929) | 0.9981<br>(CI:0.9962-1.0000) | <b>0.9977</b><br>(CI:0.9961-0.9994) | <b>0.9560</b><br>(CI:0.9264-0.9856) |
| CLBBB LBBB     | Ribeiro et al.  | 1.000                        | 1.000                        | 1.000                               | 1.000                               |
|                | Li et al.       | 1.000                        | 0.996                        | 0.998                               | N.A.                                |
|                | HuBERT-ECG BASE | 0.9000<br>(CI:0.8333-0.9667) | 1.0000<br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) |
| CRBBB RBBB     | Ribeiro et al.  | 1.000                        | 0.995                        | 0.997                               | 0.895                               |
|                | Li et al.       | 0.971                        | 0.971                        | 0.989                               | N.A.                                |
|                | HuBERT-ECG BASE | 0.9400<br>(CI:0.8529-1.0000) | 0.9943<br>(CI:0.9899-0.9987) | <b>0.9986</b><br>(CI:0.9974-0.9998) | <b>0.9740</b><br>(CI:0.9513-0.9967) |
| SBRAD          | Ribeiro et al.  | 0.938                        | 0.996                        | 0.967                               | 0.782                               |
|                | Li et al.       | 1.000                        | 0.955                        | 0.967                               | N.A.                                |
|                | HuBERT-ECG BASE | 1.0000<br>(CI:0.9999–1.0000) | 0.9919<br>(CI:0.9864-0.9975) | <b>0.9980</b><br>(CI:0.9954-1.0000) | <b>0.8860</b><br>(CI:0.7504-1.0000) |
| STACH          | Ribeiro et al.  | 0.937                        | 0.997                        | 0.985                               | 0.923                               |
|                | Li et al.       | 0.974                        | 0.970                        | 0.989                               | N.A.                                |
|                | HuBERT-ECG BASE | 0.9730<br>(CI:0.9729–0.9731) | 0.9950<br>(CI:0.9913-0.9987) | <b>0.9990</b><br>(CI:0.9983-0.9997) | <b>0.9640</b><br>(CI:0.9342-0.9938) |
| MACRO-AVERAGED | Ribeiro et al.  | 0.935                        | 0.997                        | 0.966                               | 0.863                               |
|                | Li et al.       | 0.9682                       | 0.9742                       | 0.9818                              | N.A.                                |
|                | HuBERT-ECG BASE | 0.9255<br>(CI:0.8943-0.9567) | 0.9969<br>(CI:0.9954-0.9983) | <b>0.9989</b><br>(CI:0.9983-0.9995) | <b>0.9630</b><br>(CI:0.9403-0.9857) |

**Table 4.2:** Fine-tuned HuBERT-ECG BASE performance on Ribeiro benchmarked against that from Ribeiro et al. [126] and Li et al. [117] according to multiple metrics. Sensitivity and Specificity are obtained setting the threshold to 0.5, as done by Ribeiro et al., for a fair comparison. Performance uncertainty is reported using 95% confidence intervals obtained via empirical bootstrapping using 1000 bootstrap iterations. Label abbreviations and corresponding diagnosis can be found in Supplementary Material – Appendix C - Table 1. A full comparison of all HuBERT-ECG model sizes, which may perform better than the BASE one on certain conditions for certain metrics, is provided in Supplementary Material – Appendix C - Table 7.

#### 4.4.4 HuBERT-ECG can be a good predictor of future cardiovascular events.

Predicting future CVEs is an exceptionally challenging task, especially given the scarcity of publicly available datasets. In this context, the SaMi-Trop [135] dataset offers an opportunity to assess HuBERT-ECG’s predictive capabilities. It includes 1632 ECGs from patients with Chagas disease monitored during a 2-year follow-up period, 104 of which are marked as “death events”. Noteworthy, this dataset is complete with respect to the mortality outcome and no censorship occurred. At the moment of the recording, 268 ECGs were labelled as “normal ” (abbreviated as NORM) recordings by the examining physician. On this dataset, the BASE and LARGE model configurations benefit more than the SMALL one from pre-trained weights. In fact, although competitive, the fine-tuned SMALL version reaches an AUROC of  $0.8705 \pm 0.0279$  while the randomly initialised counterpart achieves a score of  $0.8883 \pm 0.0308$ . In contrast, the BASE ( $0.8627 \pm 0.0422$ ) and LARGE ( $0.8622 \pm 0.0451$ ) configurations perform marginally better than their trained-from-scratch counterparts. We attribute this behaviour to the low number of events within SaMi-Trop, where a single misclassification can result in significant changes. For comparison purposes, we benchmark against the state-of-the-art work by Ferreira et al. [149]. While we use only the raw ECG traces, they used a combination of handcrafted ECG features, sociodemographic variables and self-reported symptoms to fit a Random Forest estimator, and evaluate their model in terms of  $G_{mean}$  score, where  $G_{mean} = \sqrt{sensitivity \cdot specificity}$ . Fixing the same classification threshold, our best model surpasses theirs, reaching a  $G_{mean}$  of 0.78 against one of 0.77, while, in terms of AUROC, no comparison is possible due to the lack of predictions and code.

#### 4.4.5 HuBERT-ECG navigates extremely complex scenarios

In real-world clinical practice, cardiologists are expected to diagnose a wide spectrum of heart-related conditions, adapting to the nuances of each patient's health profile, regardless of how frequently a condition appears. To emulate this complexity, we employ the newly assembled, comprehensive Cardio-Learning dataset, which contains more than 2.4 million ECGs and 164 overlapping conditions, where the ECG is either the primary diagnostic tool, a non-primary supportive diagnostic tool, or used to estimate the risk of future CVEs. The fine-tuned HuBERT-ECG BASE provides a macro-averaged AUROC of  $0.8517 \pm 0.0193$ , while HuBERT-ECG SMALL and LARGE follow with scores of  $0.8453 \pm 0.0058$  and  $0.8403 \pm 0.0378$ , respectively (Figure 4.4a-c). When the conditions are grouped according to the ECG diagnostic role—primary diagnostic tool or informative support for diagnosis—the BASE configuration achieves averaged AUROCs of  $0.84079 \pm 0.019$  and  $0.8777 \pm 0.0197$ , respectively. Most interestingly, without any additional recordings and by leveraging inter-relationship among conditions, HuBERT-ECG raises AUROC values in predicting future deaths for SaMi-Trop patients to  $0.9096 \pm 0.0782$ , significantly surpassing the best model presented for this task in the previous paragraph.

#### 4.4.6 Benchmarking HuBERT-ECG against reported human performance

To contextualize the clinical relevance of HuBERT-ECG, we retrospectively compare the performance of its fine-tuned version on the Ribeiro dataset with that of human readers reported in Ribeiro et al. [126]. That study involved two 4th-year cardiology residents, two 3<sup>rd</sup>-year emergency residents, and two 5th-year medical students, whose sensitivity and specificity values were averaged across the six diagnostic categories. Using the same dataset and label set, and the balanced accuracy as aggregation metric, we find that always at least two fine-tuned HuBERT-ECG configurations outperform medical residents/students on all the six diagnostic categories, and all configurations always surpass at least two groups of residents/students (see Table 4.3). However, it is important to note that this is not a direct head-to-head comparison, and no statistical testing can be performed due

to the unavailability of individual predictions from the human annotators. Nonetheless, the results suggest that HuBERT-ECG, when fine-tuned, is capable of matching or surpassing human-level performance on key diagnostic tasks, highlighting its potential in downstream clinical applications.

| Condition  | Predictor            | Sensitivity                  | Specificity                  | Balanced Accuracy                    |
|------------|----------------------|------------------------------|------------------------------|--------------------------------------|
| AF         | HuBERT-ECG SMALL     | 1.0000<br>(CI:0.9999–1.0000) | 0.9754<br>(CI:0.9656-0.9853) | 0.9877<br>(CI:0.9827-0.9926)         |
|            | HuBERT-ECG BASE      | 1.0000<br>(CI:0.999–1.000)   | 1.0000<br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000)  |
|            | HuBERT-ECG LARGE     | 1.0000<br>(CI:0.9999–1.000)  | 0.9748<br>(CI:0.9644–0.9853) | 0.9874<br>(CI:0.9821-0.9926)         |
|            | Cardiology residents | 0.7690                       | 0.9960                       | 0.8825                               |
|            | Emergency residents  | 0.6150                       | 0.9980                       | 0.8065                               |
|            | Medical students     | 0.9230                       | 0.9890                       | 0.9560                               |
| IAVB       | HuBERT-ECG SMALL     | 0.9290<br>(CI:0.8580-1.0000) | 0.9531<br>(CI:0.9387-0.9675) | 0.9411<br>(CI:0.8983-0.9837)         |
|            | HuBERT-ECG BASE      | 0.7500<br>(CI:0.6071-0.8929) | 0.9981<br>(CI:0.9962-1.0000) | 0.8741<br>(CI:0.8016-0.9464)         |
|            | HuBERT-ECG LARGE     | 1.0000<br>(CI:0.9990–1.0000) | 0.9485<br>(CI:0.9337–0.9637) | <b>0.97425</b><br>(CI:0.9663-0.9818) |
|            | Cardiology residents | 0.6790                       | 0.9970                       | 0.838                                |
|            | Emergency residents  | 0.8210                       | 0.9840                       | 0.9025                               |
|            | Medical students     | 0.929                        | 0.9790                       | 0.9540                               |
| CLBBB LBBB | HuBERT-ECG SMALL     | 1.0000<br>(CI:0.9999–1.0000) | 0.9940<br>(CI:0.9893-0.9987) | <b>0.997</b><br>(CI:0.9946-0.9993)   |
|            | HuBERT-ECG BASE      | 0.9000<br>(CI:0.8333-0.9667) | 1.0000<br>(CI:0.9999–1.0000) | 0.9500<br>(CI:0.9166-0.9833)         |
|            | HuBERT-ECG LARGE     | 0.9330<br>(CI:0.8993–0.9667) | 0.9950<br>(CI:0.9913–0.9987) | 0.9640<br>(CI:0.9453-0.9827)         |
|            | Cardiology residents | 0.9000                       | 1.0000                       | 0.9500                               |
|            | Emergency residents  | 0.8670                       | 0.9990                       | 0.9330                               |
|            | Medical students     | 0.9000                       | 0.9970                       | 0.9485                               |
| CRBBB RBBB | HuBERT-ECG SMALL     | 1.0000<br>(CI:0.9999–1.000)  | 0.9850<br>(CI:0.9776-0.9924) | 0.9672<br>(CI:0.9887-0.9962)         |
|            | HuBERT-ECG BASE      | 0.9400<br>(CI:0.8529-1.0000) | 0.9943<br>(CI:0.9899-0.9987) | 0.9792<br>(CI:0.9214-0.9993)         |
|            | HuBERT-ECG LARGE     | 0.9710<br>(CI:0.9420–1.0000) | 0.9874<br>(CI:0.9798–0.9950) | 0.9825<br>(CI:0.9609-0.9975)         |
|            | Cardiology residents | 0.9710                       | 0.9940                       | 0.8820                               |
|            | Emergency residents  | 0.7650                       | 0.9990                       | 0.9685                               |
|            | Medical students     | 0.9410                       | 0.9960                       | <b>0.9904</b>                        |
| SBRAD      | HuBERT-ECG SMALL     | 1.0000<br>(CI:0.9999–1.0000) | 0.9808<br>(CI:0.9716-0.9901) | 0.9960<br>(CI:0.9857-0.9950)         |
|            | HuBERT-ECG BASE      | 1.0000<br>(CI:0.9999–1.0000) | 0.9919<br>(CI:0.9864-0.9975) | <b>0.9990</b><br>(CI:0.9931-0.9987)  |
|            | HuBERT-ECG LARGE     | 1.0000<br>(CI:0.9999–1.0000) | 0.9796<br>(CI:0.9704–0.9889) | 0.9670<br>(CI:0.9851-0.9944)         |
|            | Cardiology residents | 0.9380                       | 0.9960                       | 0.9355                               |
|            | Emergency residents  | 0.8750                       | 0.9960                       | 0.8725                               |
|            | Medical students     | 0.7500                       | 0.9950                       | 0.9773                               |
| STACH      | HuBERT-ECG SMALL     | 0.9730<br>(CI:0.9729–0.9731) | 0.9816<br>(CI:0.9722-0.9911) | 0.9773<br>(CI:0.9725-0.9821)         |
|            | HuBERT-ECG BASE      | 0.9730<br>(CI:0.9729–0.9731) | 0.9950<br>(CI:0.9913-0.9987) | <b>0.9840</b><br>(CI:0.9821-0.9859)  |
|            | HuBERT-ECG LARGE     | 0.9730<br>(CI:0.9728–0.9732) | 0.9829<br>(CI:0.9734–0.9924) | 0.9790<br>(CI:0.9731-0.9828)         |
|            | Cardiology residents | 0.8110                       | 0.9990                       | 0.9050                               |
|            | Emergency residents  | 0.9460                       | 0.9970                       | 0.9715                               |
|            | Medical students     | 0.8380                       | 0.9960                       | 0.9170                               |

**Table 4.3:** Fine-tuned HuBERT-ECG performance on Ribeiro benchmarked against two 4<sup>th</sup> year cardiology residents (Cardiology residents), two 3<sup>rd</sup> year emergency residents (Emergency residents), two 5<sup>th</sup> year medical students (Medical students) involved in the study by Ribeiro et al. [126]. Both residents' and students'

performance are averages, as reported in the original study. Label abbreviations and corresponding diagnoses can be found in Supplementary Material – Appendix C - Table 1.

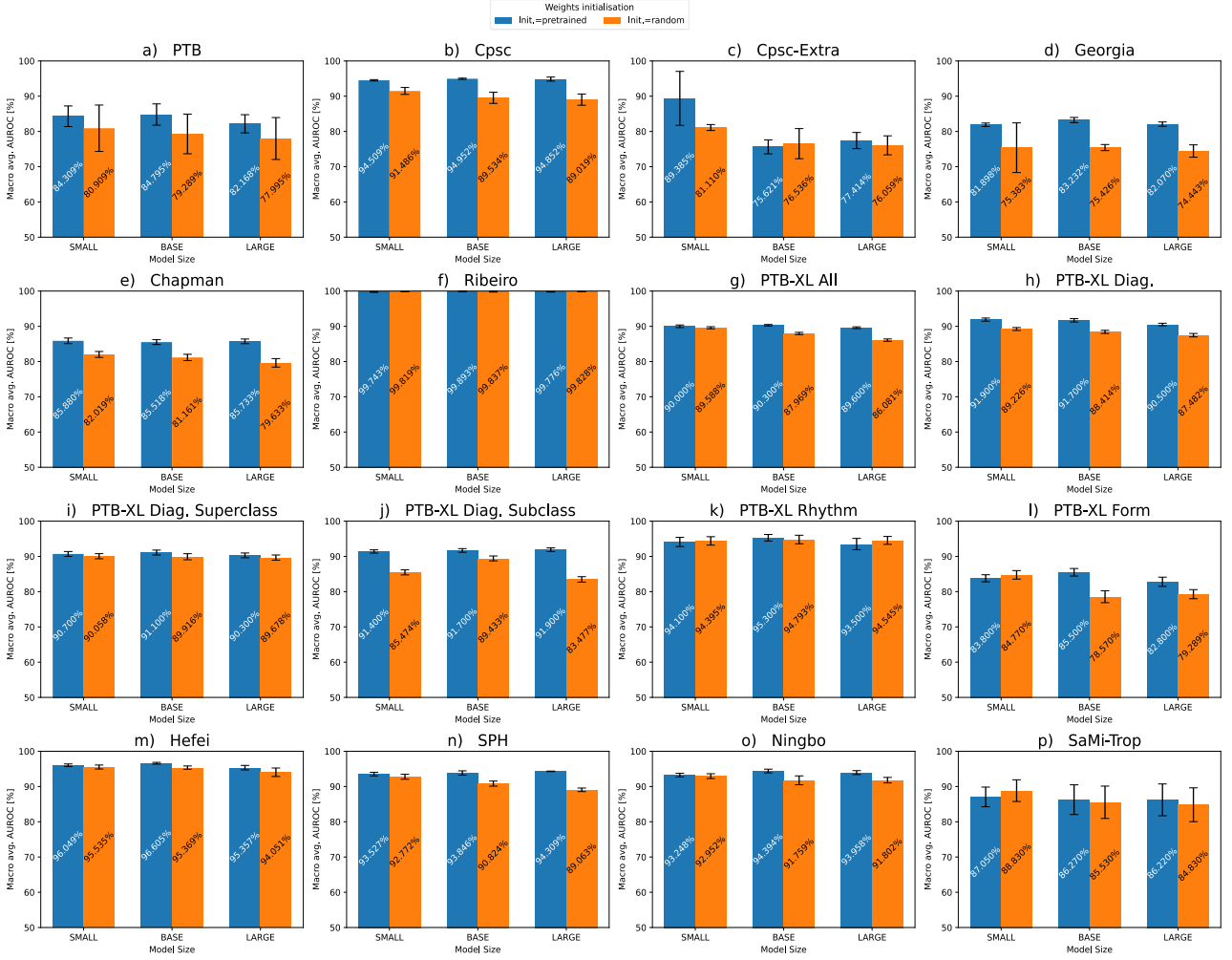
#### 4.4.7 Fine-tuning HuBERT-ECG requires less compute time than training from scratch

While acquiring sufficient labelled data is already a known bottleneck, compute time represents an equally critical constraint in deep learning workflows, particularly in domains where large datasets and complex models necessitate resource-intensive experimentation, hyperparameter optimization, and ablation studies. As demonstrated in Table 4.4, fine-tuning the pre-trained HuBERT-ECG is markedly more efficient across most scenarios. The aforementioned steps, not included in the presented results, are already performed during the self-supervised pre-training of HuBERT-ECG and can therefore be skipped by future users, offering a clear advantage over workflows that require developing models entirely from the ground up.

| Dataset                 | Model size                     |                               |                                |
|-------------------------|--------------------------------|-------------------------------|--------------------------------|
|                         | HuBERT-ECG SMALL<br>(FT / TfS) | HuBERT-ECG BASE<br>(FT / TfS) | HuBERT-ECG<br>LARGE (FT / TfS) |
| PTB                     | <b>3.3k</b> / 3.5k             | 2.3k / 2.15k                  | 1.8k / 1.5k                    |
| Cpsc                    | <b>5.8k</b> / 6.7k             | 4.8k / 4.4k                   | 3.2k / 3.1k                    |
| Cpsc-Extra              | 9.2k / 8.15k                   | 6.7k / 3.9k                   | 5.3k / 2.8k                    |
| Georgia                 | <b>5.2k</b> / 8.13k            | <b>5.4k</b> / 6.3k            | <b>2.8k</b> / 4.8k             |
| Chapman                 | <b>5k</b> / 9.22k              | <b>4.4k</b> / 5.3k            | <b>3.3k</b> / 3.5k             |
| Ribeiro                 | <b>55k</b> / 60k               | <b>45k</b> / 55k              | <b>50k</b> / 40k               |
| PTB-XL All              | <b>12.5k</b> / 27k             | <b>9k</b> / 19k               | <b>8.5k</b> / 19k              |
| PTB-XL Diag.            | <b>11k</b> / 21k               | <b>9.5k</b> / 24.5k           | <b>7.5k</b> / 20k              |
| PTB-XL Diag. Superclass | <b>7k</b> / 18.5k              | <b>6k</b> / 15.5k             | <b>4.5k</b> / 16.5k            |
| PTB-XL Diag. Subclass   | <b>8k</b> / 12.5k              | <b>6.5k</b> / 24.5k           | 8k / 8k                        |
| PTB-XL Rhythm           | <b>7.5k</b> / 19.5k            | <b>8.5k</b> / 25.5k           | <b>5.5k</b> / 20k              |
| PTB-XL Form             | <b>6.5k</b> / 17k              | <b>6.5k</b> / 9k              | <b>5.5k</b> / 9k               |
| Hefei                   | <b>10.1k</b> / 18.2k           | <b>5.8k</b> / 15.2k           | <b>5.2k</b> / 13.7k            |
| SPH                     | <b>10.5k</b> / 27k             | <b>5.5k</b> / 14.5k           | <b>7.5k</b> / 21k              |
| Ningbo                  | <b>15.8k</b> / 35.5k           | <b>12.6k</b> / 21.7k          | <b>10.1k</b> / 20k             |
| SaMi-Trop               | 0.9k / 0.8k                    | 1.1k / 0.6k                   | 0.8k / 0.5k                    |
| Cardio-Learning         | <b>30k</b> / 30.1k             | <b>27.8k</b> / 42.1k          | 32.5k / 27.4k                  |

**Table 4.4:** Comparison between fine-tuning (FT) and training-from-scratch time (TfS), on average, across datasets and model sizes. Time is expressed as number of training steps.

HuBERT-ECG performance on downstream datasets varying weights initializations and model size

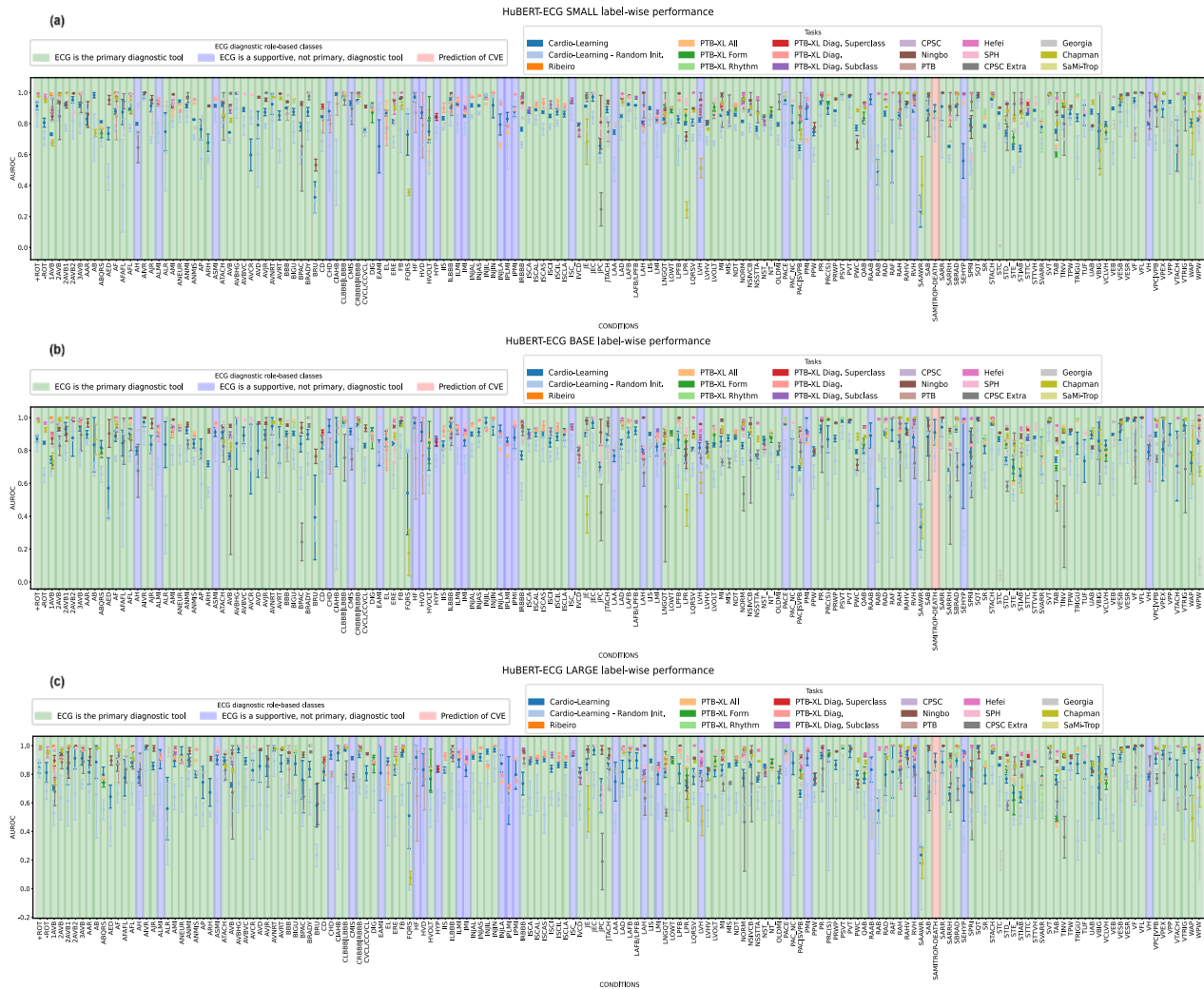


**Figure 4.3:** HuBERT-ECG macro-averaged performance across 16 downstream datasets varying weights initialization and model size. Performance uncertainty on datasets with no predefined test set is reported through standard deviations across fold performance, while on datasets with predefined test sets uncertainty is assessed using 95% confidence intervals obtained via empirical bootstrapping with 1000 bootstrap iterations.

## 4.5 Discussion and Conclusions

In this chapter, we introduce HuBERT-ECG, a novel foundation model for comprehensive ECG analysis available in three scalable configurations (SMALL, BASE, and LARGE) to address diverse clinical needs through a targeted fine-tuning. Pre-trained via self-supervision on an extensive dataset of 9.1 million ECGs from four countries and diverse populations, HuBERT-ECG demonstrates strong performance in identifying 164 cardiovascular conditions. This is rigorously validated across 16 individual datasets and their aggregation (Cardio-Learning), encompassing small to very large ECG collections, annotated with few to many overlapping conditions. Across all evaluations, the three model configurations consistently achieve AUROCs above 0.9 for an average of 136, 138, and 135 conditions and surpass 0.95 AUROC for an average of 95, 94, and 94 conditions, respectively (Figure 4.5a). These results underscore the generalisability of HuBERT-ECG’s self-supervised representations to highly different clinical contexts. Looking ahead, our performance analysis aims not only to shed light on the data requirements for effective fine-tuning but also to guide performance optimisation on new ECG-based datasets.

In diagnostic contexts with small-sampled datasets, which, requiring limited collection efforts, are common, HuBERT-ECG demonstrates both efficiency and accuracy, even when the number of conditions outpace the number of examples, often resulting in highly imbalanced label distributions. In these contexts, the SMALL and BASE configurations are preferable, demonstrating superior generalisation capabilities, whereas the LARGE configuration is less recommended. As the number of patients, ECGs and conditions increases, simulating more realistic diagnostic environments, HuBERT-ECG delivers remarkable results regardless of diagnostic granularity, albeit requiring adequate data collection efforts for fine-tuning. Furthermore, scaling up model size proves advantageous on larger sources, although smaller configurations remain competitive, thus providing desirable flexibility for complicated diagnostic fine-tuning. Notably, on Ribeiro, the largest individual dataset we use, HuBERT-ECG showcases the capacity to achieve excellent precision-sensitivity and specificity-sensitivity trade-offs (Table 4.2), matching or surpassing reported human performance (Table 4.3). Although Figure 4.3f shows that the benefits of pre-training are marginal at the extremely high performance levels obtained from such large datasets with only a few conditions, this does not diminish the value of our approach. In fact, such extensive data collection and labelling efforts—



**Figure 4.4:** HuBERT-ECG label-wise performance for (a) SMALL, (b) BASE, and (c) LARGE configurations across all datasets, including Cardio-Learning. The reported AUROC values are averages obtained from cross-validation and are presented with standard deviations if a label comes from a dataset with no predefined test set. Otherwise, if a label comes from a dataset with a predefined test set, uncertainty is expressed with 95% confidence intervals obtained through empirical bootstrapping (N=1000). Label backgrounds are coloured based on the diagnostic role of the ECG. Supplementary Material – Appendix C - Table 1 provides label abbreviations and their corresponding diagnoses. (**Zoom in for better view**).

combined with only six conditions—are uncommon in real-world settings. Nonetheless, when benchmarking our performance against prior studies, we demonstrate our straightforward fine-tuning consistently leads to high performance, frequently approaching or surpassing that of models specifically developed for individual datasets or condition groups (e.g., arrhythmias). In addition, aiming at learning efficiency, we show that fine-tuning the pre-trained HuBERT-ECG generally requires less compute time than training from scratch, which also demands costly hyper-parameter tuning and ablation studies. Ultimately, when comparing with other methods proposed as foundation models, HuBERT-ECG showcases matching or stronger performance even under more challenging learning conditions.

These findings underscore the high generalisability and utility of its self-supervised ECG representations, especially when data is overall scarce, certain classes are underrepresented, or when compute time is a considerable constraint.

Furthermore, HuBERT-ECG exhibits robust performance in highly complex clinical scenarios exemplified by Cardio-Learning, which aggregates ECGs from diverse sources and countries without task-specific separation and creates a realistic representation of heterogeneous real-world data. In this challenging context, HuBERT-ECG effectively identifies conditions for which the ECG serves as either a primary or informative support signal for diagnosis. Crucially, fine-tuning on Cardio-Learning enables HuBERT-ECG to leverage important inter-relationships among conditions, leading to substantial improvements in AUROC for label prediction—even for conditions that appear only in single subsets. For example, when fine-tuning the BASE model on PTB, the AUROCs for Coronary Heart Disease (CHD) and Heart Failure (HF) are 0.7656 and 0.7514, respectively. Remarkably, when fine-tuned on Cardio-Learning, these scores increase significantly to  $0.9488 \pm 0.0399$  for CHD and  $0.9674 \pm 0.0566$  for HF (Figure 4.5b).

Similarly, in the highly challenging task of predicting future mortality events, HuBERT-ECG achieves mean AUROCs of 0.8784, 0.9096, and 0.8857 for the SMALL, BASE, and LARGE configurations, respectively, with average improvements of 1.31, 4.69, and 2.35 percentage points (Figure 4.5c). These findings strongly suggest that training a model to recognize a broad spectrum of ECG-derived diagnostic labels can significantly enhance its performance in complex prediction tasks, including those where the ECG plays a non-primary but potentially informative role. Importantly, although the model is trained on labels from ECG readings, its rich representations can be further fine-tuned using high-quality clinical endpoints (e.g. electronic health records, imaging and/or genomic data), potentially unlocking even greater performance on tasks beyond traditional ECG applications.

Another relevant finding from HuBERT-ECG’s label-wise performance on Cardio-Learning is its low prediction variance (Figure 4.4). This suggests that merged labels within Cardio-Learning refer to consistent concepts, thus (1) excluding major label inconsistencies, and (2) supporting both our label harmonisation process and HuBERT-ECG’s robustness to label noise. Furthermore, a visual inspection of the fine-tuned HuBERT-ECG attention mechanism confirms its ability to attend to clinically relevant ECG segments rather than arbitrary noise (see Supplementary Material – Appendix C – Section 3).

While HuBERT-ECG has been rigorously tested across diverse cardiovascular conditions, our findings suggest its considerable potential to extend its application toward more personalised clinical insights. These include predicting patient response to therapies or identifying non-cardiac conditions through ECG analysis, paving the way for truly personalised cardiovascular care. For example, in the management of heart failure, HuBERT-ECG could be fine-tuned to predict patient responsiveness to resynchronisation therapy, thereby optimising patient selection criteria [150]. This capability would facilitate a more tailored therapeutic approach, aimed not only at predicting adverse events but also at enhancing outcomes, reducing hospitalisations, and improving quality of life. Such advancements

could fundamentally transform cardiovascular precision medicine, with the use of foundation ECG models for helping clinicians tailor therapies with greater precision [151].

Despite the highly encouraging results, our approach has certain limitations. First, although the data collection spans more countries than in previous studies, it does not include large and highly populated regions (e.g., Africa and India), which we aim to incorporate in future versions. Second, the scarcity of accessible ECG datasets for predicting future CVEs partially limits our ability to fully assess the HuBERT-ECG's potential in this still underexplored domain. Broader access to suitable datasets for testing the model in this challenging field is desirable to accelerate research. Third, direct and fair comparisons with previous systems, both deep-learning-based and rule-based, remain challenging. Many existing works focus on limited subsets of available datasets (often reducing the actual number of conditions), do not open-source their implementations or evaluation datasets, or do not follow standard practices in metric computation, hampering reproducibility and equitable benchmarking. Lastly, we observe that while some conditions are better classified in Cardio-Learning than in their original datasets, others show lower performance. This discrepancy may stem from erroneous labels or incomplete assignments. For example, in PTB-XL, a major label source, key terms in reports are extracted to retrospectively infer the physician's diagnostic confidence, with labels assigned even at low-confidence levels. Conversely, other datasets may rely on more robust labelling protocols, potentially resulting in identical ECGs from different sources receiving different diagnoses. Additionally, less experienced physicians may overlook certain conditions, while overly confident ones might exclude or place less emphasis on others, as also found by Li et al. [117]. These inaccuracies can reduce the data efficiency, requiring more data to compensate for their effects.

# Chapter 5 - Toward a 7M-Parameter Any-Lead ECG Foundation Model: Methodology and Preliminary Results

---

This chapter is based on an ongoing research project led by Edoardo Coppola and co-authored with Stefano Fiorini (Huawei Research), Pietro Liò (University of Cambridge), Mattia Savardi, and Alberto Signoroni (University of Brescia). The work is currently in preparation for submission to ICML 2026. Although the study is not yet finalised, it is included in this thesis because it represents a principled extension of the representation-learning paradigm developed in Chapter 4, addressing a fundamental and underexplored dimension of trustworthiness: structural robustness under acquisition constraints. Specifically, this chapter investigates whether representations learned at scale remain usable and reliable when critical parts of the input structure are missing or altered at deployment time. By focusing on lead-agnostic and any-lead ECG modelling at scale, the chapter explores the limits of transferability, adaptability, and reliability when standard assumptions about data completeness no longer hold. As such, this work is presented not as a conclusive result, but as a boundary case that exposes open challenges and informs future directions for trustworthy medical foundation models.

---

## Abstract

In this in-progress work, we present the methodology and preliminary results of a 7M-parameter foundation model for any-lead electrocardiogram (ECG) representation learning. Prior self-supervised ECG models are largely tied to the 12-lead standard and often rely on complete-graph attention, limiting scalability and efficiency. We instead encode ECGs as sparse, physiology-grounded spatiotemporal graphs, and leverage masked node modelling together with stochastic lead sampling to learn general-purpose, lead-invariant representations. Stage-1 pre-training distills the rich information content of 9.2M 12-lead ECGs from five countries, yielding a tiny 2-layer Graph Attention Network-based model that surpasses HuBERT-ECG BASE (iteration 1) in linear evaluation despite 90M fewer parameters. Future work will complete self-supervised pre-training, present a broad downstream evaluation on 21 clinical and wearable datasets, and introduce quantitative lead-importance analysis across 164 cardiovascular tasks. We will release pretrained models and code.

## 5.1 Introduction

Cardiovascular diseases are the leading cause of death worldwide, accounting for roughly a third of all global deaths<sup>9</sup>. Early detection and monitoring are therefore central to public health, and the electrocardiogram (ECG) is one of the most accessible and clinically informative tools for diagnosing and stratifying the wide spectrum of cardiac conditions [99].

Motivated by the success of foundation models in other domains, the ECG literature is now rapidly moving in the same direction [45], [114], [115], [116], [152]. Foundation models are trained on broad data through self-supervision and can be adapted to many downstream tasks with minimal effort [21], [23]. Most existing solution still focus on the standard 12-lead clinical setting, where data availability is higher and diagnostic coverage is richer. For instance, Tian et al. [116] developed a ECG-Text model pre-trained in a contrastive fashion but limited its application scope to rhythm and form conditions only; Vaid et al. [115] scaled up self-supervised pre-training on millions of ECG images, but considered only three evaluation tasks; Zhang et al. [114] pre-trained on multi-centre data and increased evaluation coverage to several datasets; Li et al. [152] trained on 10M ECGs but lacked

---

<sup>9</sup> [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))

self-supervision and scoped limited transferability. Finally, HuBERT-ECG met the foundation-model definition provided by the authoritative work of Bommasani et al. [21] (“... *trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks*”), being pre-trained through self-supervision on 9.1M ECGs from 4 countries and evaluated on 17 independent datasets covering up to 164 different tasks.

However, real-world ECG acquisition is often not based on 12 leads. Emergency settings, low-resource clinics, and especially wearables provide fewer leads, missing leads, or non-standard configurations. A powerful any-lead foundation model would therefore be transformative in practice: it could reuse knowledge distilled from large 12-lead corpora to support reliable inference on arbitrary lead subsets, enabling robust deployment across heterogeneous clinical workflows and consumer devices. Two recent works [110], [153] are important early steps in this direction. The work by Oh et al. [110] is, to our knowledge, the earliest attempt at any-lead ECG modelling. They demonstrated feasibility but relied on zero-padding missing leads, which preserved transferability at the cost of computational inefficiency and artefact injection. Monachino et al. [153], building on a similar architecture but with a different pre-training objective, improved efficiency by avoiding padding and introducing a lead-adaptive front end that maps variable-lead inputs into a fixed embedding set. This is a meaningful improvement at the input level, although the resulting fixed-size embedding sequence is still fully processed in later layers, leaving headroom for efficiency gains in the encoder itself.

In this in-progress work, we build on this emerging line of work and present the methodology and preliminary results of a tiny and efficient graph neural network-based foundation model that operates on sparse, clinically grounded spatiotemporal ECG graphs. We aim to pre-train such model in two stages on 9.2M 12-lead ECGs from five countries, covering up to 164 cardiovascular tasks, using self-supervised Masked Node Modelling (MNM) objective together with stochastic lead sampling. Lead sampling contribution is twofold: to serve as a strong structured augmentation, and as the key mechanism through which we distill the rich information content from 12-lead recordings into lead-invariant ECG representation. This training recipe ultimately aims to support transfer to any downstream lead configuration. After one pre-training stage, even with its small footprint (~7M parameters), the model already surpasses HuBERT-ECG BASE (iteration-1) under linear evaluation, despite the latter’s much larger capacity and longer pre-training. These preliminary results, presented and motivated in the following sections, are already highly encouraging, but the completion of the second pre-training stage and the broad downstream evaluation on 21 independent datasets with varying lead count and configuration will complete the picture. In addition, we will quantify, condition by condition, how performance varies with the number and identity of leads, effectively producing a lead-importance map analogous to feature-importance scores. Eventually, we will release code and pre-trained models to support the community in any-lead ECG research and analysis.

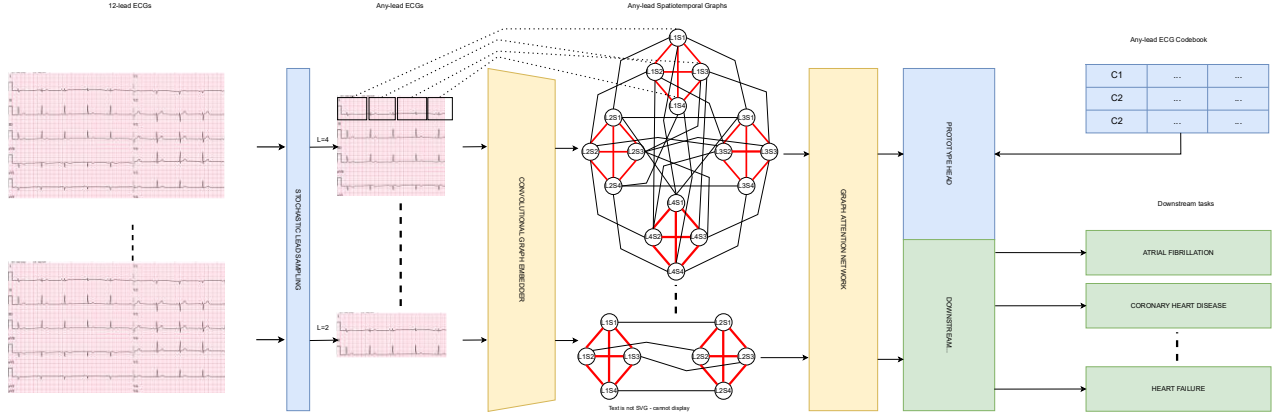
The remaining of this work is organised as follows: Section 5.2 describes our methodology and architecture; Section 5.3 presents the datasets used and experimental details; Section 5.4 showcases and motivates the highly promising results obtained so far; eventually, Section 5.5 concludes the chapter and discusses potential implications and uses of this work.

## 5.2 Methods

A pictorial representation of our architecture, spatiotemporal topology, self-supervised pre-training and supervised fine-tuning is provided in Figure 5.1.

### 5.2.1 Learning a fine-grained codebook for ECG segments

Pre-training a model to predict pre-computed frame-level prototypes from a  $C$ -class codebook learnt by a model  $\mathcal{K}$  has proven effective for speech [32] and, more recently, for 12-lead ECGs [45], [154]. This codebook is defined as a dictionary of prototypes, where each of the  $C$  classes represents a unique, characteristic geometric pattern or template recurrently found across the ECG segments. However, prior works assume a fixed 12-lead configuration or necessitate presenting the multi-lead ECG as a single-channel signal before learning the codebook. This lead rigidity may prevent generalization to arbitrary or missing leads, while the 12-lead flattening procedure discards essential structural properties such as inter-lead correlations and temporal synchrony. To overcome these limitations, we generalize the prototype discovery to operate independently within each lead, preserving spatial and temporal structure while supporting arbitrary number and combination of leads.



**Figure 5.1:** Overall methodology. During pre-training, random lead configurations are sampled from 12-lead ECGs and converted into spatiotemporal any-lead ECG graphs by a learnable convolutional embedder. In these graphs, red edges model intra-lead relationships between nodes corresponding to segments originating from the same lead. Black edges instead model temporal synchrony and spatial heart projections. The resulting spatiotemporal ECG graphs are processed by a Graph Attention Network on top of which a prototype head is placed. During fine-tuning, this head is replaced by a dataset/task specific head and 12-lead ECGs can be used as before or replaced by less-than-12-lead recordings.

Let an any-lead ECG be represented as  $X \in R^{Q \times T}$ , where  $Q \geq 1$  is the number of available leads and  $T$  the number of temporal samples. Each individual lead  $X_i \in R^T$  can be first expressed as the concatenation of  $S$  contiguous ECG segments  $X_i = ||_{j=1}^S x_{ij}$ . By means of a feature extractor  $\phi: R^{T/S} \rightarrow R^D$ , each segment  $x_{ij}$  can be described by a  $D$ -dimensional feature descriptor such that  $h_{ij} = \phi(x_{ij})$ . The sequence of feature descriptors for lead  $i$  is then  $H_i = [h_{i1}, \dots, h_{iS}] \in R^{S \times D}$ . The feature descriptors from all channels are aggregated to form the set  $H = \{H_i: H_i = \phi(X_i), i \in [1, Q]\}$ . The aggregate set  $H$  is used to fit a prototyping model  $\mathcal{K}: R^D \rightarrow [C]$ , where  $[C]$  is a  $C$ -class categorical variable that assigns each segment feature descriptor  $h_{ij}$  to one of the  $C$  learnt prototypes. The prototype sequence for an individual lead is  $\tilde{Y}_i = [\tilde{y}_{i1}, \dots, \tilde{y}_{iS}]$  and  $\tilde{Y} = \{\tilde{Y}_i \in [C]^S: i \in [1, Q]\}$  is the set of all prototype assignments. This lead-preserving segmentation and prototyping generates segment-level assignments maintaining the original multi-lead structure and synchrony of the ECG. Consequently, the learnt codebook captures patterns that remain consistent across arbitrary lead combinations, laying the foundation for any-lead ECG representation learning

### 5.2.2 Any-lead ECG Representation Learning via Masked Node Modelling

While sequence-based models such as HuBERT-ECG can process variable-length inputs through padding and attention masking, they assume a single, linearly ordered sequence. Multi-lead ECGs, however, are inherently multi-relational: each lead captures a temporally synchronized but spatially distinct view of cardiac activity. Flattening these channels into a single sequence not only erases explicit inter-lead structure but also imposes an artificial ordering across leads, forcing the model to

infer spatial dependencies implicitly. Moreover, padding to a fixed number of leads increases memory consumption and misrepresents missing leads, which correspond to absent spatial entities rather than missing temporal frames. To address this structural mismatch, we reformulate the pre-training task over ECG graphs, where nodes correspond to local lead segments and edges encode both intra-lead temporal dependencies and inter-lead synchrony and spatiality. Since 12-lead ECGs are the most widely used clinical configuration, we start from this setup to pre-train an any-lead ECG model. Specifically, for each 12-lead ECG, we sample a random subset of leads  $\mathcal{L} \subseteq \{I, II, III, aVL, aVR, aVF, V1, V2, V3, V4, V5, V6\}^{10}$  with cardinality  $L = |\mathcal{L}|$  drawn from a uniform distribution  $L \sim \mathcal{U}(1,12)$ , and construct the corresponding  $L$ -lead ECG  $X_{\mathcal{L}} \in R^{L \times T}$ . This stochastic lead sampling, analogous to random cropping in computer vision or modality masking in multimodal self-supervision [155], acts as a structured augmentation and encourages the model to develop invariance to lead count and configuration. After lead sampling, a convolutional embedder  $\psi$  embeds each retained lead  $X_i \in \mathcal{L}$  independently into  $S$  local temporal embeddings  $e_{ij}$ :  $\psi(X_i) = E_i = \{e_{ij} \in R^F, \forall j \in [1, S]\}$ . Each sequence of embeddings from the same lead forms a sub-graph with node feature matrix  $E_i \in R^{S \times F}$ , and the embeddings from all sampled leads are concatenated to produce the complete node feature matrix  $E \in R^{L \cdot S \times F}$ . The corresponding graph is denoted as  $\mathcal{G}_{\mathcal{L}}(\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  is the set of graph nodes taking values from  $E$  and  $\mathcal{E}$  is the set of edges encoding temporal, synchrony and spatial relations (see section 5.2.3).

Pre-training proceeds via Masked Node Modelling (MNM), where, unlike HuBERT-ECG, which applies masking over the entire flattened 12-lead sequence regardless of lead boundaries, we apply masking *independently within each individual lead sub-graph*, ensuring that a fixed proportion  $p_{\text{MASK}}$  of nodes is masked *per lead*. This guarantees that all leads contribute masked tokens, avoiding cases where leads may be entirely unmasked. The union of masked nodes, the embedding of which is replaced by a learnable embedding  $e_{\text{MASK}}$ , defines the overall masked node set  $M \subset \mathcal{V}$ . The resulting masked node feature matrix  $E'$  is then consumed by a GNN-based encoder  $\phi$  to produce a new node-level output representation  $\phi(E', \mathcal{E}) = O \in R^{L \cdot S \times F'}$ . For each output node, a probability distribution over the  $C$  codebook entries discovered by the prototyping model  $\mathcal{K}$  is computed:  $p(\widetilde{y}_{ij}^E E', i, j)$ ,  $\forall i \in [1, L], \forall j \in [1, S]$  and used in a standard masked cross-entropy objective:  $\mathcal{L}_{\text{MNM}} = \sum_{v_{ij} \in M} \log(p(\widetilde{y}_{ij}^E E', i, j))$ . Additionally, one can also explore multi-task pre-training utilising multiple codebooks at once to learn useful and complementary representations at varying granularities. The general-case loss objective becomes:  $\mathcal{L}_{\text{MNM}}^{\Gamma} = \sum_{\gamma=1}^{\Gamma} \sum_{v_{ij} \in M} \log(p^{\gamma}(\widetilde{y}_{ij}^{\gamma} E', i, j))$ , where  $\Gamma$  denotes the total number of codebooks and  $p^{\gamma}$  is the softmax distribution over the entries of the  $\gamma$ -th codebook. Through this pre-training scheme, the model learns to infer masked prototypes from both intra- and inter-lead context while continually exposed to randomly sampled lead configurations. The results in a topologically aware, lead-invariant representation space that preserves ECG relational structure and generalizes to arbitrary combinations and counts of leads without retraining or architectural changes.

### 5.2.3 The spatiotemporal topology

Modelling multi-channel time series such as multi-lead ECGs as graphs provides a natural way to encode relationships among signal entities, beyond what is available to sequential models [156]. Prior ECG graph constructions have mostly followed two extremes. Visibility-graph approaches [157], [158] build point-wise graphs over beats or short sequences, where each sample is a node with scalar voltage. While effective on short segments, they become prohibitively large on long recordings and tend to capture highly local, isolated dependencies. At the other extreme, lead-as-node graphs

<sup>10</sup> These are the names of the six limb leads and the six chest leads

represent each lead as a single node and define edges from known physiological lead relations [159], [160]. These scale well and incorporate spatial priors, but are too coarse to express intra-lead temporal dependencies such as relations among waves within the same cardiac cycle and distant beats. Our goal is to retain physiological grounding and scalability while making intra-lead temporal structure explicit.

Building on the node definitions and lead-wise subgraphs introduced above, we now specify the edge set  $\mathcal{E}$  that defines our spatiotemporal topology. Edges are partitioned into intra-lead temporal relations and inter-lead synchrony/spatial relations:  $\mathcal{E} = \mathcal{E}_{intra} \cup \mathcal{E}_{inter}$ . Specifically,  $\mathcal{E}_{intra} = \bigcup_{i=1}^L \{(v_{ij}, v_{ik}) : j \neq k, j, k \in [1, S]\}$ , which means that, for each sampled lead  $i \in [1, L]$ , we connect all nodes that belongs to the same lead, yielding a complete subgraph. Edges are undirected, thus message passing is bidirectional across the entire lead duration and trajectory. This design deliberately avoids restricting connectivity to local neighbourhoods: it allows the encoder to directly model both local (within-beat) dependencies among adjacent segments and global (across-beats/long-range) regularities along the full ECG duration. Physiologically, such dense intra-lead coupling reflects that ECG morphology is governed by recurrent cardiac cycles and rate-dependent “memory” effects—e.g., repolarization hysteresis and restitution—where the interpretation of any temporal segment benefits from context distributed across preceding and subsequent beats. Next, because all leads are recorded simultaneously, we connect nodes aligned in time across distinct lead:  $\mathcal{E}_{inter} = \{(v_{ij}, v_{i',j}) : i \neq i', i, i' \in [1, L], j \in [1, S]\}$ . These edges encode temporal synchronisation by construction and serve even as a spatial prior, since different leads are simultaneous projections of the same underlying cardiac electrical field. In practice, they enable the encoder to exploit cross-lead co-manifestations of events (e.g. coordinated ST-segment shifts or dyssynchronous activation patterns) without flattening leads into an artificial 1D order.

With nodes ordered by concatenating the  $S$  nodes of each lead (as in  $E \in R^{L \cdot S \times F}$ ), the resulting  $L$ -lead ECG adjacency matrix  $A \in \{0,1\}^{LS \times LS}$  has the following symmetric block form:

$$A = \begin{bmatrix} \mathbf{1} & \mathbf{I} & \cdots & \mathbf{I} \\ \mathbf{I} & \mathbf{1} & \cdots & \mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{I} & \cdots & \mathbf{1} \end{bmatrix}$$

Where each diagonal block  $\mathbf{1} \in \{1\}^{S \times S}$  is an all-ones matrix encoding the intra-lead complete subgraph, and each off-diagonal block is the identity  $\mathbf{I}$  of size  $S$ , enforcing time-aligned inter-lead coupling. For clarity,  $\mathbf{1}$  indicates within-lead full connectivity, whereas  $\mathbf{I}$  indicates cross-lead synchronisation at matching temporal indices.

This topology mirrors how ECGs are generated and interpreted. The complete intra-lead subgraphs allow the model to relate any two temporal segments within the same lead, capturing both short-range wave-to-wave structure within a beat and long-range regularities across beats over the full recording; their bidirectionality is consistent with cyclic cardiac conduction and electrophysiological “memory” phenomena, where morphology is context-dependent across time. Inter-lead identity connections encode the simultaneity of body-surface potentials and their spatially distinct projections, supporting modelling of clinically relevant cross-lead patterns (e.g., localized ischemia or conduction asynchrony). Overall, the spatiotemporal topology preserves native multi-lead synchrony and temporal structure while providing a scalable inductive bias aligned with cardiac electrophysiology.

Ultimately, from a computational point of view, the two edge sets that our spatiotemporal topology induces,  $\mathcal{E}_{intra}$  and  $\mathcal{E}_{inter}$ , contribute to the edge count as follows:  $|\mathcal{E}_{intra}| = L \frac{S(S-1)}{2} = \mathcal{O}(LS^2)$

and  $|\mathcal{E}_{\text{inter}}| = S \frac{L(L-1)}{2} = \mathcal{O}(SL^2)$ . Hence,  $|\mathcal{E}| = \mathcal{O}(LS^2 + SL^2)$ . Since  $L \sim \mathcal{U}(1,12)$ , we can derive best-, worst- and expected-case message passing costs. In the best case  $L = 1$ ,  $|\mathcal{E}| = \mathcal{O}(S^2)$ ; in the worst-case  $L = 12$ ,  $|\mathcal{E}| = \mathcal{O}(12S^2 + 144S)$ . Under uniform sampling,  $\mathbb{E}[|\mathcal{E}|] = \frac{S(S-1)}{2} \mathbb{E}[L] + \frac{S}{2} \mathbb{E}[L(L-1)] = \mathcal{O}(S^2 \mathbb{E}[L] + S \mathbb{E}[L^2])$ . This means that lead sampling reduces both intra-lead and inter-lead edges in expectation, by factors  $\frac{\mathbb{E}[|\mathcal{E}_{\text{intra}}|]}{|\mathcal{E}_{\text{intra}}|_{12}} = \mathbb{E}[L]/12$  and  $\frac{\mathbb{E}[|\mathcal{E}_{\text{inter}}|]}{|\mathcal{E}_{\text{inter}}|_{12}} = \mathbb{E}[L(L-1)]/(12 \cdot 11)$ , respectively. In addition, Transformer-based models such as HuBERT-ECG [45] and many other models from the literature (e.g. [110], [120], [148]) implicitly operate on a complete token graph over the same  $N = L \cdot S$  nodes, requiring all-pairs self-attention with per-layer cost  $\mathcal{O}(N^2) = \mathcal{O}(L^2 S^2)$ , whereas our spatiotemporal message passing costs is  $\mathcal{O}(LS^2 + SL^2)$  (up to the feature dimension  $F$ ). Their cost ratio  $\frac{LS^2 + SL^2}{L^2 S^2} = \frac{1}{L} + \frac{1}{S}$  shows that our informed adjacency yields asymptotically cheaper per-layer computation than complete-graph attention in realistic multi-lead, multi-segment regime (i.e.  $L, S > 1$ ), with savings increasing as ECG duration ( $S$ ) and lead count ( $L$ ) grow.

### 5.2.4 Transfer-guided Codebook Refinement

Following Hsu et al. [32] and Coppola et al. [45], we adopt an iterative codebook refinement scheme to update prototype supervision and continue pre-training. After the first pre-training stage, we freeze all model parameters and extract unmasked latent node representations for fitting a new prototyping model  $\mathcal{K}^{(2)}$ . This replaces the original codebook and provided refined targets for a second MNM stage.

Since our encoder is GNN-based, node representations are repeatedly update by message passing and therefore tend to smooth over network depth and training time. Also, the small diameter of our topology accelerates mixing across nodes. While this facilitates MNM by allowing masked nodes to be inferred from broad context, it may also shrink the variance of latent representations over time. To avoid refining the codebook on overly smoothed latents, we select the checkpoint used for latent extraction via an early-stopping-by-transfer criterion: we save checkpoints during stage-1, evaluate their transfer quality on a downstream proxy task, and choose the earliest checkpoint  $t^*$  beyond which further pre-training yields negligible gains. We use  $\varphi_{t^*}^{(1)}$  to extract latents and build the refined codebook.

Finally, because refinement operates on topology-shaped latents, the refined targets can still reflect a smooth manifold. To limit additional smoothing and keep representations expressive during stage-2 MNM, we consider randomly dropping intra-lead edges with probability  $p_{\text{drop}}$  [161]. This weakens mixing during stage-2 message passing and acts as a simple variance-preserving regularisation while training against the refined codebook. We do not consider the edge dropping during the first pre-training stage to maximize contextual information and stabilize MNM under the initial codebook.

### 5.2.5 Architecture

The first component of our architecture is a shallow convolutional embedder that maps each ECG lead into a sequence of node feature vectors. Specifically, the embedder processes  $L$ -lead ECGs sampled at 100 Hz and outputs node representations at a temporal resolution of 0.25 seconds per node. This framing yields  $S = 20$  nodes per lead and captures shorter, wave-based segments that are more granular than those produced by HuBERT-ECG. From these representations, we construct the spatiotemporal ECG graph. During pre-training, we randomly mask nodes with probability  $p_{\text{MASK}}$  meaning that  $p_{\text{MASK}}$  of nodes in each lead subgraph are replaced by a special learnable node feature

vector. After masking, the resulting graph is processed by a Graph Attention Network (GAT) encoder [162]. Through message passing layers, the encoder generates node representations that are projected by linear layer and softmax into prototype class scores. We observe no benefit from employing learnable lookup embedding matrices with similarity measures, which also require careful tuning of the temperature that scales the logits. After pre-training, we attach a randomly initialised linear layer followed by a task/dataset specific activation (e.g. softmax for multi-class problems, sigmoid for multi-label/multi-task datasets) and fine-tune the resulting new model on downstream data. Table 5.1 summarises our model architecture.

|                                       |                  |                  |
|---------------------------------------|------------------|------------------|
| 1D Convolutional embedder             | strides          | 3, 2, 2, 2       |
|                                       | kernels width    | 10, 3, 3, 2      |
|                                       | channels         | 768              |
|                                       | activation       | GeLU             |
|                                       | normalization    | LayerNorm        |
| GAT encoder                           | layers           | 2                |
|                                       | hidden channels  | 768              |
|                                       | normalization    | LayerNorm        |
|                                       | attention heads  | 8                |
|                                       | activation       | GeLU             |
|                                       | Dropout          | 0.1              |
| Prototype classification head         | Input dimension  | 768              |
|                                       | Output dimension | $C$              |
|                                       | Bias             | False            |
| Downstream task/dataset-specific head | Input dimension  | 768              |
|                                       | Output dimension | # classes/labels |
| Number of parameters                  |                  | ~7M              |

**Table 5.1:** Summary of model architecture.

## 5.3 Experimental details

### 5.3.1 Data and preprocessing

To learn generalisable any-lead ECG representations across arbitrary lead configurations and enable any-lead inference, we design a strategy that consists in randomly sampling leads from the 12 *standard* derivations during training. This choice leverages the broadly accepted and empirically proved information redundancy within 12-lead ECG signals [163] and aims at distilling global 12-lead knowledge into any-lead representation learning. We utilise a highly diverse collection of 12-lead ECG datasets encompassing a wide range of cardiovascular conditions, even co-occurring. These sources are aggregated from five countries and comprise both labelled and unlabelled partitions, totalling 13 distinct datasets for initial representation learning: Ribeiro (CODE) [126], IKEM[164], PTB-XL[129], CPSC and CPSC-Extra [127], Chapman [131], Georgia [130], Hefei, MIMIC-IV[134], Ningbo [132], PTB [128], SaMi-Trop [135], and SPH [133]. Notably, we surpass the pre-training size and diversity of HuBERT-ECG with the addition of IKEM. For supervised fine-tuning and evaluation, following Coppola et al. [45], we retain all previously listed datasets except the unlabelled sets of Ribeiro, IKEM, and MIMIC-IV. We also consider Cardio-Learning as defined in HuBERT-ECG’s paper as additional, cardiologist-level benchmark. To further strengthen evaluation coverage, we incorporate two additional sources: the “clinical data” introduced by Tian et al. [116], and the Pediatric dataset curated by Tan et al.[165]. Finally, to assess the transferability of our model to ECG scenarios with substantially reduced lead count, including ECGs from wearable devices, we

include OffTest<sup>11</sup> [166] and MITHSDB [167]. We plan to include more wearable ECG datasets and reduced-lead sources in the near future. An overview of the datasets used in this work is provided in Table 5.2. We note that test sets used for evaluations are held out and not used for training at any stage. For data preprocessing, we follow consolidated practices and apply rescaling into  $[-1, 1]$  range and band-pass filtering ( $[0.05, 47]$  Hz) [45], [144], [145], [146].

| Dataset              | #Conditions         | #ECGs   | #Leads | SSL<br>Pre-training | SFT &<br>Evaluation | Country of<br>Origin |
|----------------------|---------------------|---------|--------|---------------------|---------------------|----------------------|
| Ribeiro – Unlabelled | Unknown             | 5877604 | 12     | ✓                   | ✗                   | Brazil               |
| Ribeiro – Labelled   | 6                   | 2322465 | 12     | ✓                   | ✓                   | Brazil               |
| IKEM                 | Unknown             | 98130   | 12     | ✓                   | ✗                   | Czech Republic       |
| PTB-XL               | (71, 44, 23, 5, 12) | 21837   | 12     | ✓                   | ✓                   | Germany              |
| CPSC                 | 9                   | 6878    | 12     | ✓                   | ✓                   | China                |
| CPSC-Extra           | 52                  | 3453    | 12     | ✓                   | ✓                   | China                |
| Chapman              | 52                  | 10246   | 12     | ✓                   | ✓                   | China                |
| Georgia              | 65                  | 10344   | 12     | ✓                   | ✓                   | United States        |
| Hefei                | 29                  | 20036   | 12     | ✓                   | ✓                   | China                |
| MIMIC-IV             | Unknown             | 800036  | 12     | ✓                   | ✗                   | United States        |
| Ningbo               | 76                  | 34905   | 12     | ✓                   | ✓                   | China                |
| PTB                  | 17                  | 515     | 12     | ✓                   | ✓                   | Germany              |
| SaMi-Trop            | 2                   | 1632    | 12     | ✓                   | ✓                   | Brazil               |
| SPH                  | 44                  | 25770   | 12     | ✓                   | ✓                   | China                |
| Cardio-Learning (*)  | 164                 | 2458081 | 12     | ✓                   | ✓                   | Miscellaneous        |
| Clinical data        | 10                  | 1784    | 12     | ✗                   | ✓                   | China                |
| Pediatric            | 19                  | 14190   | 9, 12  | ✗                   | ✓                   | China                |
| OffSet               | 60                  | 7000    | 12     | ✗                   | ✓                   | China                |
| MITHSDB              | (2, 5)              | 409     | 1      | ✗                   | ✓                   | United States        |

**Table 5.2:** Overview of the datasets used in this study, detailing number of conditions/tasks or classes, number of recordings and available leads, participation in pre-training, fine-tuning & evaluation, and country of origin. “\*” indicates that a specific dataset has been produced artificially merging other sources.

### 5.3.2 Unsupervised codebook discovery

As in HuBERT-ECG, our pre-training is organized into two stages, each preceded by an unsupervised codebook discovery. Before stage 1, we build an initial codebook from hand-crafted descriptors. Specifically, for each node (i.e., ECG segment) we extract a 39-dimensional MFCC feature vector [121]. Unlike HuBERT-ECG, which clusters descriptors from a flattened 12-lead sequence, we compute MFCCs independently within each lead. This avoids introducing spurious feature discontinuities at artificial lead junctions and preserves the native multi-lead structure during prototype discovery. We fit a codebook by clustering all MFCC descriptors with k-means++. The codebook size  $C$  is selected by first restricting candidate values via the elbow method, and then validating a small set of candidates through ablation. Concretely, we consider  $C \in \{50, 70, 100\}$  and all possible combinations, and retain the best-performing configuration (stage-1 uses  $C = 50$ ). After

<sup>11</sup> Although OffSet comprises 12-lead ECGs, recordings are acquired with a wearable device.

stage-1 pre-training, we refine the codebook by clustering latent node representations extracted from the transfer-optimal checkpoint  $t^*$ , following the iterative refinement protocol described in Section 5.2.4. This yields an updated prototyping model  $\mathcal{K}^{(2)}$  with refined vocabulary size  $C'$ , the exact value of which will be found in the near future. Because the full set of descriptors (MFCCs in stage 1 and latents in stage 2) does not fit in memory, we employ a batched version of  $k$ -means++, using 20 random restarts to improve centroid initialization and stability.

### 5.3.3 Self-supervised Pre-training

Similarly to HuBERT-ECG, we split the unlabelled corpus into 90% for pre-training and 10% for internal validation. Both partitions include ECGs from IKEM [164] to increase geographic and recording-condition diversity. Stage-1 pre-training is performed for up to 150k steps with an effective batch size of 512, masking ratio  $p_{\text{MASK}} = 40\%$  (applied per-lead), learning rate  $1e-3$  with linear warm-up over 8% of steps, and initial weight decay of 0.01 and dropout of 0.1. Both regularisation terms are adapted via dynamic regularisation (penalty-count=1, validation-frequency=5000 steps) [45]. To select the stage-1 checkpoint used for latent extraction and codebook refinement, we adopt the early-stopping-by-transfer criterion described in Section 5.2.4: among the saved stage-1 checkpoints, we choose the earliest checkpoint  $t^*$  beyond which downstream linear-evaluation improvements  $\Delta$  become negligible (defined as  $\Delta < 10^{-3}$  AUROC). This transfer-aware selection is motivated by the tendency of GNNs on small-diameter graphs to produce increasingly smoothed representations over training [168]. Figures 5.3 and 5.4 supports this choice by relating checkpoint progression to neighbourhood similarity, latent diversity and transferability. During stage 2, we then plan to resume pre-training using the model checkpoint after 75k steps and the same hyper-parameters outlined above, but dropping edges of the spatiotemporal topology with a probability  $p_{\text{drop}}$  found via ablation.

### 5.3.4 Supervised fine-tuning and downstream evaluation

To evaluate our GAT-based any-lead model and assess its downstream utility in both clinically standard 12-lead settings and non-standard ones (e.g., wearable ECGs), we fine-tune the stage-2 pre-trained encoder on each dataset listed in Table 5.2. By benchmarking transfer from standard clinical recordings to reduced-lead or wearable domains, we specifically test whether representations learned from abundant 12-lead data remain effective when lead count and placement change. This is practically relevant because wearable ECG configurations are inherently constrained in lead coverage and large, well-labelled wearable datasets are harder to obtain. This would allow leveraging clinically standard pretrained representation as a natural way to improve performance in data-limited settings [169], [170].

For fair comparison, we adopt the same train/validation/test splits used by HuBERT-ECG<sup>12</sup> where available, and we reconstruct Cardio-Learning following the original protocol. For newly introduced datasets, we create a held-out test set comprising 10% of available ECGs and use the remaining 90% for training and internal validation.

Fine-tuning is performed by attaching a randomly initialized linear head to the encoder and updating all model weights, including the convolutional embedder. Given the small parameter budget ( $\sim 7\text{M}$ ), full fine-tuning allows maximal adaptation to each downstream distribution. During fine-tuning we disable MNM masking but retain stochastic lead sampling so that the model continues to learn lead-invariant representations aligned with task labels. We additionally apply temporally aligned random

---

<sup>12</sup> [HuBERT-ECG/reproducibility at master · Edoar-do/HuBERT-ECG](#)

cropping as data augmentation within each available lead and reapply this procedure as test-time augmentation.

At evaluation, we always test on the corresponding held-out partition while randomly sampling leads once per ECG (i.e., without aggregating predictions across multiple lead draws, so not as test-time augmentation). On Cardio-Learning, because it contains 164 tasks, we conduct two extra analyses: (i) we evaluate the same fine-tuned model under exactly  $L = 1, 2, \dots, 12$  randomly sampled leads to quantify how performance scales with lead count; and (ii) we plan to repeat this while fixing the specific lead(s) sampled, enabling a task-wise, disease-wise notion of lead importance analogous to feature-importance analyses. This type of analysis greatly expands the set possible conclusions that can be drawn from the 2021 Physionet Challenge [125]

We report performance using AUROC throughout, in line with prior literature [45], [113], [120], [146], [147], [148].

## 5.4 Preliminary Results

In this section we report a set of preliminary, but already highly encouraging, results. We first present an extensive ablation study about the role of topology and encoder choice (Section 5.4.1), comparing our solution fairly with HuBERT-ECG BASE after the first pre-training iteration. Second, we explore the roles of temporal positional encoding [119], a pivotal ingredient in Transformer-based models to provide order among tokens, and learnable lead embeddings, which can enforce lead identity among graph nodes (Section 5.4.2). Third, we identify the optimal per-lead masking ratio (Section 5.4.3). Next, we investigate the effects of codebook size and multi-task pre-training (Section 5.4.3), which proved advantageous for both HuBERT [32] and HuBERT-ECG. Finally, we analyse node representation quality through the lenses of oversmoothing to guide the second stage of pre-training.

### 5.4.1 Ablation on graph encoders and ECG topologies

This ablation has two goals: (i) identify the most effective GNN encoder for our masked node modelling objective, and (ii) validate the proposed spatiotemporal topology against standard alternatives. In particular, we compare our topology with: (a) a fully connected graph, considered both with fixed edges (e.g., Graph Isomorphism Networks) and with learned edge weights via attention/restriction mechanisms (e.g., transformer-like and sheaf-based models), and (b) the absence of any topology, i.e., processing ECG nodes independently as an MLP baseline. For clarity and fairness, all models in this first study are constrained to a similar parameter budget ( $\sim 4\text{M}$  parameters). This keeps compute manageable while ensuring that differences in performance can be attributed primarily to architectural inductive bias and graph structure rather than model capacity. We evaluate a diverse set of graph encoders representative of current practice in graph representation learning [171]: Multi-Layer Perceptron (MLP), Graph Transformer [172], Graph Isomorphism Network (GIN) [173], GAT [162], Sheaf Neural Network [174], Sheaf Attention (v1) [175], and Sheaf Attention v2 (our implementation using GATv2-style [176] attention). These models span fixed-topology message passing, attention-based graph learning, and sheaf-based generalizations, providing a broad testbed for assessing both encoder suitability and topology design. Results are reported in Table 5.3.

After pre-training all models for 60k steps, we linearly evaluate them on Ribeiro-dev [45] to disentangle the effects of the pre-trained representations from the downstream learning capabilities of the models. Unsurprisingly, we note that using solely node features is not enough for effective any-lead ECG representation learning, likely because utilising no topology does not allow to model relationships among different segments. Second, we find that our topology consistently delivers

stronger performance than those from the fully connected one. This result underlie the value of imposing a physiologically informed, sparse inductive bias: by explicitly encoding intra-lead temporal relations and inter-lead synchrony and spatiality, the graph promotes information exchange along clinically meaningful axes rather than dispersing attention uniformly across all nodes. Ultimately, we find the GAT to be the best encoder, even outperforming sheaf-based generalisation that are theoretically more expressive. Most interestingly, we highlight that this solid performance (0.937) surpasses that from HuBERT-ECG BASE (first iteration), a model that was pre-trained for 20k extra steps, with *~90 million additional parameters* and the full lead set.

|                       | Spatiotemporal | Fully Connected | No topology |
|-----------------------|----------------|-----------------|-------------|
| GAT                   | <b>0.937</b>   | 0.854           | /           |
| Graph Transformer     | 0.907          | 0.853           | /           |
| GIN                   | 0.900          | 0.820           | /           |
| Sheaf                 | 0.808          | 0.799           | /           |
| Sheaf Attention v1    | 0.811          | 0.752           | /           |
| Sheaf Attention v2    | 0.822          | 0.700           | /           |
| MLP                   | /              | /               | 0.747       |
| HuBERT-ECG BASE it. 1 | 0.933          |                 |             |

**Table 5.3:** Ablation study over GNN encoder architectures and graph topologies. Performance is reported after linear evaluation in terms of AUROC. HuBERT-ECG BASE it. 1 is reported as baseline.

Behind this stronger performance lie our design choices: (i) preserving lead identity and temporal synchrony through the spatiotemporal graph, which injects a physiology-aligned sparse prior so message passing follows clinically meaningful inter-lead and intra-lead relations instead of complete-graph mixing; (ii) lead-wise codebook discovery, which fits prototypes within each lead’s native distribution and avoids artefacts introduced by flattening ECG derivations; (iii) lead-balanced masking, which applies a fixed masking ratio per lead and prevents shortcut solutions based on a subset of unmasked leads, forcing the encoder to use both intra-lead and cross-lead context; and (iv) stochastic lead sampling, which exposes the model to arbitrary lead subsets during MNM and promotes lead-invariant representations that transfer better even to standard 12-lead evaluation. Overall, these results show that a physiology-grounded inductive bias can outweigh brute-force model capacity in early ECG representation learning.

Next, we study the effect of encoder depth and width on downstream performance (Table 5.4). With efficiency and scalability in mind, we sweep over the number of GAT layers and hidden channels while keeping the pre-training protocol fixed. We observe that increasing width yields larger gains than increasing depth: a 2-layer, 768-channel GAT improves downstream AUROC by about 2 points over the 3-layer 512-channel model presented above. This suggests that richer node features are more beneficial than deeper message passing, likely because two layers already provide global context exchange in a graph with diameter 2.

|        | Hidden Channels |       |              |
|--------|-----------------|-------|--------------|
| Layers | 256             | 512   | 768          |
| 2      | 0.924           | 0.934 | <b>0.951</b> |
| 3      | 0.928           | 0.937 | 0.936        |
| 4      | 0.923           | 0.929 | 0.930        |

**Table 5.4:** Exploration on the number of layers and input/output channels in the GAT encoder.

### 5.4.2 Ablation on temporal encodings and lead embeddings

GNNs (including GATs) are permutation-invariant over node order: message passing aggregates a multiset of neighbour features and does not depend on any imposed index ordering. Thus, if nodes correspond to temporal segments, explicit positional signals could in principle help disambiguate their order. We test this by adding sinusoidal temporal positional encodings (PE) to node features before message passing, as in the original Transformer[119]. Sinusoidal encodings are a strong, low-parameter default for fixed-length sequences and allow the model to recover relative offsets via linear operations. We choose sinusoidal PEs over more elaborate schemes (e.g., T5-style [108] or Rotary Positional Encodings [177]) for two practical reasons: (i) our ECG inputs have a known, fixed duration of 5 seconds, so extrapolation to longer lengths, one key advantage of more recent approaches, is not required; (ii) we want to isolate the effect of explicit temporal injection itself, rather than coupling it to architecture-specific mechanisms developed for long-context NLP.

We also test learnable lead embeddings to explicitly encode lead identity into node features. We evaluate two downstream evaluation settings: when such embeddings remain frozen (i.e., pure linear evaluation), and when they are fine-tuned jointly with the linear classification head (i.e. a partial model fine-tuning). This distinction is introduced to test whether lead embeddings are useful only when adapted to the downstream distribution, rather than assuming a priori that they should be frozen or tuned.

Results in Table 5.5 show that neither temporal sinusoidal encodings nor lead embeddings improve transfer; all variants underperform the baseline without additional encodings. This indicates that the graph inductive bias already supplies sufficient temporal and lead information through complete intra-lead connectivity and time-aligned inter-lead synchrony edges, making extra encodings redundant. Notably, this suggests that for long 5-second ECGs, effective representations can emerge without explicit temporal PEs—an outcome that is both computationally favourable and conceptually non-trivial given that ECGs are strongly temporal signals and we are not merely considering local isolated beats.

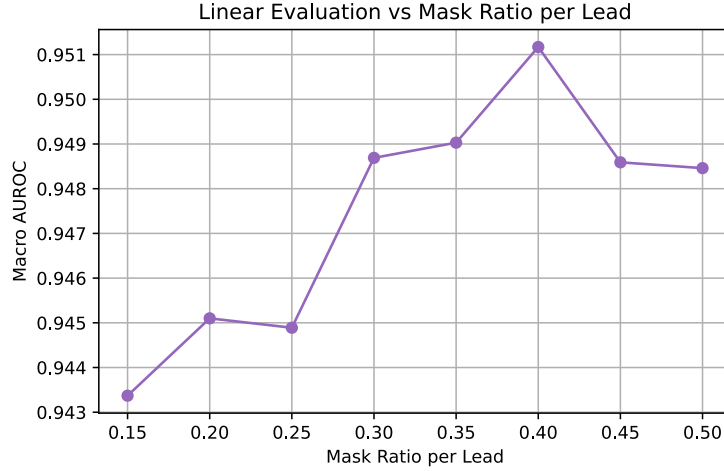
|     | None         | Lead* | Lead  | Temporal | Lead* + Temporal | Lead + Temporal |
|-----|--------------|-------|-------|----------|------------------|-----------------|
| GAT | <b>0.951</b> | 0.918 | 0.933 | 0.931    | 0.916            | 0.930           |

**Table 5.5:** Effects of positional temporal encodings and learnable lead embeddings. “\*” indicates that the learnable lead embeddings remain frozen during the downstream evaluation, meaning that a simple linear evaluation is conducted. If this symbol is not reported, it means lead embeddings are fine-tuned jointly with the linear classification head.

### 5.4.3 Ablation on the masking ratio

Optimising masked modelling objectives, including MNM, requires selecting an appropriate masking ratio  $p_{\text{MASK}}$ . We therefore conduct an ablation study on this hyper-parameter while keeping all other factors fixed. Specifically, we pre-train multiple GAT instances for 60k steps using the same spatiotemporal topology and the best hyper-parameters identified above, without temporal PEs or learnable lead embeddings. We vary only  $p_{\text{MASK}}$  in 5% increments, which in our discretization corresponds to masking  $\pm 1$  node per lead subgraph. Figure 5.2 shows that masking between 30% and 50% nodes per lead yields comparable linear evaluation AUROC, with 40% (8 masked nodes per lead) providing the best result. We thus adopt  $p_{\text{MASK}} = 0.4$  as default ratio. Notably, this is higher than the 33% masking used by HuBERT-ECG on flattened 12-lead sequences. A plausible explanation is that, in our setting, masked nodes can be reconstructed from a more structured and informative context: complete intra-lead connectivity and time-aligned inter-lead edges provide direct

physiological pathways for message passing, allowing the model to tolerate heavier masking without collapsing performance.



**Figure 5.2:** Linear evaluation performance on Ribeiro-dev after pre-training with varying per lead masking ratios.

#### 5.4.4 Investigation on codebook size and multi-task pre-training

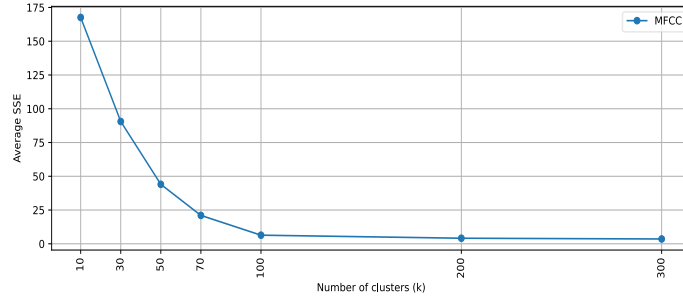
The codebook size  $C$  controls the granularity of the discrete prototypes assigned to ECG nodes. If  $C$  is too small, prototypes become overly coarse and the MNM objective risks collapsing into a trivial prediction task; if  $C$  is too large, prototypes can fragment stable modes in the latent space, injecting label noise and hurting transfer. We therefore tune  $C$  for stage-1. We first apply the elbow method to the SSE curve obtained from clustering MFCC descriptors (Figure 5.3), which indicates a reasonable range between  $C = 50$  and  $C = 100$ . We then pre-train multiple GAT instances under an identical protocol and select the final value based on downstream linear evaluation, finding  $C = 50$  to be optimal (Table 5.5).

We also test multi-task pre-training during stage 1 by predicting multiple codebooks simultaneously (or larger combined vocabularies), since this has improved performance in HuBERT and HuBERT-ECG. In our case, however, multi-tasking provides no benefit: larger or multiple codebooks do not outperform  $C = 50$ , and at best match it (e.g.,  $C = 100, 50$ ). This suggests that our spatiotemporal topology and GAT encoder already induce a strongly denoised and low-variance latent manifold at stage-1, supporting only a limited number of stable clusters. Increasing  $C$  or adding parallel codebooks therefore tends to over-partition these modes without adding informative supervision, yielding redundant or noisy targets rather than richer representations.

| Codebook size – Stage 1 | Lin. Eval AUROC |
|-------------------------|-----------------|
| $C = 50$                | <b>0.951</b>    |
| $C = 70$                | 0.943           |
| $C = 100$               | 0.948           |
| $C = 70, 50$            | 0.946           |
| $C = 100, 50$           | 0.951           |
| $C = 100, 70$           | 0.949           |
| $C = 100, 70, 50$       | 0.947           |

**Table 5.6:** Linear GAT models pre-trained on codebooks of varying size.

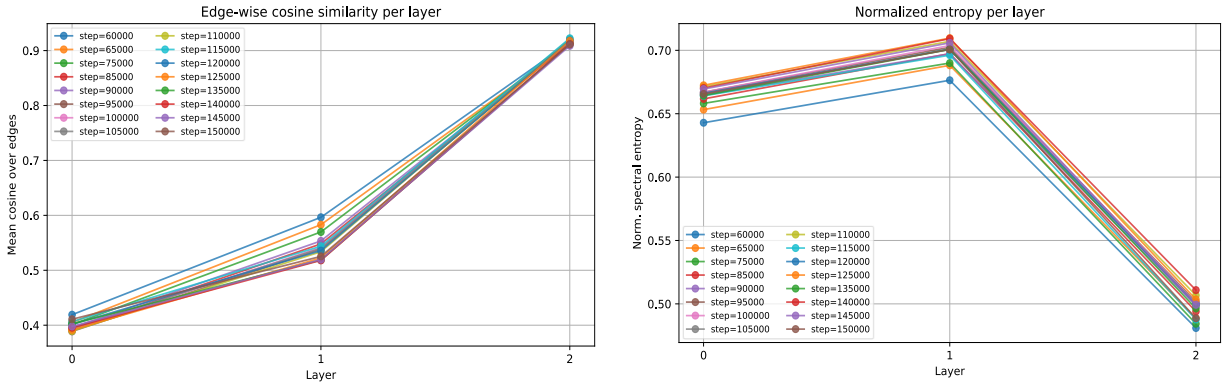
evaluation performance of



**Figure 5.3:** MFCC-based clustering quality in terms of SSE.

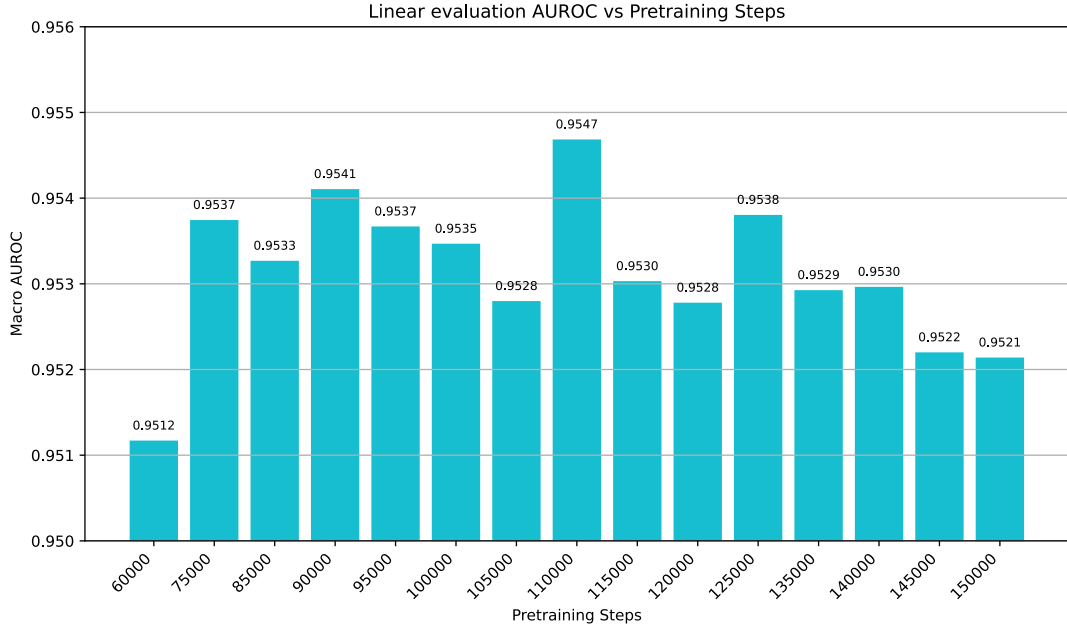
#### 5.4.5 Analysing cluster quality across encoding layers

To guide iterative codebook refinement, we analyse how the quality and diversity of node representations evolve across encoding layers and training checkpoints. In particular, we track two neighbourhood-faithful oversmoothing indicators over all stage-1 checkpoints from 60k to 150k steps: edge-wise cosine similarity per layer and normalized spectral entropy per layer (Figure 5.4-left and Figure 5.4-right). These measures quantify, respectively, how much representational variance remains at each layer and how similar neighbouring nodes become under message passing. Since the refined codebook is obtained by clustering latent node vectors, preserving sufficient latent diversity is essential for stable and informative prototypes.



**Figure 5.4:** (Left) Edge-wise cosine similarity across GAT layers and model checkpoints. (Right) Normalized entropy across GAT layers and model checkpoints. Both graphs provide insights into model’s risk of oversmoothing node representations and which encoding layer is the most suitable to extract latent representation for codebook refinement.

Across all stage-1 checkpoints (60k–150k), the layer-wise diagnostics remain consistent: normalised entropy stays high at the input and after the first GAT layer, then drops sharply at layer 2, while edge-wise cosine similarity rises mildly at layer 1 and jumps to just above 0.9 at layer 2. This points to the final layer as a potential oversmoothing bottleneck, i.e., a stage where neighbourhood representations tend to collapse into a low-variance manifold. Given this behaviour, the relevant selection problem is not “which checkpoint minimizes MNM loss,” but rather “which checkpoint maximizes transfer while still preserving a meaningful latent space for refinement”. We therefore track downstream linear evaluation performance across checkpoints and observe that transfer improves early and then plateaus (Figure 5.5) even as MNM continues to decrease. Interpreting this plateau together with the oversmoothing theory [168], we select the refinement checkpoint via early-stopping-by-transfer, choosing the earliest checkpoint delivering a non-negligible AUROC improvement. This yields  $t^* = 75\text{k}$ . This analysis also sheds light onto which encoding layer is most suited for latent extraction. In our case, layer 1 of the  $t^*$  checkpoint retains substantially higher entropy while already encoding spatiotemporal context, representing the best candidate for extraction and drive the second pre-training stage.



**Figure 5.5:** Linear evaluation performance of multiple model checkpoint pre-training for increasing steps.

A practical advantage of our selection approach is that it provides a cheap, theory-driven criterion for choosing the extraction layer. In HuBERT-ECG, the layer selection had to be narrowed using prior NLP heuristics [178] and then finalised with clustering-quality proxies. Here, instead, oversmoothing theory [168], tracked through entropy and neighbourhood cosine, directly indicates where latent diversity collapses, yielding an objective, architecture-agnostic guide to extraction. While our current encoder has only three layers, the same diagnostics can scale to deeper GNNs at essentially no extra cost, offering a general and computation-light alternative to exhaustive layer sweeps or ad-hoc clustering metrics.

## 5.5 Discussion and future steps

In this chapter we introduce a tiny ( $\sim 7\text{M}$  parameters) any-lead ECG foundation model built around a set of strong, physiologically informed inductive biases. The first is the spatiotemporal topology: each

lead is represented as a complete intra-lead subgraph of temporal segments, while leads are linked through time-aligned inter-lead edges. This yields sparse, clinically meaningful message-passing pathways that preserve both intra-lead temporal coherence across the full 5-second recording and inter-lead synchrony/spatial coupling, avoiding the unstructured complete-graph mixing typical of Transformer-based models. The second bias is stochastic lead sampling during pre-training, which acts as a structured augmentation: the model must solve the task under arbitrary lead subsets, promoting lead invariance and forcing maximal use of whichever projections are available. The third is Masked Node Modelling (MNM), which further encourages extracting information from context within leads and across synchronized leads. Together, these choices define a compact yet highly capable pre-training recipe that already surpasses much larger baselines (e.g., HuBERT-ECG BASE, ~90M fewer parameters), reinforcing the takeaway that the right inductive bias can beat brute-force capacity in early ECG representation learning.

The completed project aims to deliver an efficient and scalable foundation model that supports any-lead ECG analysis across the clinical and wearable spectrum. Practically, this would enable strong transfer from abundant 12-lead hospital data to less-than-12-lead domains (including wearables), where labelled datasets are smaller, noisier, and less diverse, thereby widening the scope of accurate wearable diagnosis beyond today’s limited arrhythmia-focused solutions [170]. Beyond performance, the any-lead setup unlocks new clinical insights: by systematically evaluating downstream models under controlled lead counts and fixed specific leads, we aim to quantify how many derivations are necessary for reliable detection of specific conditions and which individual derivations carry the most diagnostic information per pathology. Ultimately, the most ambitious target is to approach the accuracy of standard 12-lead models while retaining the flexibility and efficiency required for real-world arbitrary-lead scenarios.

To reach these goals, the next phases of work will aim to (i) complete the second pre-training stage with latent-refined prototypes; (ii) perform full fine-tuning across the planned suite of clinical and wearable datasets to quantify transfer beyond linear evaluation; (iii) finalise the lead-importance and lead-count sensitivity analyses on Cardio-Learning and related benchmarks. These steps will consolidate the current evidence into a complete any-lead foundation model and clarify its practical and clinical impact.

## Chapter 6 - Conclusions and Future Works

This thesis advances a focused technical route to trustworthy medical AI by developing methods that remain clinically meaningful, robust to real deployment variability, and scalable when supervision and compute is limited. The contributions are organised as a progression across domains and modalities, and each chapter adds a concrete building block toward medical models that clinicians can scrutinise, adapt, and reuse across settings.

Chapter 2 introduces an explainable multimodal framework for COVID-19 prognosis from adequately recovered, early chest X-rays and triage attributes. The method describes how real-world data quality can be recovered and how the resulting evidence can be combined in practice: fusing heterogeneous inputs in a controlled way so that individual modalities retain interpretable roles, and producing explanations at both model and patient level. The chapter shows that, in a safety-critical settings, multimodal performance is most valuable when the path from evidence to prediction is transparent enough to be inspected by clinicians.

Chapter 3 addresses what happens when models face the variability of real clinical environments. It proposes efficient adaptation strategies for multi-modal multiple sclerosis lesion segmentation models under distribution shift, explicitly targeting minimal supervision and resistance to forgetting. On top of adaptation, the chapter introduces Latent Distance Analysis, a label-free reliability estimator grounded in latent-space geometry, so that a deployed model can flag likely failure cases even when target labels are unavailable. Together, these contributions emphasise that deployment trust depends on both the ability to recover performance under drift and the ability to warn when recovery may be uncertain.

Chapter 4 demonstrates that large-scale self-supervision can provide generalisable ECG representations in a domain where robust unimodal foundations are still absent. By presenting HuBERT-ECG as a foundation backbone for scalable cardiac applications, the chapter shows that transferable physiological features can be learned from heterogeneous unlabelled ECG data and adapted efficiently to many downstream tasks. The evaluation stresses an important principle that runs through the thesis: generality must be supported by systematic cross-scenario evidence rather than inferred from scale alone.

Chapter 5 extends this foundational effort to settings far from the 12-lead clinical standard. It proposes a lightweight any-lead ECG foundation model based on sparse, clinically grounded spatiotemporal ECG graphs, masked node modelling, and stochastic lead sampling to learn lead-invariant representations distilled from large-scale 12-lead ECG corpora. Even while the work is still being completed, early results indicate that sharp inductive priors and lead invariance enable a compact GNN-based model to surpass much larger foundational baselines at early pretraining stages, pointing toward scalable transfer to reduced-lead and wearable domains where trustworthy foundations are most needed. More broadly, this chapter highlights open challenges and opportunities in developing ECG foundation models that remain diagnostically meaningful under arbitrary lead configurations, suggesting a path toward systems capable of preserving clinically relevant signals, even when standard acquisition assumptions no longer hold, without positioning these results as a definitive endpoint.

Taken together, the thesis contributes a set of methods that make trustworthiness operational and, more broadly, supports a pragmatic path toward trustworthy medical foundation models. In medicine, the prerequisites for fully general multimodal systems remain hard to satisfy. Progress therefore comes from building strong, carefully validated backbones and pairing them with robustness and

reliability mechanisms that can anticipate deployment realities and challenges. The methods developed here show that trust is not a property that emerges from or accuracy alone, but from deliberate alignment between model structure and inductive bias, clinical evidence, and the conditions under which a system is expected to operate. Across the thesis, different modelling paradigms are therefore adopted deliberately to prioritise distinct but complementary dimensions of trustworthiness, reflecting the fact that no single mechanism is sufficient across all clinical data regimes and deployment scenarios.

Looking ahead, we plan to complete the any-lead ECG foundation model and to extend the thesis contributions toward broader settings, including multimodal representation alignment and reliability estimation across domains.

## References

- [1] K. Lekadir et al., “FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare,” *BMJ*, vol. 388, p. e081554, Feb. 2025.
- [2] C. I. Eke, L. Shuib, and M. Phiri, “A comprehensive survey of AI trust frameworks in healthcare: current state and future directions,” *Health Serv Outcomes Res Methodol*, Sep. 2025.
- [3] M. Goisauf et al., “Trust, Trustworthiness, and the Future of Medical AI: Outcomes of an Interdisciplinary Expert Workshop,” *J Med Internet Res*, vol. 27, p. e71236, Jun. 2025.
- [4] M. Mora-Cantalops, E. García-Barriocanal, and M.-Á. Sicilia, “Trustworthy AI Guidelines in Biomedical Decision-Making Applications: A Scoping Review,” *Big Data and Cognitive Computing*, vol. 8, no. 7, p. 73, Jul. 2024.
- [5] A. S. Albahri et al., “A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion,” *Information Fusion*, vol. 96, pp. 156–191, Aug. 2023.
- [6] F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi, “Review of multimodal machine learning approaches in healthcare,” *Information Fusion*, vol. 114, p. 102690, Feb. 2025.
- [7] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, “Deep Multimodal Learning with Missing Modality: A Survey,” 2024. Preprint at <https://arxiv.org/abs/2409.07825>.
- [8] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Med Image Anal*, vol. 79, p. 102470, Jul. 2022.
- [9] R.-K. Sheu and M. S. Pardeshi, “A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System,” *Sensors*, vol. 22, no. 20, p. 8068, Oct. 2022.
- [10] I. D. Mienye et al., “A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges,” *Inform Med Unlocked*, vol. 51, p. 101587, 2024.
- [11] Z. Sadeghi et al., “A review of Explainable Artificial Intelligence in healthcare,” *Computers and Electrical Engineering*, vol. 118, p. 109370, Aug. 2024.
- [12] Y. Brima and M. Atemkeng, “Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis,” *BioData Min*, vol. 17, no. 1, p. 18, Jun. 2024.
- [13] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, “Transparency of deep neural networks for medical image analysis: A review of interpretability methods,” *Comput Biol Med*, vol. 140, p. 105111, Jan. 2022.
- [14] E. H. Houssein, A. M. Gamal, E. M. G. Younis, and E. Mohamed, “Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review,” *Cluster Comput*, vol. 28, no. 7, p. 469, Sep. 2025.
- [15] K. N. Alam, P. B. Zadeh, and A. Sheikh-Akbari, “Attribution-Based Explainability in Medical Imaging: A Critical Review on Explainable Computer Vision (X-CV) Techniques and Their Applications in Medical AI,” *Electronics (Basel)*, vol. 14, no. 15, p. 3024, Jul. 2025.
- [16] J. Silva-Rodríguez, F. Shakeri, H. Bahig, J. Dolz, and I. Ben Ayed, “Few-Shot, Now for Real: Medical VLMs Adaptation Without Balanced Sets or Validation,” 2026, pp. 237–247..
- [17] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk, “Domain Generalization for Medical Image Analysis: A Review,” *Proceedings of the IEEE*, vol. 112, no. 10, pp. 1583–1609, Oct. 2024.
- [18] E. Pachetti and S. Colantonio, “A systematic review of few-shot learning in medical imaging,” *Artif Intell Med*, vol. 156, p. 102949, Oct. 2024.
- [19] L. Julian Lechuga Lopez, S. Elsharief, A. D. Jorf, F. Darwish, C. Ma, and F. E. Shamout, “Uncertainty Quantification for Machine Learning in Healthcare: A Survey,” 2025.

- [20] L. Han, “Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings,” *JAMA Netw Open*, vol. 8, no. 6, p. e2513705, Jun. 2025.
- [21] R. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” Aug. 2021.
- [22] S. Matta et al., “A systematic review of generalization research in medical image classification,” *Comput Biol Med*, vol. 183, p. 109256, Dec. 2024.
- [23] M. Moor et al., “Foundation models for generalist medical artificial intelligence,” *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023.
- [24] X. Liu et al., “Self-Supervised Learning: Generative or Contrastive,” *IEEE Trans Knowl Data Eng*, vol. 35, no. 1, pp. 857–876, Jan. 2023.
- [25] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training,” 2018. Preprint at <https://openai.com/research/language-unsupervised>
- [26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. Preprint at <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>
- [27] T. B. Brown et al., “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, 2020.
- [28] OpenAI et al., “GPT-4 Technical Report,” Mar. 2023.
- [29] M. Caron et al., “Emerging Properties in Self-Supervised Vision Transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.
- [30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 16000–16009. Preprint at <http://arxiv.org/abs/2111.06377>
- [31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” *Advances in neural information systems*, vol. 33, pp. 12449–12460, 2020.
- [32] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 3451–3460, 2021.
- [33] R. Schäfer et al., “Overcoming data scarcity in biomedical imaging with a foundational multi-task model,” *Nat Comput Sci*, vol. 4, no. 7, pp. 495–509, Jul. 2024.
- [34] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, and J. Qadir, “Privacy-preserving artificial intelligence in healthcare: Techniques and applications,” *Comput Biol Med*, vol. 158, p. 106848, May 2023.
- [35] M. Y. Lu et al., “A visual-language foundation model for computational pathology,” *Nat Med*, vol. 30, no. 3, pp. 863–874, Mar. 2024.
- [36] H. Xu et al., “A whole-slide foundation model for digital pathology from real-world data,” *Nature*, vol. 630, no. 8015, pp. 181–188, Jun. 2024.
- [37] E. Vorontsov et al., “A foundation model for clinical-grade computational pathology and rare cancers detection,” *Nat Med*, vol. 30, no. 10, pp. 2924–2935, Oct. 2024.
- [38] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, “Knowledge-enhanced visual-language pre-training on chest radiology images,” *Nat Commun*, vol. 14, no. 1, p. 4542, Jul. 2023.
- [39] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” *Nat Biomed Eng*, vol. 6, no. 12, pp. 1399–1406, Sep. 2022.
- [40] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment anything in medical images,” *Nat Commun*, vol. 15, no. 1, p. 654, Jan. 2024.
- [41] M. Christensen, M. Vukadinovic, N. Yuan, and D. Ouyang, “Vision–language foundation model for echocardiogram interpretation,” *Nat Med*, vol. 30, no. 5, pp. 1481–1488, May 2024.

- [42] K. Zhang et al., “A generalist vision–language foundation model for diverse biomedical tasks,” *Nat Med*, Aug. 2024.
- [43] E. Coppola, D. Ferrari, M. Savardi, and A. Signoroni, “Explainable AI for COVID-19 prognosis from early chest x-ray and clinical data in the context of the COVID-CXR international hackathon,” in *Medical Imaging 2023: Computer-Aided Diagnosis*, K. M. Iftekharuddin and W. Chen, Eds., SPIE, Apr. 2023, p. 71.
- [44] E. Coppola, M. Savardi, and A. Signoroni, “Towards Robust and Reliable Multi-Modal 3D Segmentation of Multiple Sclerosis Lesions,” *Pattern Recognit Lett*, Dec. 2025, doi: 10.1016/j.patrec.2025.12.008.
- [45] E. Coppola, M. Savardi, M. Massussi, M. Adamo, M. Metra, and A. Signoroni, “HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications,” Nov. 18, 2024. Preprint at <https://www.medrxiv.org/content/10.1101/2024.11.14.24317328v3>.
- [46] M. Roberts et al., “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” *Nat Mach Intell*, vol. 3, no. 3, pp. 199–217, Mar. 2021.
- [76] A. Signoroni et al., “BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset,” *Med Image Anal*, vol. 71, p. 102046, Jul. 2021.
- [48] F. E. Shamout et al., “An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department,” *NPJ Digit Med*, vol. 4, no. 1, p. 80, May 2021.
- [49] P. Soda et al., “AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study,” *Med Image Anal*, vol. 74, p. 102216, Dec. 2021.
- [50] Z. Jiao et al., “Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study,” *Lancet Digit Health*, vol. 3, no. 5, pp. e286–e294, May 2021.
- [51] J. Cheng et al., “COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data,” *Eur Radiol*, vol. 32, no. 7, pp. 4446–4456, Jul. 2022.
- [52] A. Borghesi and R. Maroldi, “COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression,” *Radiol Med*, vol. 125, no. 5, pp. 509–513, May 2020.
- [53] F. Shi et al., “Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19,” *IEEE Rev Biomed Eng*, vol. 14, pp. 4–15, 2021.
- [54] Y. Oh, S. Park, and J. C. Ye, “Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets,” *IEEE Trans Med Imaging*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.
- [55] F. Sardanelli and G. Di Leo, “Assessing the Value of Diagnostic Tests in the Coronavirus Disease 2019 Pandemic,” *Radiology*, vol. 296, no. 3, pp. E193–E194, Sep. 2020.
- [56] S. M. Lundberg and L. Su-in, “A unified approach to interpreting model prediction,” *Advances in Neural Information Processing Systems*, 2017.
- [57] O. Troyanskaya et al., “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, Jun. 2001.
- [58] I. Au-Yong et al., “Chest Radiograph Scoring Alone or Combined with Other Risk Scores for Predicting Outcomes in COVID-19,” *Radiology*, vol. 302, no. 2, pp. 460–469, Jan. 2022.
- [59] D. R. van Nderpelt et al., “Scanner-specific optimisation of automated lesion segmentation in MS,” *Neuroimage Clin*, vol. 44, p. 103680, 2024.
- [60] T. Wiltgen et al., “LST-AI: A deep learning ensemble for accurate MS lesion segmentation,” *Neuroimage Clin*, vol. 42, p. 103611, 2024.
- [61] E. López-Varela, N. O. Pascual, J. Quezada-Sánchez, C. Oreja-Guevara, E. S. Bueso, and N. Barreira, “Enhanced multiple sclerosis diagnosis using high-resolution 3D OCT volumes with synthetic slices,” *Pattern Recognit Lett*, vol. 189, pp. 99–105, Mar. 2025.

- [62] P. D. Molyneux et al., “Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility,” *Neuroradiology*, vol. 41, no. 12, pp. 882–888, Dec. 1999.
- [63] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham, “A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions,” *Neuroimage*, vol. 49, no. 2, pp. 1524–1535, Jan. 2010.
- [64] E. Roura et al., “A toolbox for multiple sclerosis lesion segmentation,” *Neuroradiology*, vol. 57, no. 10, pp. 1031–1043, Oct. 2015.
- [65] K. Selvaganesan et al., “Robust, atlas-free, automatic segmentation of brain MRI in health and disease,” *Heliyon*, vol. 5, no. 2, p. e01226, Feb. 2019.
- [66] A. M. Hindsholm et al., “Scanner agnostic large-scale evaluation of MS lesion delineation tool for clinical MRI,” *Front Neurosci*, vol. 17, May 2023.
- [67] R. Kushol, A. H. Wilman, S. Kalra, and Y.-H. Yang, “DSMRI: Domain Shift Analyzer for Multi-Center MRI Datasets,” *Diagnostics*, vol. 13, no. 18, p. 2947, Sep. 2023.
- [68] M. Salem et al., “Multiple Sclerosis Lesion Synthesis in MRI Using an Encoder-Decoder U-NET,” *IEEE Access*, vol. 7, pp. 25171–25184, 2019.
- [69] S. Gaj, B. Thoomukuntla, D. Ontaneda, and K. Nakamura, “Subject-Based Transfer Learning in Longitudinal Multiple Sclerosis Lesion Segmentation,” *Journal of Neuroimaging*, vol. 35, no. 1, Jan. 2025.
- [70] J. Zhang et al., “Harmonization-enriched domain adaptation with light fine-tuning for multiple sclerosis lesion segmentation,” in *Medical Imaging 2024: Clinical and Biomedical Imaging*, B. S. Gimi and A. Krol, Eds., SPIE, Apr. 2024, p. 97.
- [71] O. Wiles, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. D. Dvijotham, and A. T. Cemgil, “A Fine-Grained Analysis on Distribution Shifts”. Preprint at <https://arxiv.org/abs/2110.11328>.
- [72] K. Fan et al., “Test-Time Linear Out-of-Distribution Detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 23752–23761.
- [73] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of distribution image detection in neural networks,” in *Proceedings of the International Conference on Learning Representations*, 2018.
- [74] D. Hendrycks and K. Gimpel, “A BASELINE FOR DETECTING MISCLASSIFIED AND OUT-OF-DISTRIBUTION EXAMPLES IN NEURAL NETWORKS,” in *Proceedings of the International Conference on Learning Representations*, 2017.
- [75] A. Rondinella et al., “Enhancing Multiple Sclerosis Lesion Segmentation in Multimodal MRI Scans with Diffusion Models,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2023, pp. 3733–3740.
- [76] S. Aslani et al., “Multi-branch convolutional neural network for multiple sclerosis lesion segmentation,” *Neuroimage*, vol. 196, pp. 1–15, Aug. 2019.
- [77] M. Hashemi, M. Akhbari, and C. Jutten, “Delve into Multiple Sclerosis (MS) lesion exploration: A modified attention U-Net for MS lesion segmentation in Brain MRI,” *Comput Biol Med*, vol. 145, p. 105402, Jun. 2022.
- [78] A. Kaur, L. Kaur, and A. Singh, “DeepCONN: patch-wise deep convolutional neural networks for the segmentation of multiple sclerosis brain lesions,” *Multimed Tools Appl*, vol. 83, no. 8, pp. 24401–24433, Aug. 2023.
- [79] O. Cetin, B. Canel, G. Dogali, and U. Sakoglu, “Enhancing precision in multiple sclerosis lesion segmentation: A U-net based machine learning approach with data augmentation,” *NeuroImage: Reports*, vol. 5, no. 1, p. 100235, Mar. 2025.
- [80] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, “Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1229–1239, May 2016.

- [81] S. R. Hashemi, S. S. Mohseni Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2019.
- [82] N. Gessert et al., "Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs," *Computerized Medical Imaging and Graphics*, vol. 84, p. 101772, Sep. 2020.
- [83] R. A. Kamraoui, V.-T. Ta, T. Tourdias, B. Mansencal, J. V. Manjon, and P. Coupé, "DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation," *Med Image Anal*, vol. 76, p. 102312, Feb. 2022.
- [84] S. Valverde et al., "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *Neuroimage*, vol. 155, pp. 159–168, Jul. 2017.
- [85] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi, "LONGITUDINAL MULTIPLE SCLEROSIS LESION SEGMENTATION USING 3D CONVOLUTIONAL NEURAL NETWORKS," in *Proceedings of the 2015 longitudinal multiple sclerosis lesion segmentation challenge*, 2015.
- [86] S. Valverde et al., "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *Neuroimage Clin*, vol. 21, p. 101638, 2019.
- [87] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based Out-of-Distribution Detection," in *Advances in Neural Information Processing Systems*, 2020.
- [88] K. Lee, K. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," in *Advances in Neural Information Processing Systems*, 2018.
- [89] J. Chen, Y. Li, Y. Liang, and S. Jha, "ATOM: Robustifying Out-of-distribution Detection Using Outlier Mining," in *European Conference on Machine Learning*, 2021.
- [90] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.
- [91] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "TENT: FULLY TEST-TIME ADAPTATION BY ENTROPY MINIMIZATION," in *Proceedings of the International Conference on Learning Representations*, 2021.
- [92] A. Rondinella et al., "ICPR 2024 Competition on Multiple Sclerosis Lesion Segmentation—Methods and Results," 2025, pp. 1–16.
- [93] A. Carass et al., "Longitudinal multiple sclerosis lesion segmentation: Resource and challenge," *Neuroimage*, vol. 148, pp. 77–102, Mar. 2017.
- [94] O. Commowick et al., "Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure," *Sci Rep*, vol. 8, no. 1, p. 13650, Sep. 2018.
- [95] Ž. Lesjak et al., "A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus," *Neuroinformatics*, vol. 16, no. 1, pp. 51–63, Jan. 2018.
- [96] O. Commowick et al., "Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset," *Neuroimage*, vol. 244, p. 118589, Dec. 2021.
- [97] N. R. Smilowitz and J. S. Berger, "Perioperative Cardiovascular Risk Assessment and Management for Noncardiac Surgery," *JAMA*, vol. 324, no. 3, p. 279, Jul. 2020.
- [98] Y. J. van de Vegte et al., "Genetic insights into resting heart rate and its role in cardiovascular disease," *Nat Commun*, vol. 14, no. 1, p. 4646, Aug. 2023.
- [99] P. A. Friedman, "The Electrocardiogram at 100 Years: History and Future," *Circulation*, vol. 149, no. 6, pp. 411–413, Feb. 2024.
- [100] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, p. 102289, Jun. 2022.
- [101] S. V. Kalmady et al., "Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level," *NPJ Digit Med*, vol. 7, no. 1, p. 133, May 2024.

- [102] W. Sun et al., “ECG for high-throughput screening of multiple diseases: Proof-of-concept using multi-diagnosis deep learning from population-based datasets,” Oct. 2022.
- [103] S. Goto et al., “Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection,” *Circulation*, vol. 146, no. 10, pp. 755–769, Sep. 2022.
- [104] M. Kim et al., “Deep learning for predicting rehospitalization in acute heart failure: Model foundation and external validation,” *ESC Heart Fail*, Jul. 2024.
- [105] N. Musa et al., “A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram,” *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, pp. 9677–9750, Jul. 2023.
- [106] S. G. Finlayson et al., “The Clinician and Dataset Shift in Artificial Intelligence,” *New England Journal of Medicine*, vol. 385, no. 3, pp. 283–286, Jul. 2021.
- [107] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North, Stroudsburg, PA, USA: Association for Computational Linguistics*, 2019, pp. 4171–4186.
- [108] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, 2020.
- [109] D. Kiyasseh, Tingting Zhu, and David A. Clifton., “Clocs: Contrastive learning of cardiac signals across space, time, and patients,” in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 5606–5615.
- [110] J. Oh, H. Chung, J. Kwon, D. Hong, and E. Choi, “Lead-agnostic self-supervised learning for local and global representations of electrocardiograms,” in *Conference on Health, Inference and Learning*, 2022, pp. 338–353.
- [111] W. Liu et al., “Dense lead contrast for self-supervised representation learning of multilead electrocardiograms,” *Inf Sci (N Y)*, vol. 634, pp. 189–205, Jul. 2023.
- [112] W. Liu, H. Zhang, S. Chang, H. Wang, J. He, and Q. Huang, “A joint cross-dimensional contrastive learning framework for 12-lead ECGs and its heterogeneous deployment on SoC,” *Comput Biol Med*, vol. 152, p. 106390, Jan. 2023.
- [113] T. Mehari and N. Strodthoff, “Self-supervised representation learning from 12-lead ECG data,” *Comput Biol Med*, vol. 141, p. 105114, Feb. 2022.
- [114] S. Zhang et al., “ECGFM: A foundation model for ECG analysis trained on a multi-center million-ECG dataset,” *Information Fusion*, vol. 124, p. 103363, Dec. 2025.
- [115] A. Vaid et al., “A foundational vision transformer improves diagnostic performance for electrocardiograms,” *NPJ Digit Med*, vol. 6, no. 1, p. 108, Jun. 2023.
- [116] Y. Tian et al., “Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG,” *Cell Rep Med*, vol. 5, no. 12, p. 101875, Dec. 2024.
- [117] J. Li et al., “An Electrocardiogram Foundation Model Built on over 10 Million Recordings with External Evaluation across Multiple Domains,” Apr. 2025. Preprint at <https://arxiv.org/pdf/2410.04133>.
- [118] S. Choi, S. Mousavi, P. Si, H. G. Yhdego, F. Khadem, and F. Afghah, “ECGBERT: Understanding Hidden Language of ECGs with Self-Supervised Representation Learning,” Jun. 2023. Preprint at <https://arxiv.org/abs/2306.06340>.
- [119] A. Vaswani et al., “Attention Is All You Need,” *Adv Neural Inf Process Syst*, Jun. 2017.
- [120] R. Hu, J. Chen, and L. Zhou, “Spatiotemporal self-supervised representation learning from multi-lead ECG signals,” *Biomed Signal Process Control*, vol. 84, p. 104772, Jul. 2023.
- [121] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, “HMM-Based Audio Keyword Generation,” 2004, pp. 566–574.
- [122] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, ACM-SIAM Symposium on Discrete Algorithms, Jan. 2007, pp. 1027–1035.

- [123] A. Fan, E. Grave, and A. Joulin, “Reducing Transformer Depth on Demand with Structured Dropout,” in Proceedings of the International Conference on Learning Representations, Sep. 2019. Preprint at <https://arxiv.org/abs/1909.11556>
- [124] E. A. Perez Alday et al., “Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020,” *Physiol Meas*, vol. 41, no. 12, p. 124003, Dec. 2020.
- [125] M. A. Reyna et al., “Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021,” in 2021 Computing in Cardiology (CinC), IEEE, Sep. 2021, pp. 1–4.
- [126] A. H. Ribeiro et al., “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nat Commun*, vol. 11, no. 1, p. 1760, Apr. 2020.
- [127] F. Liu et al., “An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection,” *J Med Imaging Health Inform*, vol. 8, no. 7, pp. 1368–1373, Sep. 2018.
- [128] R. Bousseljot, D. Kreiseler, and A. Schnabel, “Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet,” *Biomedizinische Technik/Biomedical Engineering*, pp. 317–318, Jul. 2009.
- [129] P. Wagner et al., “PTB-XL, a large publicly available electrocardiography dataset,” *Sci Data*, vol. 7, no. 1, p. 154, May 2020.
- [130] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals,” *Circulation*, vol. 101, no. 23, Jun. 2000.
- [131] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, “A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients,” *Sci Data*, vol. 7, no. 1, p. 48, Feb. 2020.
- [132] J. Zheng et al., “Optimal Multi-Stage Arrhythmia Classification Approach,” *Sci Rep*, vol. 10, no. 1, p. 2898, Feb. 2020.
- [133] H. Liu et al., “A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements,” *Sci Data*, vol. 9, no. 1, p. 272, Jun. 2022.
- [134] B. Gow et al., “MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset.”
- [135] C. S. Cardoso et al., “Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile,” *BMJ Open*, vol. 6, no. 5, p. e011181, May 2016.
- [136] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep Clustering for Unsupervised Learning of Visual Features,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018.
- [137] S. Azizi et al., “Big Self-Supervised Models Advance Medical Image Classifications,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3478–3488.
- [138] S. Pai et al., “Foundation model for cancer imaging biomarkers,” *Nat Mach Intell*, vol. 6, no. 3, pp. 354–367, Mar. 2024.
- [139] F. Haghighi, M. R. H. Taher, M. B. Gotway, and Ji. Liang, “DiRA: Discriminative, Restorative, and Adversarial Learning for Self-supervised Medical Image Analysis,” in Proceedings of the IEEE/CVF international conference on computer vision, 2022, pp. 20824–20834.
- [140] T. Richter, M. Bahrami, Y. Xia, D. S. Fischer, and F. J. Theis, “Delineating the effective use of self-supervised learning in single-cell genomics,” *Nat Mach Intell*, vol. 7, no. 1, pp. 68–78, Dec. 2024.
- [141] J. Tang, T. Xia, Y. Lu, C. Mascolo, and A. Saeed, “Electrocardiogram Report Generation and Question Answering via Retrieval-Augmented Self-Supervised Modeling,” in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Apr. 2025, pp. 1–5.
- [142] B. VanBerlo, J. Hoey, and A. Wong, “A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound,” *BMC Med Imaging*, vol. 24, no. 1, p. 79, Apr. 2024.

- [143] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform Manifold Approximation and Projection,” *J Open Source Softw*, vol. 3, no. 29, p. 861, Sep. 2018.
- [144] A. Natarajan et al., “A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification,” in *Computing in Cardiology*, Dec. 2020.
- [145] S. Hong, W. Zhang, C. Sun, Y. Zhou, and H. Li, “Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020,” *Front Physiol*, vol. 12, Jan. 2022.
- [146] T. Mehari and N. Strodthoff, “Towards Quantitative Precision for ECG Analysis: Leveraging State Space Models, Self-Supervision and Patient Metadata,” *IEEE J Biomed Health Inform*, vol. 27, no. 11, pp. 5326–5334, Nov. 2023.
- [147] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, “Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL,” *IEEE J Biomed Health Inform*, vol. 25, no. 5, pp. 1519–1528, May 2021.
- [148] Y. Na, M. Park, Y. Tae, and S. Joo, “Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram,” in *Proceedings of the International Conference on Learning Representations*, 2024.
- [149] A. M. Ferreira et al., “Two-year death prediction models among patients with Chagas Disease using machine learning-based methods,” *PLoS Negl Trop Dis*, vol. 16, no. 4, p. e0010356, Apr. 2022.
- [150] P. C. Wouters et al., “Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy,” *Eur Heart J*, vol. 44, no. 8, pp. 680–692, Feb. 2023.
- [151] C. Leclercq et al., “Wearables, telemedicine, and artificial intelligence in arrhythmias and heart failure: Proceedings of the European Society of Cardiology Cardiovascular Round Table,” *EP Europace*, vol. 24, no. 9, pp. 1372–1383, Oct. 2022.
- [152] G. Monachino et al., “Self-DANA: A Resource-Efficient Channel-Adaptive Self-Supervised Approach for Foundation Models,” 2025.
- [153] J. Jin, H. Wang, H. Li, J. Li, J. Pan, and S. Hong, “Reading your heart: Learning ecg words and sentences via pre-training ecg language model,” in *Proceedings of the International Conference on Learning Representations*, 2025.
- [154] Y. Tian, D. Krishnan, and P. Isola, “Contrastive Multiview Coding,” 2020, pp. 776–794.
- [155] M. Jin et al., “A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection,” *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 12, pp. 10466–10485, Dec. 2024.
- [156] D. EPMoghaddam, A. Muguli, M. Razavi, and B. Aazhang, “A graph-based cardiac arrhythmia classification methodology using one-lead ECG recordings,” *Intelligent Systems with Applications*, vol. 22, p. 200385, Jun. 2024.
- [157] K. Zeinalipour and M. Gori, “Graph Neural Networks for Topological Feature Extraction in ECG Classification,” 2023, pp. 17–27.
- [158] Y. Qiang, X. Dong, X. Liu, Y. Yang, Y. Fang, and J. Dou, “Conv-RGNN: An efficient Convolutional Residual Graph Neural Network for ECG classification,” *Comput Methods Programs Biomed*, vol. 257, p. 108406, Dec. 2024.
- [159] T. Chen, Y. Ma, Z. Pan, W. Wang, and J. Yu, “Fusion of multi-scale feature extraction and adaptive multi-channel graph neural network for 12-lead ECG classification,” *Comput Methods Programs Biomed*, vol. 265, p. 108725, Jun. 2025.
- [160] Y. Rong, W. Huang, T. Xu, and J. Huang, “DropEdge: Towards Deep Graph Convolutional Networks on Node Classification,” in *Proceedings of the International Conference on Learning Representations*, 2020.
- [161] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, “Graph Attention Networks,” in *Proceedings of the International Conference on Learning Representations*, 2018.

- [162] F. Mason, A. C. Pandey, M. Gadaleta, E. J. Topol, E. D. Muse, and G. Quer, “AI-enhanced reconstruction of the 12-lead electrocardiogram via 3-leads with accurate clinical assessment,” *NPJ Digit Med*, vol. 7, no. 1, p. 201, Aug. 2024.
- [163] M. Seják, J. Sido, and D. Žahour, “ElectroCardioGuard: Preventing patient misidentification in electrocardiogram databases through neural networks,” *Knowl Based Syst*, vol. 280, p. 111014, Nov. 2023.
- [164] J. Tan et al., “A pediatric ECG database with disease diagnosis covering 11643 children,” *Sci Data*, vol. 12, no. 1, p. 867, May 2025.
- [165] J. Lai et al., “Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset,” *Nat Commun*, vol. 14, no. 1, p. 3741, Jun. 2023.
- [166] C. Liu et al., “An open access database for the evaluation of heart sound algorithms,” *Physiol Meas*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016.
- [167] Nicolas Keriven, “Not too little, not too much: a theoretical analysis of graph (over)smoothing,” in *Advances in Neural Information Processing Systems*, 2022.
- [168] E. Knight, E. K. Oikonomou, A. Aminorroaya, A. F. Pedroso, and R. Khera, “Wearable-Echo-FM: An ECG-echo foundation model for single lead electrocardiography,” Jun. 11, 2025.
- [169] A. Imane, B. Abdelmajid, O. H. Hanaa, E. D. Y. Frederic, and D. L. Hilal, “A Focused Survey of ECG Datasets for Artificial Intelligence-Based Atrial Fibrillation Detection,” *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 7, 2025.
- [170] Y. Ye, L. Xia, and C. Huang, “Graph Masked Autoencoder for Sequential Recommendation,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2023, pp. 321–330.
- [171] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, “Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 1548–1554.
- [172] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How Powerful are Graph Neural Networks,” in *Proceedings of the International Conference on Learning Representations*, 2019.
- [173] C. Bodnar, Francesco Di Giovanni, B. P. Chamberlain, P. Liò, and M. Bronstein, “Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs,” in *Advances in Neural Information Processing Systems*, 2022.
- [174] F. Barbero, C. Bodnar, H. S. de Ocariz Borde, and P. Liò, “Sheaf Attention Networks,” *NeurIPS2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.
- [175] S. Brody, U. Alon, and E. Yahov, “How Attentive are Graph Attention Networks?,” in *International Conference on Learning Representations*, 2022.
- [176] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, “RoFormer: Enhanced transformer with Rotary Position Embedding,” *Neurocomputing*, vol. 568, p. 127063, Feb. 2024.
- [177] A. Rogers, O. Koleva, and A. Rumshisky, “A primer in BERTology: What we know about how BERT works,” *Trans Assoc Comput Linguist*, vol. 8, pp. 842–866, 2021.
- [178] A. Vedaldi and S. Soatto, “Quick Shift and Kernel Methods for Mode Seeking,” 2008, pp. 705–718.
- [179] M. Xu, L. Y. Duan, J. Cai, L. T. Chia, C. Xu, and Q. Tian, “HMM-based audio keyword generation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3333, 2004.
- [180] A. K. Singh and S. Krishnan, “ECG signal feature extraction trends in methods and applications,” *Biomed Eng Online*, vol. 22, no. 1, p. 22, Mar. 2023.
- [181] M. Boussaa, I. Atouf, M. Atibi, and A. Bennis, “ECG signals classification using MFCC coefficients and ANN classifier,” in *2016 International Conference on Electrical and Information Technologies (ICEIT)*, IEEE, May 2016, pp. 480–484.
- [182] R. L. Thorndike, “Who belongs in the family?,” *Psychometrika*, vol. 18, no. 4, pp. 267–276, Dec. 1953.

- [183] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Trans Pattern Anal Mach Intell*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
- [184] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Commun Stat Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.
- [185] S. Abnar and W. Zuidema, “Quantifying Attention Flow in Transformers,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 4190–4197.

# Supplementary Material

## Appendix A

| Clinical feature abbreviation | Meaning                                                              |
|-------------------------------|----------------------------------------------------------------------|
| Sex                           | Sex of the subject                                                   |
| DaysFever                     | Days of fever at hospitalization                                     |
| Hospital                      | Hospital ID                                                          |
| PaCO2                         | Partial pressure of CO2 in the arterial blood                        |
| Temp_C                        | Subject temperature in Celsius degrees                               |
| pH                            | Blood pH                                                             |
| Brixia-score confidence       | Aggregated BSNet confidence in formulating the regional Brixia-score |
| DifficultyInBreathing         | Subject's difficulty in breathing                                    |
| RBC                           | Red Blood Cell count                                                 |
| D dimer                       | Fibrin degradation product                                           |
| Glucose                       | Blood glucose concentration                                          |
| PCT                           | Procalcitonin                                                        |
| Age                           | Subject's age                                                        |
| CRP                           | C-Reactive protein                                                   |
| SaO2                          | Atrial oxygen saturation                                             |
| WBC                           | White Blood Cell count                                               |
| PaO2                          | Partial pressure of oxygen in arterial blood                         |
| Brixia-score                  | Six-region severity score compute by BSNet                           |
| LDH                           | Lactate dehydrogenase                                                |

**Supplementary Material – Appendix A - Table 1:** Selected clinical features for model training and corresponding meaning.

### Explainability maps creation algorithm

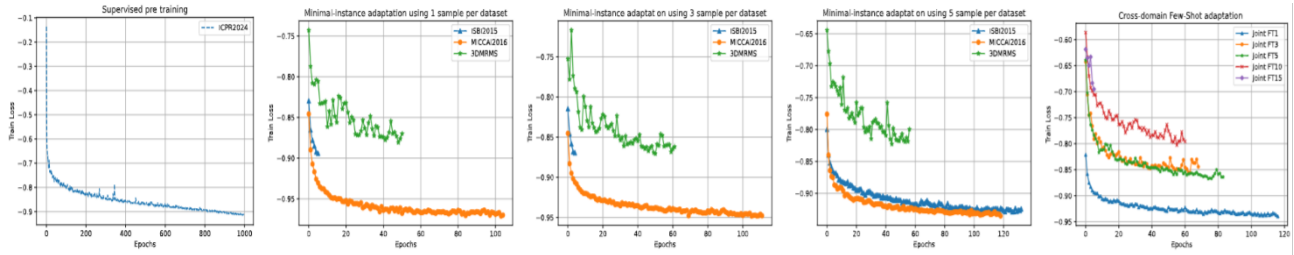
The algorithm for creating the explainability maps derives from the one described in the paper of Signoroni et al. [47] and is reported here for completeness: “The creation of the explainability map starts with the input image division into  $N$  super-pixels, i.e., regions that share similar intensity and pattern, extracted as in [179] . Starting from the input image, we create  $N$  image replicas in which a single super-pixel  $i$  (from 1 to  $N$ ) is masked to zero. We call  $p_0$  the probability map that the model produces starting from the original image ( $2 \times 3 \times 4$  values, i.e., one for each of the 4 severity classes in every lung sector that composes the Brixia-score). We call instead  $p_i$  the probability map produced from the  $i$ -th replica. We then accumulate for the differences between all the super-pixel masked predictions  $p_i$  and the original prediction  $p_0$ . Given  $S$  the set of super-pixels, the output explanation map  $E$  is obtained as:

$$E = \sum_i^{|S|} S_i \cdot (p_i - p_0)$$

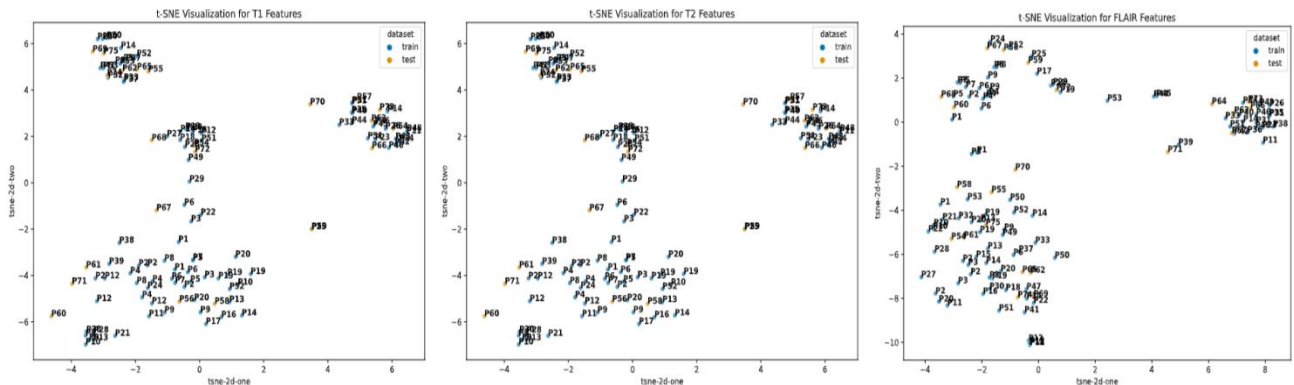
Intuitively, the obtained maps highlight the regions that most account for the score in exam. ... the more intense the colour, the more important the region contribution to the score decision”.

To produce our map, we fused the above described score-related probability maps by retaining, for every super-pixel, the colour and saturation corresponding to the maximum confidence value related to every possible score (from 0 to 3).

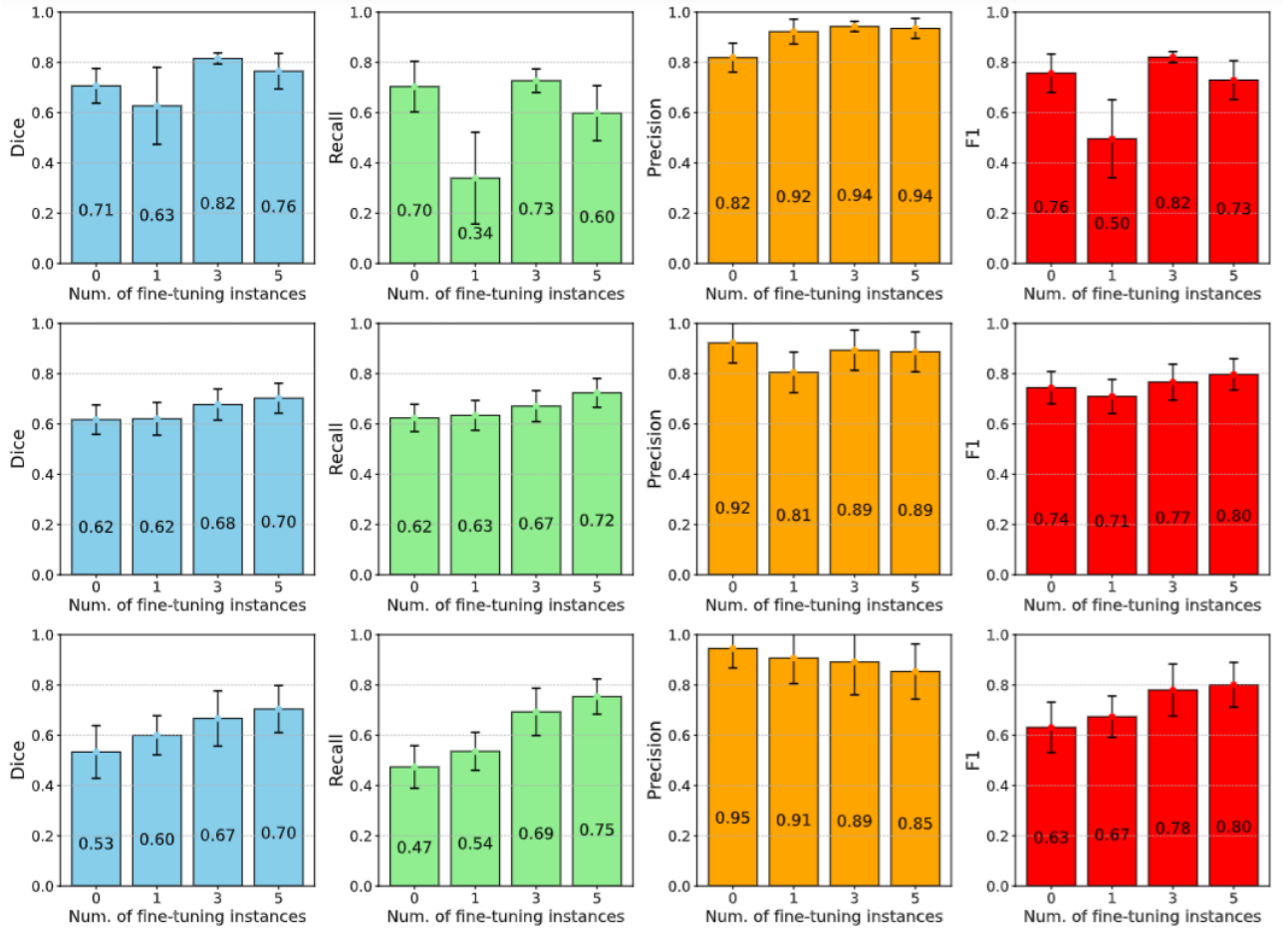
## Appendix B



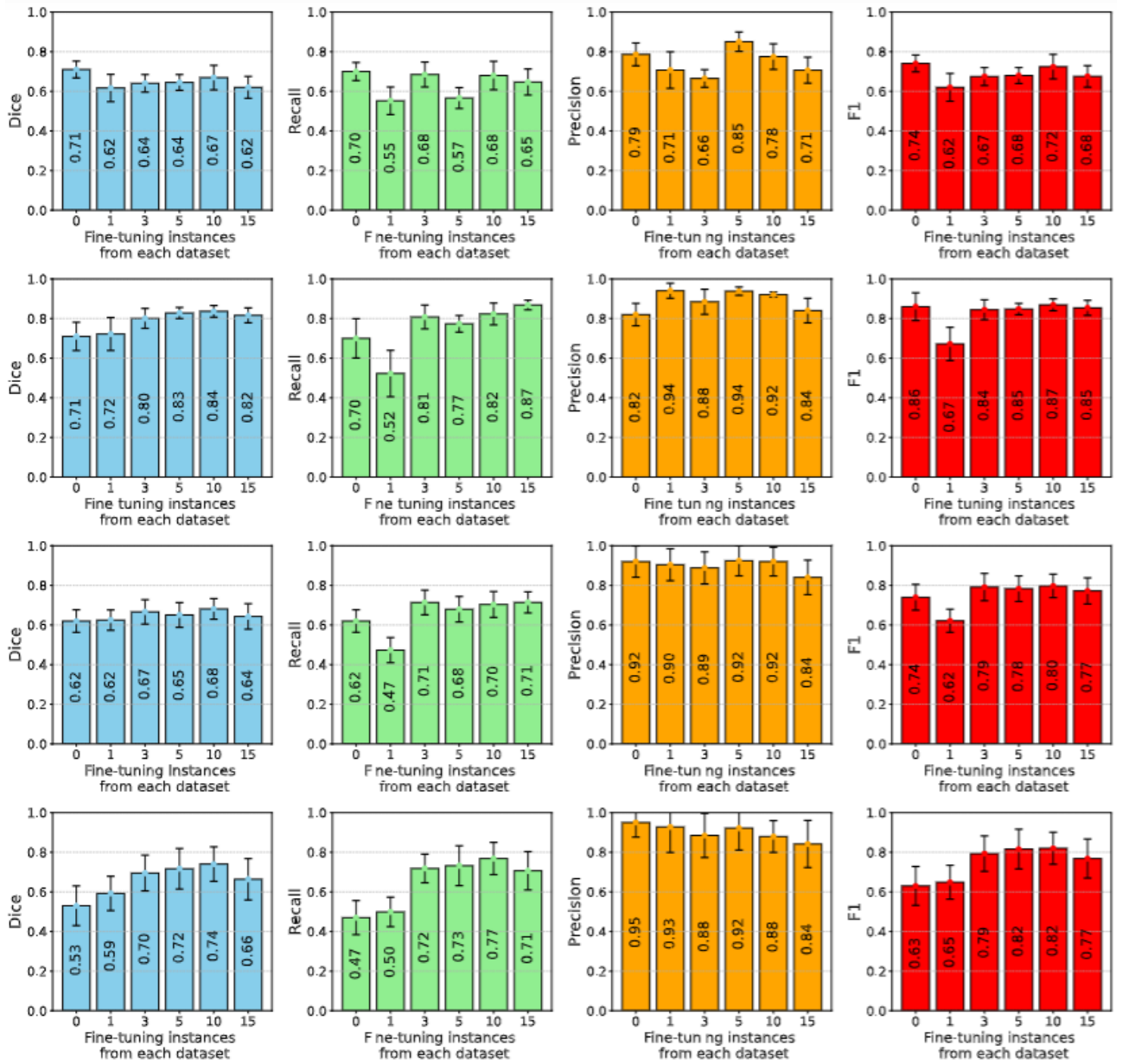
**Supplementary Material -Appendix B - Figure 1:** Plots of the training loss when pre-training, adapting to individual datasets and adapting to multiple datasets simultaneously. The first graph reports the training loss during the supervised pre-training of nnUNet. The second, third and fourth graphs show the training loss during the adaptation to individual datasets using 1, 3, 5 scans from each target source, respectively. The last graph depicts the training loss during the cross-domain few-shot adaptation when 1, 3, 5, 10, 15 scans from each dataset are sampled and used jointly for fine-tuning. “FT” stands for fine-tuning. (Zoom in for better view).



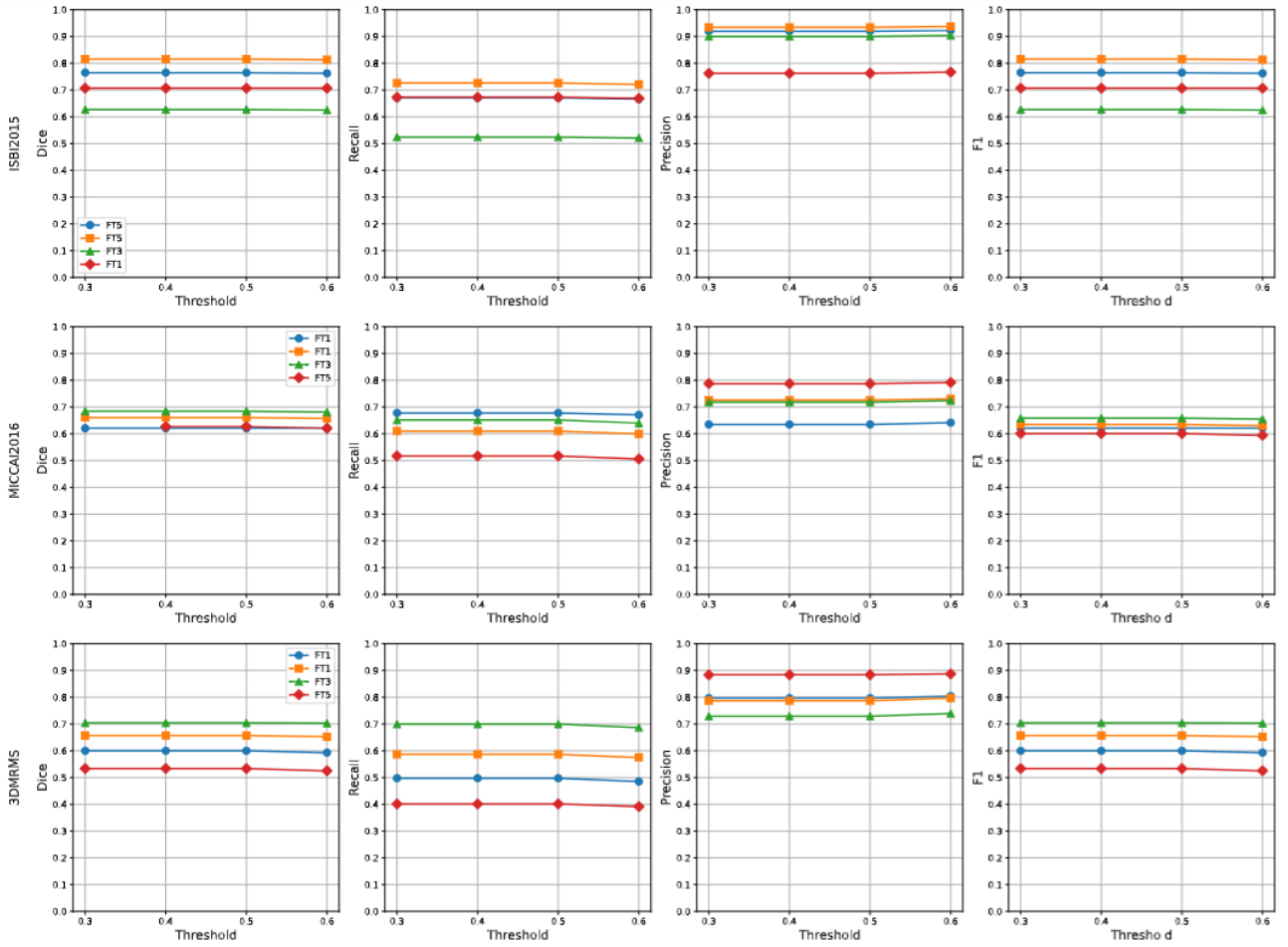
**Supplementary Material – Appendix B - Figure 2:** t-SNE visualisation developed in the context of the ICPR2024’s MSLesSeg Competition of intensity-based features extracted from raw MRI scans included in ICPR2024 dataset. On the left, 2D-compressed T1w features; in the centre, 2D-compressed T2w features; on the right, 2D-compressed FLAIR features. Instance names are show above the corresponding two-dimensional point. Three clusters are discernible and used to instantiate the k-means model used in the Latent Distance Analysis.



**Supplementary Material – Appendix B - Figure 3:** Adapted model performance on individual unseen datasets. From top to bottom, segmentation performance on ISBI2015, MICCAI2016, and 3DMRMS after fine-tuning on 1, 3, 5 volumes from each individual target dataset. Fine-tuning on 0 instances indicates direct transfer of the pretrained model without any adaptation and serves as baseline. Error bars are reported with 95% confidence intervals.



**Supplementary Material – Appendix B - Figure 4:** Model performance after cross-dataset adaptation. From top to bottom, segmentation performance on ICPR2024, ISBI2015, MICCAI2016, and 3DMRMS after cross-dataset adaptation using 1 to 15 instances from each target dataset jointly. Fine-tuning on 0 instances indicates direct transfer of the pretrained model without sampling from any target dataset and serves as baseline. Error bars are reported with 95% CI.



**Supplementary Material – Appendix B - Figure 5:** Sensitivity analysis over the segmentation threshold when adapting the pre-trained model to multiple datasets individually. “FT” means fine-tuning and the subsequent number indicates the number of instances used to adapt the model to a new dataset. When decreasing the segmentation threshold to 0.4 and 0.3, there is no significant change across metrics. In contrast, increasing the threshold to 0.6 results in a small but consistent decrease across all metrics.

## Appendix C

| Diagnosis                                    | Abbreviation | Found Abbreviated Synonyms | Super-conditions |
|----------------------------------------------|--------------|----------------------------|------------------|
| 2nd Degree Av Block                          | 2AVB         | IIAVB                      | AVB              |
| Second Av Block Mobitz Type I                | 2AVB1        | IIAVB I, MOI               | AVB, 2AVB        |
| Mobitz Type II Atrioventricular Block        | 2AVB2        | IIAVB II,                  | AVB, 2AVB        |
| Av Block Varying Conduction                  | AVBVC        |                            | AVB              |
| Av Block Advanced (High Grade)               | AVBHG        |                            | AVB              |
| Av Block Complete (Third Degree)             | 3AVB         | IIIAVB, CHB                | AVB              |
| 1st Degree Av Block                          | 1AVB         | IAVB                       | AVB              |
| Anterior Myocardial Infarction               | ANMI         |                            | MI               |
| Extensive Anterior Myocardial Infarction     | EAMI         |                            | ANMI, MI         |
| St Elevation                                 | STE          | STE                        | MI               |
| Inferoposterolateral Myocardial Infarction   | IPLMI        |                            | MI               |
| Posterior Myocardial Infarction              | PMI          |                            | MI               |
| Inferolateral Myocardial Infarction          | ILMI         |                            | MI               |
| Inferoposterior Myocardial Infarction        | IPMI         |                            | MI               |
| Anterolateral Myocardial Infarction          | ALMI         |                            | MI               |
| Lateral Myocardial Infarction                | LMI          |                            | MI               |
| Acute Myocardial Infarction                  | AMI          |                            | MI               |
| Anteroseptal Myocardial Infarction           | ASMI         |                            | MI               |
| Inferior Myocardial Infarction               | IMI          |                            | MI               |
| Ischemic In Anterior Leads (Subclass)        | ISCA         |                            | MIS              |
| Anterior Ischemia                            | ANMIS        | ISCAN                      | MIS              |
| Inferior Ischaemia                           | IIS          | ISCIN                      | MIS              |
| Subendocardial Injury In Inferior Leads      | INJIN        |                            | MIS              |
| Ischemic In Inferior Leads                   | ISCI         |                            | MIS              |
| Subendocardial Injury In Inferolateral Leads | INJIL        |                            | MIS              |
| Ischemic In Inferolateral                    | ISCIL        |                            | MIS              |
| Subendocardial Injury In Lateral Leads       | INJLA        |                            | MIS              |
| Ischemic In Lateral Leads                    | ISCLA        |                            | MIS              |
| Lateral Ischaemia                            | LIS          |                            | MIS              |
| Subendocardial Injury In Anteroseptal Leads  | INJAS        |                            | MIS              |
| Ischemic In Anteroseptal Leads               | ISCAS        |                            | MIS              |
| Subendocardial Injury In Anterolateral Leads | INJAL        |                            | MIS              |
| Ischemic In Anterolateral Leads              | ISCAL        |                            | MIS              |
| Left Atrial Abnormality                      | LAA          |                            |                  |
| Left Atrial Hypertrophy                      | LAH          | LAO, LAE                   | AH               |
| Right Atrial Abnormality                     | RAAB         |                            |                  |
| Right Atrial Hypertrophy                     | RAH          | RAO, RAE                   | AH               |
| Left Ventricular Hypertrophy                 | LVH          |                            | VH               |

|                                                                |            |             |            |
|----------------------------------------------------------------|------------|-------------|------------|
| Voltage Criteria (QRS) For Left Ventricular Hypertrophy        | VCLVH      |             | VH         |
| Septal Hypertrophy                                             | SEHYP      |             | VH         |
| Right Ventricular Hypertrophy                                  | RVH        |             | VH         |
| ST-T Change Due To Ventricular Hypertrophy                     | STTVH      |             | VH         |
| Incomplete Right Bundle Branch Block                           | IRBBB      |             | BBB        |
| Complete Left Bundle Branch Block   Left Bundle Branch Block   | CLBBB LBBB | CLBBB, LBBB | BBB        |
| Incomplete Left Bundle Branch Block                            | ILBBB      |             | BBB        |
| Complete Right Bundle Branch Block   Right Bundle Branch Block | CRBBB RBBB | CRBBB, RBBB | BBB        |
| Transient Ischemic Attack                                      | TIA        |             |            |
| Atrial Hypertrophy                                             | AH         |             |            |
| Myocardial Infarction                                          | MI         |             |            |
| Myocardial Ischemia                                            | MIS        |             |            |
| Ventricular Hypertrophy                                        | VH         |             |            |
| Coronary Heart Disease                                         | CHD        |             |            |
| Chronic Myocardial Ischemia                                    | CMIS       | CMI         |            |
| Heart Failure                                                  | HF         |             |            |
| Heart Valve Disorder                                           | HVD        |             |            |
| Left Ventricular Strain                                        | LVS        |             |            |
| Counterclockwise Rotation                                      | -ROT       | CCR         | CVCL/CCVCL |
| Clockwise Rotation                                             | +ROT       | CR          | CVCL/CCVCL |
| Accelerated Atrial Escape Rhythm                               | AAR        |             |            |
| Atrial Bigeminy                                                | AB         |             |            |
| Abnormal QRS                                                   | ABQRS      |             |            |
| Atrial Escape Beat                                             | AED        |             |            |
| Atrial Fibrillation                                            | AF         | AFIB        |            |
| Atrial Fibrillation And Flutter                                | AFAFL      |             |            |
| Atrial Flutter                                                 | AFL        | AFLT        |            |
| Accelerated Idioventricular Rhythm                             | AIVR       |             |            |
| Accelerated Junctional Rhythm                                  | AJR        |             |            |
| Suspect Arm Ecg Leads Reversed                                 | ALR        |             |            |
| Atrial Pacing Pattern                                          | AP         |             |            |
| Atrial Rhythm                                                  | ARH        |             |            |
| Atrial Tachycardia                                             | ATACH      |             |            |
| Av Block                                                       | AVB        | _AVB        |            |
| Atrioventricular Dissociation                                  | AVD        |             |            |
| Atrioventricular Junctional Rhythm                             | AVJR       |             |            |
| Atrioventricular Node Reentrant Tachycardia                    | AVNRT      |             |            |
| Atrioventricular Reentrant Tachycardia                         | AVRT       |             |            |
| Bundle Branch Block                                            | BBB        |             |            |
| Blocked Premature Atrial Contraction                           | BPAC       |             |            |
| Bradycardia                                                    | BRADY      |             |            |
| Brugada                                                        | BRU        |             |            |
| Brady Tachy Syndrome                                           | BTS        |             |            |
| Cardiac Dysrhythmia                                            | CD         |             |            |

|                                                                 |            |                     |    |
|-----------------------------------------------------------------|------------|---------------------|----|
| Congenital Incomplete Atrioventricular Heart Block              | CIAHB      |                     |    |
| Clockwise Or Counterclockwise Vectorcardiographic Loop          | CVCL/CCVCL |                     |    |
| Diffuse Intraventricular Block                                  | DIB        |                     |    |
| Early Repolarization                                            | ERE        |                     |    |
| Fusion Beats                                                    | FB         |                     |    |
| Fqrs Wave                                                       | FQRS       |                     |    |
| High T-Voltage                                                  | HTV        |                     |    |
| Indeterminate Cardiac Axis                                      | ICA        |                     |    |
| Idioventricular Rhythm                                          | IR         |                     |    |
| Junctional Escape                                               | JE         |                     |    |
| Junctional Premature Complex                                    | JPC        |                     |    |
| Junctional Tachycardia                                          | JTACH      |                     |    |
| Left Axis Deviation                                             | LAD        |                     |    |
| Left Posterior Fascicular Block                                 | LPFB       |                     |    |
| Prolonged Pr Interval                                           | LPR        |                     |    |
| Low Qrs Voltages                                                | LQRSV      |                     |    |
| Prolonged Qt Interval                                           | LNGQT      | LQT                 |    |
| Left Ventricular High Voltage                                   | LVHV       |                     |    |
| Nonspecific Intraventricular Conduction Disorder                | NSIVCB     |                     |    |
| Normal Sinus Rhythm                                             | NORM       |                     |    |
| Nonspecific St T Abnormality                                    | NSSTTA     |                     |    |
| Old Myocardial Infarction                                       | OLDMI      |                     | MI |
| Premature Atrial Contraction   Supraventricular Premature Beats | PAC SVPB   | PAC, SVPB, SVPB PAC |    |
| Prolonged P Wave                                                | PPW        |                     |    |
| Pacing Rhythm                                                   | PR         |                     |    |
| Poor R Wave Progression                                         | PRWP       |                     |    |
| Paroxysmal Supraventricular Tachycardia                         | PSVT       |                     |    |
| Paroxysmal Ventricular Tachycardia                              | PVT        |                     |    |
| P Wave Change                                                   | PWC        |                     |    |
| Qwave Abnormal                                                  | QAB        | QAVE                |    |
| R Wave Abnormal                                                 | RAB        |                     |    |
| Right Axis Deviation                                            | RAD        |                     |    |
| Rapid Atrial Fibrillation                                       | RAF        |                     |    |
| Right Atrial High Voltage                                       | RAHV       |                     |    |
| Sinus Arrhythmia                                                | SARRH      | SA                  |    |
| Sinus Atrium To Atrial Wandering Rhythm                         | SAAWR      |                     |    |
| Sinoatrial Block                                                | SAB        |                     |    |
| Sinus Arrest                                                    | SARR       |                     |    |
| Sinus Bradycardia                                               | SBRAD      | SB                  |    |
| Sinus Node Dysfunction                                          | SND        |                     |    |
| Shortened Pr Interval                                           | SPRI       |                     |    |
| Decreased Qt Interval                                           | SQT        |                     |    |
| Sinus Tachycardia                                               | STACH      | ST                  |    |
| S T Changes                                                     | STC        |                     |    |
| St Depression                                                   | STD        | STD                 |    |
| St Interval Abnormal                                            | STIAB      |                     |    |
| Supraventricular Bigeminy                                       | SVB        |                     |    |

|                                                                     |           |               |          |
|---------------------------------------------------------------------|-----------|---------------|----------|
| Supraventricular Tachycardia                                        | SVT       | SVTACH        |          |
| T Wave Abnormal                                                     | TAB       | TAB_          |          |
| T Wave Inversion                                                    | TINV      | INVT          |          |
| Tall P Wave                                                         | TPW       |               |          |
| U Wave Abnormal                                                     | UAB       |               |          |
| Ventricular Bigeminy                                                | VBIG      |               |          |
| Ventricular Ectopics                                                | VEB       |               |          |
| Ventricular Escape Beat                                             | VESB      |               |          |
| Ventricular Escape Rhythm                                           | VESR      |               |          |
| Ventricular Fibrillation                                            | VF        |               |          |
| Ventricular Flutter                                                 | VFL       |               |          |
| Premature Ventricular Contractions  <br>Ventricular Premature Beats | VPC VPB   | VPC, VPB, PVC |          |
| Ventricular Pre Excitation                                          | VPEX      |               |          |
| Ventricular Pacing Pattern                                          | VPP       |               |          |
| Paired Ventricular Premature<br>Complexes                           | VPVC      |               |          |
| Ventricular Tachycardia                                             | VTACH     |               |          |
| Ventricular Trigeminy                                               | VTRIG     |               |          |
| Wandering Atrial Pacemaker                                          | WAP       |               |          |
| Wolff-Parkinson-White Pattern                                       | WPW       |               |          |
| Low Voltage                                                         | LVOLT     |               |          |
| TU fusion                                                           | TUF       |               |          |
| Atrial premature complexes non-<br>conducted                        | PAC_NC    |               | PAC SVPB |
| Av conduction ration N:D                                            | AVCR      |               |          |
| Left anterior fascicular block                                      | LAFB      | LANFB         |          |
| Junctional escape complex(es)                                       | JEC       |               |          |
| ST deviation with T-wave change                                     | STTC      |               |          |
| Left anterior/posterior fascicular<br>block                         | LAFB/LPFB |               |          |
| Non specific intraventricular<br>conduction disturbance             | IVCD      |               |          |
| Non specific t wave changes                                         | NT_       |               |          |
| Sinus rhythm                                                        | SR        |               |          |
| Digitalis effect                                                    | DIG       |               |          |
| Premature complex(es)                                               | PRC(S)    |               |          |
| Supraventricular arrhythmia                                         | SVARR     |               |          |
| Trigeminal pattern (unknown origin<br>SV or Ventricular)            | TRIGU     |               |          |
| Low amplitude t waves                                               | LOWT      |               |          |
| Electrolytic disturbance or drug<br>(former EDIS)                   | EL        |               |          |
| Bigeminal pattern (unknown origin<br>SV or Ventricular)             | BIGU      |               |          |
| Normal functioning artificial<br>pacemaker                          | PACE      |               |          |
| Non-diagnostic t abnormalities                                      | NDT       |               |          |
| ST-T changes compatible with<br>ventricular aneurysm                | ANEUR     |               |          |
| Non specific ST changes                                             | NST_      |               |          |
| Non specific ischemic                                               | ISC       |               |          |
| Hypertrophy                                                         | HYP       |               |          |

|                  |       |  |  |
|------------------|-------|--|--|
| High qrs voltage | HVOLT |  |  |
|------------------|-------|--|--|

**Supplementary Material – Appendix C – Table 1:** This table lists the harmonised diagnoses and the corresponding abbreviations used throughout this study. For conditions that appear with different names across various datasets, the table also includes found synonyms. When generating the Cardio-Learning dataset, condition appearing under various synonyms are standardised to a unique name. The ‘Super-conditions’ column indicates hierarchically broader diagnostic categories. When creating the Cardio-Learning dataset, if broader and more specific conditions come from different constituent datasets, the broader conditions augment the annotation of an ECG that is labelled with the more specific condition. If an ECG is labelled with a more general condition, its annotation is not augmented with a more specific diagnostic category, unless explicitly confirmed in the original annotation. Importantly, these hierarchical augmentations are performed only when the hierarchy involves diagnostic labels coming from different constituent datasets. This means that if a more specific condition and its broader diagnostic category appear in the same dataset, but an ECG from that source is explicitly labelled only the more specific one, the hierarchical mapping is not performed to respect the distinction made by the original annotators.

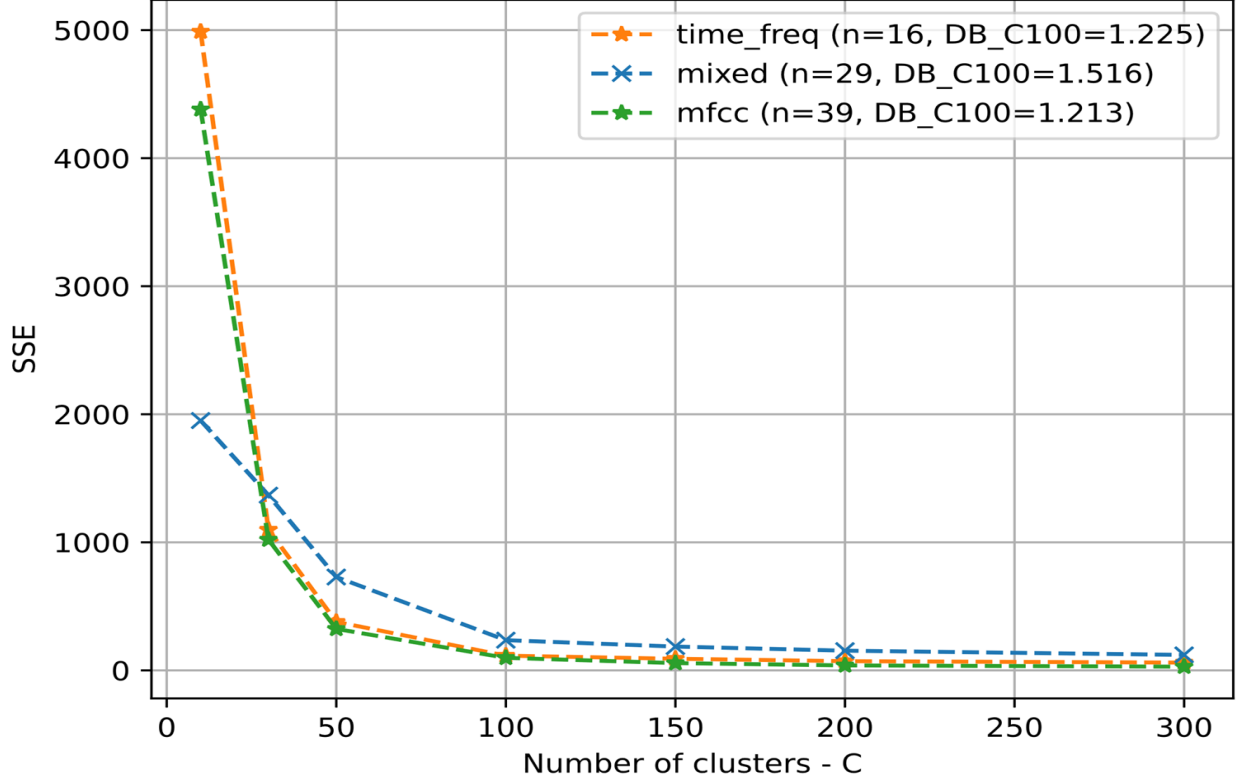
## 1 Pre-training analysis

This section presents an analysis of how pre-training with different ECG feature descriptors and sampling rates impact downstream performance. We then investigate the interesting effects of using our dynamic regularisation during pre-training. Finally, we examine the clustering quality across both HuBERT-ECG encoding layers and pre-training iterations to measure their downstream impact. To these aims, we pre-train HuBERT-ECG BASE for one iteration with fixed hyperparameters and running configuration that are described in each of the following sections. Then, to evaluate the impact of different choices, we simply perform linear evaluations, i.e., training of a randomly initialised linear layer atop a frozen pre-trained model, on a development set that we extract from the labelled dataset used by Ribeiro et al. [126]. We build such a set, hereafter referred to as *Ribeiro-dev*, by deliberately excluding all normal ECGs, as we are more interested in assessing the ability to detect cardiac abnormalities rather than normal ECGs. Linear evaluation results represent a conservative estimate, as they are derived using fewer trainable parameters and development set that include more than 300k instances.

### 1.1 Exploring feature descriptors for ECG segments

In order to identify the most appropriate ECG segment descriptor to use in the initial pre-training iteration, we perform k-means clustering on the entire training set with 3 different feature descriptors extracted from the ECG segments. It is important to mention that we choose k-means over other clustering approaches (e.g. DBSCAN, hierarchical clustering, Gaussian Mixture models) for its computational efficiency and scalability, two crucial ingredients when working in large-scale setups as ours. The first descriptor takes into account 16 simple time-frequency features, while the second one is based on 39 Mel Frequency Cepstral coefficients[180] (MFCCs) following Hsu et al. [32]. These descriptors have been successfully used in ECG analysis [181], [182]. The third descriptor combines the first 13 MFCCs and all the previous time-frequency features. For each descriptor, we run k-means with 10, 30, 50, 100, 150, 200 and 300 clusters and compute the corresponding sum of squared errors (SSE), or inertia (Supplementary Material – Appendix C - Figure 1). These clusters correspond to the pseudo-labels used as target variable during the self-supervised pre-training of HuBERT-ECG. We observe that MFCC- and time-frequency based descriptors provide comparable inertia and a global mean squared error in the order of  $1e-5$ , and thus appear as promising candidates. Following the elbow method [183], and driven by the intuition that more clusters capture too fine and specific ECG patterns for the first pre-training iteration, we set  $C = 100$  as the best number of clusters to use irrespective of the descriptor. Given this number of clusters, we compute the Davies-Bouldin [184] index to have an additional selection criterion and obtain extremely low scores. Interestingly,

while clustering on MFCCs yields higher inertia than clustering on time-frequency features, clusters on the former are slightly more compact, as suggested by a marginally lower Davies-Bouldin score. Therefore, we experiment with the use of both MFCC- and time-frequency based k-means models during pre-training and, then, linearly evaluate the resulting models on *Ribeiro-dev* until performance plateau or a maximum number of 80k training steps is reached. The linear evaluation results, presented in Supplementary Material – Appendix C -Table 2, show that pre-training with pseudo-labels generated by an MFCC-based k-means model yields better downstream performance.



**Supplementary Material – Appendix C - Figure 1:** Performance comparison of multiple k-means clustering runs in terms of Sum of Squared Errors (SSE). Once fixed the optimal number of clusters ( $C = 100$ ), the Davies-Bouldin index (DB) is computed and reported as “DB\_C100”. Lower DB values indicate better clustering.

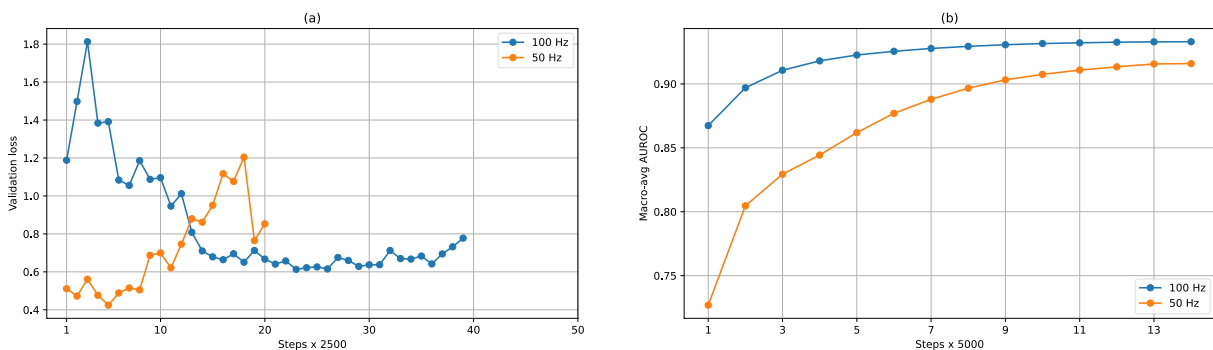
| Label generator                          | Macro-averaged AUROC |
|------------------------------------------|----------------------|
| MFCC-based k-means – $C = 100$           | 0.933                |
| Time-frequency based k-means – $C = 100$ | 0.913                |

**Supplementary Material – Appendix C - Table 2:** Linear evaluation performance of pre-trained HuBERT-ECG BASE models when using pseudo-labels generated by MFCC- and time-frequency based k-means models.

## 1.2 How the sampling rate affects downstream performance

The sampling rate of an ECG depends on the specific settings of the machine recording the electrical activity of the heart and there is no standard practice in this regard. For example, ECGs from the Ribeiro et al. [126] dataset are sampled at 400 Hz, while ECGs from PTB-XL [129] and SPH [133] are sampled at 400 and 500 Hz, respectively. We investigate the effect that the ECGs sampling rate

has on downstream performance. This parameter is important as it regulates the dimensionality of the data, therefore the computation speed and the memory occupation, and the degree of dilution of the information content. While it is true that training with ECGs sampled at lower frequencies is computationally less expensive, it is also true that sampling ECGs at lower frequencies may not provide enough samples to capture significant features and nuances in the input signal. Increasing the sampling rate may remedy this issue, at the cost of slower training and higher memory occupation. However, an oversampled ECG may contain redundant samples that overly dilute the information content and may be perceived as noise by a deep learning model. For these reasons, after band-pass filtering our ECGs to exclude frequencies outside  $[0.05, 47]$  Hz, which is reported to contain the dominant components of P waves, T waves and QRS complexes [145], we investigate the effects that sampling a 12-lead ECG at 50 and 100 Hz on linear evaluation performance over *Ribeiro-dev*. For these experiments, to determine a suitable sampling rate to work with, we linearly evaluate pre-trained models until the performance plateau or 80k training steps are performed. To always have the same number of embeddings being processed by the Transformer encoder, we design waveform convolutional embedders that become increasingly shallow as the sampling rate decreases, while also maintaining the same number of ECG segments. All the other design choices and running configurations are fixed across the experiments, as shown in Supplementary Material – Appendix C - Table 3. Supplementary Material – Appendix B - Figure 2a illustrates the validation loss curves observed during pre-training at 50 and 100 Hz while Supplementary Material – Appendix B – Figure 2b shows the corresponding linear evaluation performance on *Ribeiro-dev*. Two discernible trends are evident from the former: as the sampling rate decreases, the validation loss curves start at lower values than those corresponding to higher sampling rates; furthermore, as the sampling rate decreases, the models tend to show earlier overfitting to the training data. We suspect that at lower sampling rates, the information needed to solve the upstream task becomes more readily accessible, thereby facilitating the learning process. However, when clinical downstream labels are introduced, there is a noticeable performance discrepancy between models operating at 50 Hz and those operating at twice this sampling rate (Supplementary Material – Appendix C - Figure 2b and Supplementary Material – Appendix C - Table 4). We attribute this discrepancy to the insufficient number of samples available to capture condition-related patterns. In summary, sampling at 100 Hz captures all the meaningful physiological information, satisfies the Nyquist-Shannon theorem, and represents a desirable trade-off between computational cost and accurate downstream performance. All results are shown in Supplementary Material – Appendix C - Table 4.



**Supplementary Material – Appendix C - Figure 2:** Pre-training and linear evaluation performance with varying sampling rate. **(a)** Validation loss curves obtained when pre-training HuBERT-ECG with ECGs sampled at 50 and 100 Hz. **(b)** Downstream performance obtained when HuBERT-ECG is linearly evaluated on *Ribeiro-dev* after being pre-trained with ECGs sampled at 50 and 100 Hz.

### 1.3 The effects of the dynamic regularisation

In order to gain a deeper insight into the impact of dynamic regularisation on training time and the time required to tune regularisation terms, including weight decay and dropout probabilities, we consider four different pre-training setups, each followed by a linear evaluation to assess the actual benefit on downstream performance. The first and the second ones consist in performing the entire first pre-training iteration (260k steps with no early stopping) with a weight decay of 0.01 and a dropout probability of 0.1 (referred to as *default*), with and without dynamic regularisation, respectively. The last two setups, instead, consist in performing the same pre-training iteration, with and without dynamic regularisation, but setting the initial weight decay and dropout probability to the maximum values found by the dynamic regularisation. All experiments consider the BASE architecture, ECGs sampled at 100 Hz, a MFCC-based k-means as pseudo-label generator ( $C = 100$ ), a batch size of 448 instances and a masking percentage  $p = 33\%$ . The results of our experiments, presented in Supplementary Material – Appendix C - Table 5 and labelled with letters for clarity, reveal interesting insights. When using dynamic regularisation with default weight decay and dropout probability (setup A), HuBERT-ECG converges in 80k steps, resulting in the best macro-averaged AUROC of 0.933 during linear evaluation. In contrast, disabling the dynamic regularisation while maintaining the default regularisation terms (setup B), significantly slows down the convergence and leads to inferior downstream results. Instead, when comparing setups C and D, we observe two opposite behaviours: setup C shows faster convergence but worse upstream performance, while setup D shows slower convergence but better upstream performance. Despite these discordant trends, both setups produce close downstream results, both surpassing those from setup B, but still falling short of those of setup A. This suggests that pre-training with the dynamic regularisation, or the maximum regularisation terms it finds, accelerate pre-training and allows the model to adapt itself to avoid overfitting, resulting in improved linear evaluation when compared to scenarios where the dynamic regularisation is not used at all. Nevertheless, we hypothesise that initiating pre-training with already high regularisation terms, irrespective of whether dynamic regularisation is employed or not (setups C and D), may impair the model's learning capability at the most crucial stage, resulting in inferior, albeit marginal, downstream performance. To investigate into the source of these benefits, it seems plausible that a dynamic dropout rate encourages the model to upweight alternative hidden paths in the Transformer's feed-forward blocks when necessary, without permanently excluding the dropped ones. In such cases, when also the weight decay increases, this complementary form of regularisation prevents significant changes to the weights of those paths, thereby maintaining training stability.

| Architectural outline   | Sampling rate                                                      |                                                                       |
|-------------------------|--------------------------------------------------------------------|-----------------------------------------------------------------------|
|                         | 50 Hz                                                              | 100 Hz                                                                |
| Convolutional embedder  | kernels = (10, 3, 3, 2)<br>strides= (4, 2, 2, 2)<br>channels = 512 | kernels = (10, 3, 3, 2)<br>strides= (4, 2, 2, 2, 2)<br>channels = 512 |
| Transformer encoder     | BASE                                                               |                                                                       |
| k-means                 | MFCC, $C = 100$                                                    |                                                                       |
| Pre-training steps      | 130k                                                               |                                                                       |
| Linear evaluation steps | 80k                                                                |                                                                       |
| Batch size              | 448                                                                |                                                                       |
| P                       | 33%                                                                |                                                                       |
| LayerDrop               | 0.1                                                                |                                                                       |
| Dropout                 | 0.1                                                                |                                                                       |
| Weight decay            | 0.01                                                               |                                                                       |

**Supplementary Material – Appendix C – Table 3:** Architecture design and running configuration when experimenting with ECGs sampled at 50 and 100 Hz.

| ECG sampling rate | macro-averaged AUROC | Linear evaluation steps at plateau |
|-------------------|----------------------|------------------------------------|
| 50 Hz             | 0.9176               | 65k                                |
| 100 Hz            | 0.9330               |                                    |

**Supplementary Material – Appendix C – Table 4:** Linear evaluation performance of HuBERT-ECG on *Ribeiro-dev* after being pre-trained at 50 and 100 Hz. Linear evaluation steps necessary to let validation AUROC plateau are also reported.

| Setups                                                    | Pre-training steps | Pre-training<br>(validation loss, validation accuracy) | macro-averaged AUROC |
|-----------------------------------------------------------|--------------------|--------------------------------------------------------|----------------------|
| Dynamic regularisation & Default regularisation Terms     | 80k                | 0.6379, 0.8205                                         | 0.933                |
| No Dynamic regularisation & Default regularisation Terms  | 235k               | 0.6493, 0.8182                                         | 0.921                |
| (C) Dynamic regularisation & Maximum regularisation Terms | 72.5k              | 0.6775, 0.8111                                         | 0.9263               |
| No Dynamic regularisation & Maximum regularisation Terms  | 90k                | 0.5860, 0.8312                                         | 0.9244               |

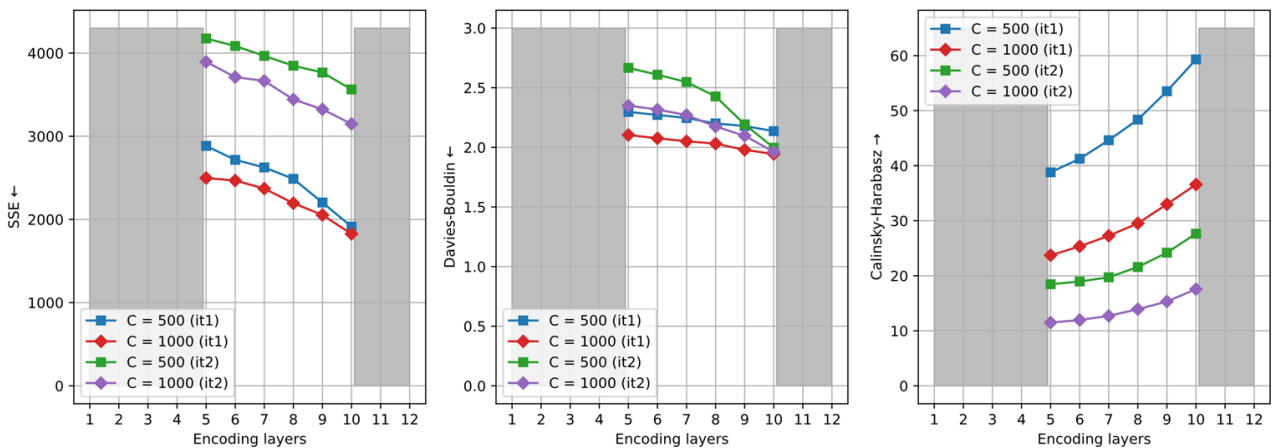
**Supplementary Material – Appendix C – Table 5:** Effects of the use of dynamic regularisation when pre-training HuBERT-ECG in terms of pre-training steps and upstream performance under the form of validation loss and validation accuracy. For each setup, linear evaluation performance on *Ribeiro-dev* is also reported.

## 1.4 The downstream impact of clustering quality across encoding layers and iterations

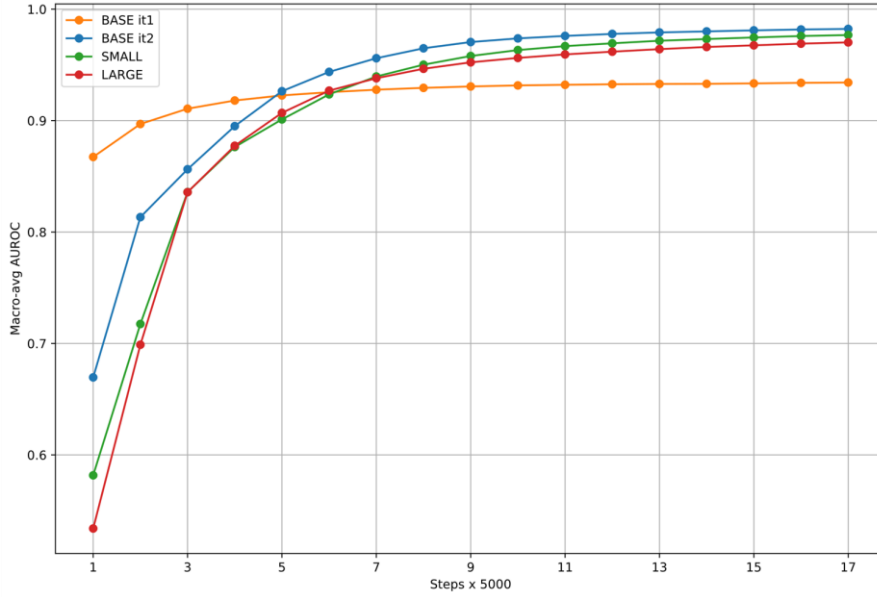
In the work of Hsu et al. [32], prior to the second and third pre-training iterations, the most suitable number of clusters was determined empirically, as well as the layer from which latent features had to be extracted. This was achieved through the analysis of the quality of clusters of hidden representations extracted from each model layer. In particular, an automatic-speech-recognition model was used to produce frame-level phonetic labels that served as targets to measure the correctness of forced-aligned cluster assignments. The same approach is not applicable to our case as, to the best of our knowledge, there is no powerful open-source ECG model that can produce frame-level forced-aligned cluster assignments. Consequently, to determine which layer’s latent features should be clustered and how many clusters are to be found, we make use once again of traditional compactness and separability metrics. In addition, to limit the set of layers to explore, we combine such metrics with findings from the NLP domain regarding layers transferability [178]. After sampling 10% of pre-training ECGs, we cluster their segments’ latent representations from the 5<sup>th</sup> – 10<sup>th</sup> Transformer layer and measure the clustering quality in terms of inertia, Davies-Bouldin index and Calinsky-Harabasz [185] index. We exclude from consideration shallower layers on the grounds that they would produce too coarse representations to be of use in clustering. Similarly, deeper layers are excluded on the grounds that they would generate representations that would lead to too task-

specific pseudo-labels for a generic pre-training. Upon completion of the first iteration, we identify 500 and 1000 clusters of latent features from each of these layers, since we consider that setting a higher number of clusters than that used in the first iteration is necessary for two reasons: 1) to try to generate much finer cluster assignments; 2) to avoid measuring, through clustering, the degree of separability of the classes learnt during the previous iteration. The results of this final analysis are presented in Supplementary Material – Appendix C - Figure 3. As can be seen, each metric shows a monotonic trend, regardless of the representations being clustered. Therefore, they do not provide an unequivocal indication of the best encoding layer to extract features from. To continue the pre-training of the BASE configuration, we then choose to cluster latent representations from the 8<sup>th</sup> layer into 500 clusters, as this point seems to mark a non-negligible change in the metrics we consider. After completing the second iteration, we re-introduce diagnostic labels and perform a linear evaluation with the same running configurations reported in previous paragraphs to quantify the relative improvement with respect to previous evaluations. As shown in Supplementary Material – Appendix C - Figure 4, when the first-iteration model saturates, the second-iteration one achieves results that are approximately 5 AUROC points better and still has room for improvement.

Once HuBERT-ECG BASE completes the second iteration, we repeat the clustering step described above and report the results in Supplementary Material – Appendix C - Figure 3. Once again, the metrics we consider show a monotonic trend that does not facilitate the choice of an acceptable number of clusters (500 or 1000) nor which layer the latent representations should be extracted from. For these reasons, to start pre-training the SMALL and LARGE model configurations, we decide to extract latent features from the 9<sup>th</sup> layer and to cluster them into 500 and 1000 clusters, respectively. We believe that pre-training these third-iteration models requires, for any configuration, to select a deeper layer than those selected for the previous iterations. Furthermore, increasing the quantity and fineness of cluster assignments can be beneficial for a model with higher learning capacity such as HuBERT-ECG LARGE. Conversely, for the SMALL model size, maintaining the same number and granularity of cluster assignments can facilitate the task of mimicking the BASE model without significant loss of performance. The benefits of these choices are also displayed in Supplementary Material – Appendix C - Figure 4. Upon completion of 362.5k- and 422.5k-step pre-trainings, the SMALL and LARGE configurations perform similarly when linearly evaluated on *Ribeiro-dev* for 80k steps: both achieve a macro-averaged AUROC slightly lower than that of the BASE configuration, but show no sign of strong saturation within the training time of this linear evaluation.



**Supplementary Material – Appendix C - Figure 3:** Clustering quality metrics (inertia, Davies-Bouldin index, Calinsky-Harabasz index) across HuBERT-ECG BASE encoding layers after the first (it1) and second (it2) pre-training iterations. Symbols  $\uparrow$  and  $\downarrow$  indicate whether a metric needs to be maximised or minimised, respectively. Gray-shaded regions refer to encoding layers not considered in this analysis.



**Supplementary Material – Appendix C - Figure 4:** Linear evaluation performance of HuBERT-ECG BASE, after first and second pre-training iteration, HuBERT-ECG SMALL and HuBERT-ECG LARGE on *Ribeiro-dev*.

## 2 Ablation study

In this section, we present a short sequence of ablation studies to investigate the effects of specific architectural choices and the impact of important hyperparameters. In particular, we study (1) the impact of our masking strategy, and (2) the effects of multi-task learning when experimenting with multiple cluster ensembles. To do so we perform multiple pre-trainings of HuBERT-ECG BASE (first iteration only) followed by linear evaluations on *Ribeiro-dev* with fixed hyperparameters. If not otherwise mentioned, the experimental configurations are the same of the previous section.

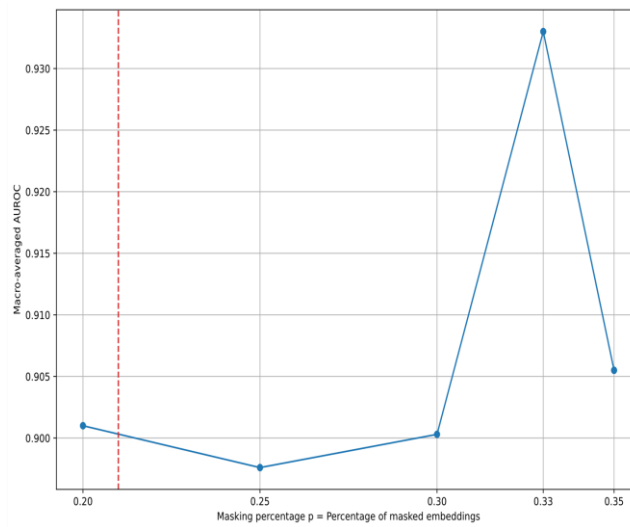
### 2.1 Impacts of the masking strategy

We consider setting the value of the masking percentage  $p$  of crucial importance to make HuBERT-ECG learn high quality representations of 12-lead ECGs. An excessively low value would generate a trivial upstream task, while an exaggeratedly high value would result into a nearly impossible one. For this reason, we experiment with multiple values of this hyperparameter in order to see how it impacts on downstream performance and plot the results of these experiments in Supplementary Material – Appendix C - Figure 5. For the sake of comparison, we also plot the exact percentage of embeddings that are masked when the masking strategy proposed by Hsu et al. [32] is used. Beyond some statistical and label noise, what emerges clearly is that setting  $p = 33\%$  guarantees the best results, as linear evaluation performance peaks at this masking percentage, while following the masking strategy used to pre-train HuBERT leads to suboptimal performance. We believe this finding is due to the more regular patterns and high information redundancy of ECGs compared to audio signals.

### 2.2 The effects of multi-task learning

To observe the downstream effects of pre-training HuBERT-ECG in a multi-task learning framework, we perform three pre-trainings with an increasing number of tasks to solve. Initially, we pre-train HuBERT-ECG using pseudo-labels generated by a single MFCC-based k-means model with 100 clusters. Then, we pre-train again with an additional clustering model with 200 clusters and,

eventually, we consider an ensemble of three k-means models with 100, 200 and 300 clusters. Supplementary Material – Appendix C - Table 6 reports the attained results. Interestingly, adding just a new clustering model does not change pre-training nor affects downstream performance. In contrast, although marginal, we see an improvement when considering an ensemble of three k-means models providing much more granular pseudo-labels. Since solving three tasks at once is harder than solving just one of them, it is not surprising to see that such performance gains follow a much longer pre-training. We also believe that the cardinality of the ensemble is not as relevant as the maximum number of clusters, which is likely more useful to the model to capture label-specific patterns in downstream data. This is analogous to the refinement of the cluster assignments that is performed prior to the second iteration, except for how such refinement is achieved, since both refinements aim to generate finer and more granular pseudo-labels for the ECG segments. Considering both the longer pre-training time and the marginal improvement obtained on the results reached after training with pseudo-labels generated by a single clustering model, we do not proceed in pre-training with cluster ensembles.



**Supplementary Material – Appendix C - Figure 5:** HuBERT-ECG BASE linear evaluation performance after pre-training with different values of the hyper-parameter  $p$  (i.e. the percentage of ECG embeddings to mask). The red dashed line indicates the percentage of ECG embeddings that would be masked if we followed the masking strategy used to pre-train HuBERT.

| K-means ensemble                        | Macro-averaged AUROC | Pre-training steps |
|-----------------------------------------|----------------------|--------------------|
| MFCC-based k-means, $C = 100$           | 0.933                | 80k                |
| MFCC-based k-means, $C = 100, 200$      | 0.930                | 75k                |
| MFCC-based k-means, $C = 100, 200, 300$ | 0.943                | 135k               |

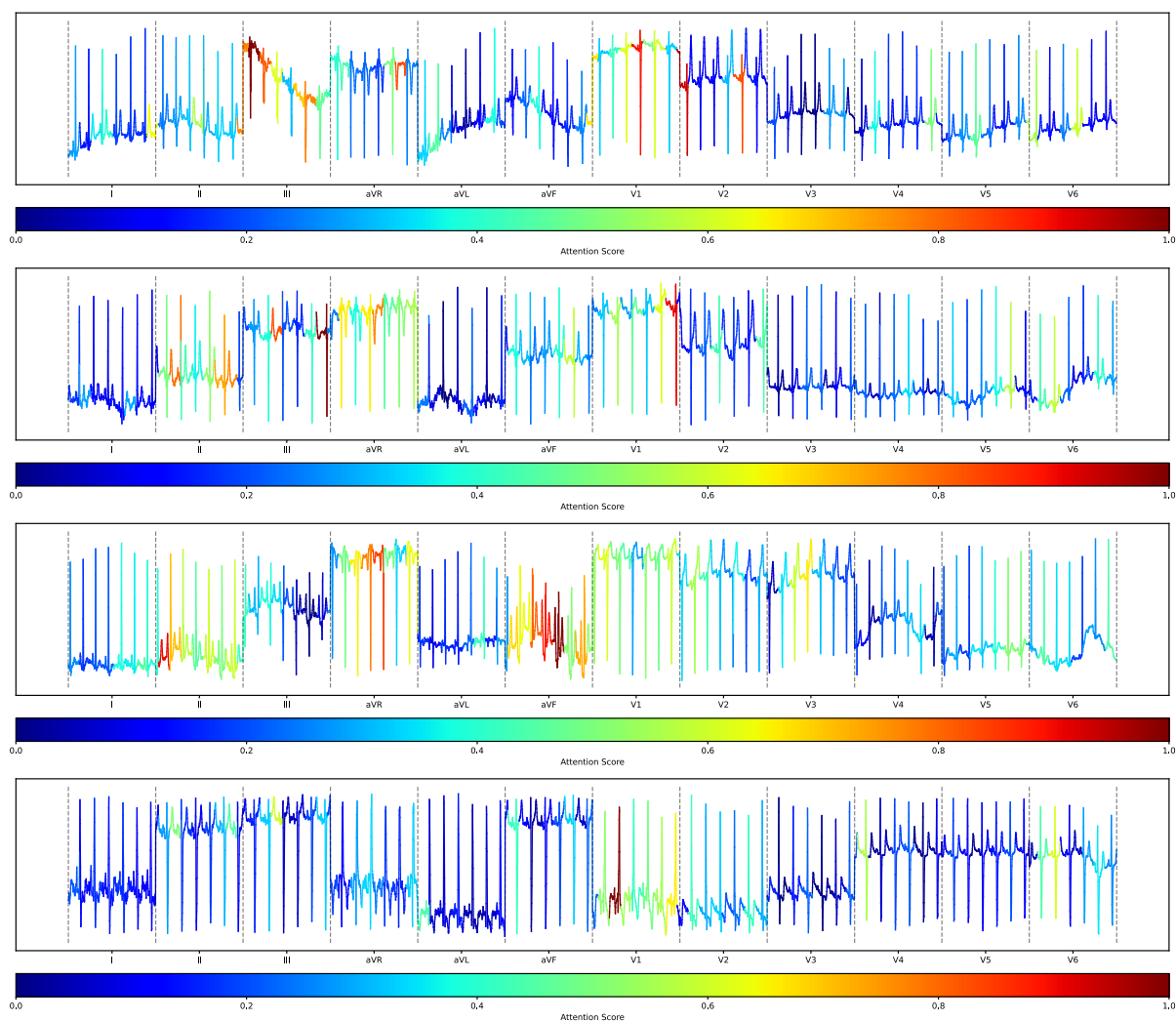
**Supplementary Material – Appendix C – Table 6:** Linear evaluation performance of HuBERT-ECG BASE after being pre-trained in a multi-task learning framework in which tasks are represented by multiple k-means models composing an ensemble of label generators. The corresponding number of pre-training steps is reported for every task/cluster ensemble.

### 3 Explaining fine-tuned diagnostic models

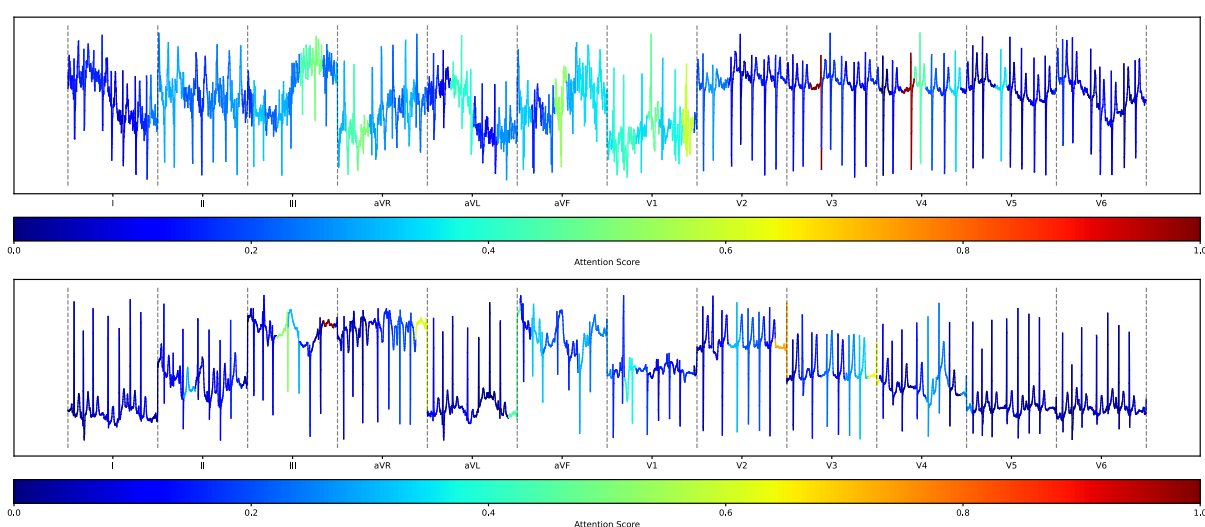
The self-supervised pre-trained HuBERT-ECG model has no knowledge of diseases or their manifestations in ECG signals. During pretraining, the model learns only the structural and statistical properties of the signal, encoding them into general-purpose representations [24]. This learning process is guided by the pre-training task of predicting pseudo-labels—generated in an unsupervised, disease-unrelated way—for masked ECG embeddings. As such, the representations learned are not directly aligned with clinically useful, human-interpretable ECG characteristics. Because of this, applying explainability techniques directly to the pretrained model would yield explanations that are clinically uninformative: the internal features are not intended to capture pathophysiological phenomena. In contrast, once HuBERT-ECG is fine-tuned on a supervised diagnostic task, its representations become aligned with disease-specific patterns, making explainability techniques more meaningful.

In this section, we apply explainability methods to a supervised fine-tuned version of HuBERT-ECG. Specifically, we use attention maps [186], which are well-suited to Transformer-based architectures, to visualize which parts of the ECG the model focuses on when making predictions. In Supplementary Material – Appendix C - Figure 6, we present four ECGs from the PTB-XL dataset labelled as Sinus Rhythm and correctly classified by the model. The attention maps show that the model focuses on multiple regions of the signal, with some variability across recordings. This variability is expected and desirable, reflecting inter-patient differences and signal-specific features that HuBERT-ECG learns during fine-tuning and leverages when predicting the output condition. In Supplementary Material – Appendix C - Figure 7, we show two ECGs labelled as Atrial Fibrillation, also correctly detected. In these cases, the attention maps highlight clinically relevant signal segments, suggesting that the model is focusing on meaningful characteristics rather than spurious patterns or noise.

While these visualizations do not provide a direct one-to-one mapping with classical ECG features (e.g., P-wave morphology or QRS duration), they offer qualitative evidence that the fine-tuned model attends to diagnostic signal regions in a task-consistent manner. The observed variability in attention patterns reflects the model’s flexibility in handling heterogeneous ECG presentations, rather than a lack of intelligent feature extraction.



**Supplementary Material – Appendix C – Figure 6:** Attention maps on predictions formulated by a fine-tuned HuBERT-ECG for correctly classified Sinus Rhythm ECGs



**Supplementary Material – Appendix C - Figure 6:** Attention maps on predictions formulated by a fine-tuned HuBERT-ECG for correctly classified Atrial Fibrillation ECGs

| Condition      | Models              | Sensitivity                  | Specificity                  | AUROC                               | AUPRC                               |
|----------------|---------------------|------------------------------|------------------------------|-------------------------------------|-------------------------------------|
| AF             | Ribeiro et al.      | 0.769                        | 1.000                        | 0.885                               | 0.773                               |
|                | ECGFounder          | 1.000                        | 0.996                        | 0.999                               | N.A                                 |
|                | HuBERT-ECG<br>SMALL | 1.0000<br>(CI:0.9999–1.0000) | 0.9754<br>(CI:0.9656–0.9853) | 1.0000<br>(CI:0.9986–1.0000)        | 0.9774<br>(CI:0.9278–1.0000)        |
|                | HuBERT-ECG<br>BASE  | 1.0000<br>(CI:0.999–1.000)   | 1.0000<br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) |
|                | HuBERT-ECG<br>LARGE | 1.0000<br>(CI:0.9999–1.000)  | 0.9748<br>(CI:0.9644–0.9853) | 0.9976<br>(CI:0.9954–0.9999)        | 0.9240<br>(CI:0.8535–0.9945)        |
| IAVB           | Ribeiro et al.      | 0.929                        | 0.995                        | 0.962                               | 0.807                               |
|                | ECGFounder          | 0.864                        | 0.957                        | 0.949                               | N.A.                                |
|                | HuBERT-ECG<br>SMALL | 0.9290<br>(CI:0.8580–1.0000) | 0.9531<br>(CI:0.9387–0.9675) | 0.9890<br>(CI:0.9817–0.9963)        | 0.8309<br>(CI:0.7366–0.9252)        |
|                | HuBERT-ECG<br>BASE  | 0.7500<br>(CI:0.6071–0.8929) | 0.9981<br>(CI:0.9962–1.0000) | <b>0.9977</b><br>(CI:0.9961–0.9994) | <b>0.9560</b><br>(CI:0.9264–0.9856) |
|                | HuBERT-ECG<br>LARGE | 1.0000<br>(CI:0.9990–1.0000) | 0.9485<br>(CI:0.9337–0.9637) | 0.9960<br>(CI:0.9935–0.9985)        | 0.8750<br>(CI:0.7929–0.9571)        |
| CLBBB LBBB     | Ribeiro et al.      | 1.000                        | 1.000                        | 1.000                               | 1.000                               |
|                | ECGFounder          | 1.000                        | 0.996                        | 0.998                               | N.A.                                |
|                | HuBERT-ECG<br>SMALL | 1.0000<br>(CI:0.9999–1.0000) | 0.9940<br>(CI:0.9893–0.9987) | 1.0000<br>(CI:0.9999–1.0000)        | 0.9950<br>(CI:0.9837–1.0000)        |
|                | HuBERT-ECG<br>BASE  | 0.9000<br>(CI:0.8333–0.9667) | 1.0000<br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) | <b>1.0000</b><br>(CI:0.9999–1.0000) |
|                | HuBERT-ECG<br>LARGE | 0.9330<br>(CI:0.8993–0.9667) | 0.9950<br>(CI:0.9913–0.9987) | 0.9983<br>(CI:0.9971–0.9995)        | 0.9760<br>(CI:0.9609–0.9911)        |
| CRBBB RBBB     | Ribeiro et al.      | 1.000                        | 0.995                        | 0.997                               | 0.895                               |
|                | ECGFounder          | 0.971                        | 0.971                        | 0.989                               | N.A.                                |
|                | HuBERT-ECG<br>SMALL | 1.0000<br>(CI:0.9999–1.000)  | 0.9850<br>(CI:0.9776–0.9924) | <b>0.9987</b><br>(CI:0.9976–0.9998) | 0.9800<br>(CI:0.9616–0.9984)        |
|                | HuBERT-ECG<br>BASE  | 0.9400<br>(CI:0.8529–1.0000) | 0.9943<br>(CI:0.9899–0.9987) | 0.9986<br>(CI:0.9974–0.9998)        | <b>0.9740</b><br>(CI:0.9513–0.9967) |
|                | HuBERT-ECG<br>LARGE | 0.9710<br>(CI:0.9420–1.0000) | 0.9874<br>(CI:0.9798–0.9950) | 0.9977<br>(CI:0.9957–0.9998)        | 0.9690<br>(CI:0.9424–0.9956)        |
| SBRAD          | Ribeiro et al.      | 0.938                        | 0.996                        | 0.967                               | 0.782                               |
|                | ECGFounder          | 1.000                        | 0.955                        | 0.967                               | N.A.                                |
|                | HuBERT-ECG<br>SMALL | 1.0000<br>(CI:0.9999–1.0000) | 0.9808<br>(CI:0.9716–0.9901) | 0.9973<br>(CI:0.9948–0.9998)        | 0.8801<br>(CI:0.7719–0.9883)        |
|                | HuBERT-ECG<br>BASE  | 1.0000<br>(CI:0.9999–1.0000) | 0.9919<br>(CI:0.9864–0.9975) | <b>0.9980</b><br>(CI:0.9954–1.0000) | 0.8860<br>(CI:0.7504–1.0000)        |
|                | HuBERT-ECG<br>LARGE | 1.0000<br>(CI:0.9999–1.0000) | 0.9796<br>(CI:0.9704–0.9889) | 0.9980<br>(CI:0.9962–0.9998)        | <b>0.8950</b><br>(CI:0.8014–0.9886) |
| STACH          | Ribeiro et al.      | 0.937                        | 0.997                        | 0.985                               | 0.923                               |
|                | ECGFounder          | 0.974                        | 0.970                        | 0.989                               | N.A.                                |
|                | HuBERT-ECG<br>SMALL | 0.9730<br>(CI:0.9729–0.9731) | 0.9816<br>(CI:0.9722–0.9911) | 0.9969<br>(CI:0.9946–0.9992)        | 0.9350<br>(CI:0.8807–0.9893)        |
|                | HuBERT-ECG<br>BASE  | 0.9730<br>(CI:0.9729–0.9731) | 0.9950<br>(CI:0.9913–0.9987) | <b>0.9990</b><br>(CI:0.9983–0.9997) | <b>0.9640</b><br>(CI:0.9342–0.9938) |
|                | HuBERT-ECG<br>LARGE | 0.9730<br>(CI:0.9728–0.9732) | 0.9829<br>(CI:0.9734–0.9924) | 0.9957<br>(CI:0.9939–0.9975)        | 0.9480<br>(CI:0.9149–0.9811)        |
| MACRO-AVERAGED | Ribeiro et al.      | 0.935                        | 0.997                        | 0.966                               | 0.863                               |
|                | ECGFounder          | 0.9682                       | 0.9742                       | 0.9818                              | N.A.                                |
|                | HuBERT-ECG<br>SMALL | 0.9806<br>(CI:0.9657–0.9955) | 0.9784<br>(CI:0.9740–0.9828) | 0.9974<br>(CI:0.9964–0.9984)        | 0.9330<br>(CI:0.9045–0.9615)        |
|                | HuBERT-ECG<br>BASE  | 0.9255<br>(CI:0.8943–0.9567) | 0.9969<br>(CI:0.9954–0.9983) | <b>0.9989</b><br>(CI:0.9983–0.9995) | <b>0.9630</b><br>(CI:0.9403–0.9857) |
|                | HuBERT-ECG<br>LARGE | 0.9770<br>(CI:0.9641–0.9899) | 0.9782<br>(CI:0.9740–0.9825) | 0.9978<br>(CI:0.9973–0.9983)        | 0.9310<br>(CI:0.9013–0.9607)        |

**Supplementary Material – Appendix C – Table 7:** Comparison of all HuBERT-ECG configurations with existing models from the literature on Ribeiro test set according to multiple metrics. Sensitivity and specificity are obtained setting the threshold to 0.5, as done by Ribeiro et al., for a fair comparison. Performance uncertainty is reported using 95% confidence intervals computed via empirical bootstrapping (N=1000). Label abbreviations and corresponding diagnosis can be found in Supplementary Material – Appendix C - Table 1.

| Models                                       | PTB-XL<br>All | PTB-XL<br>Form | PTB-XL<br>Rhythm | PTB-XL<br>Diag. | PTB-XL<br>Diag.<br>Subclass | PTB-XL<br>Diag.<br>Superclass |
|----------------------------------------------|---------------|----------------|------------------|-----------------|-----------------------------|-------------------------------|
| Macro-averaged AUROC                         |               |                |                  |                 |                             |                               |
| HuBERT-ECG<br>SMALL                          | 0.900±0.006   | 0.838±0.012    | 0.941±0.013      | 0.919±0.005     | 0.914±0.005                 | 0.907±0.007                   |
| HuBERT-ECG<br>BASE                           | 0.903±0.002   | 0.855±0.012    | 0.953±0.010      | 0.917±0.005     | 0.917±0.006                 | 0.911±0.008                   |
| HuBERT-ECG<br>LARGE                          | 0.896±0.003   | 0.828±0.018    | 0.935±0.016      | 0.905±0.002     | 0.919±0.006                 | 0.903±0.007                   |
| Hu et al.[120] <sup>(*)</sup>                | 0.947         | ~ 0.895        | ~ 0.980          | ~ 0.950         | ~ 0.940                     | ~ 0.938                       |
| Liu et al.[111] <sup>(*)</sup>               | N.A.          | N.A.           | N.A.             | N.A.            | N.A.                        | 0.912                         |
| Liu et al.[112] <sup>(*)</sup>               | N.A.          | N.A.           | N.A.             | N.A.            | N.A.                        | 0.877                         |
| Mehari & Strodthoff<br>[113] <sup>(**)</sup> | 0.942         | N.A.           | N.A.             | N.A.            | N.A.                        | N.A.                          |
| Strodthoff et<br>al.[147] <sup>(***)</sup>   | 0.925         | 0.896          | 0.957            | 0.937           | 0.929                       | 0.928                         |
| Na et al.[148] <sup>(***)</sup>              | N.A.          | N.A.           | N.A.             | N.A.            | N.A.                        | 0.933                         |
| Zhang et al.[114] <sup>(*)</sup>             | 0.889         | N.A.           | N.A.             | N.A.            | N.A.                        | N.A.                          |

**Supplementary Material – Appendix C – Table 8:** Fine-tuned HuBERT-ECG performance on PTB-XL datasets against models from the literature. Performance uncertainty is reported using 95% confidence intervals obtained through empirical bootstrapping using 1000 bootstrap iterations. “N.A.” stands for “Not Available”. <sup>(\*)</sup> Code not available. Results preceded by ‘~’ symbol are estimated by looking at paper graphs because of the lack of unequivocal numerical values. <sup>(\*\*)</sup> Based on the available code, the AUROC is calculated on individual batches and then averaged—introducing batch-size dependency—with custom handling for any NaN values that arise due to class imbalances within batches. <sup>(\*\*\*)</sup> Based on the available code, the AUROC is computed after aggregating predictions and corresponding ground-truth labels across all batches—a standard approach we also adopt.