Self-DANA: A Resource-Efficient Channel-Adaptive Self-Supervised Approach for Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Foundation Models (FMs) are large-scale models trained on extensive datasets that can be adapted to a wide range of downstream tasks with minimal fine-tuning. They have recently also gained attention in Electrocardiogram (ECG) signal analysis. One of the key properties of FMs is their transferability to a wide range of downstream scenarios. However, the adaptation of ECG FMs to downstream scenarios with fewer available channels (i.e., wearable and portable devices) still has to be properly investigated. In this work, we propose Self-DANA, an easy-to-integrate solution that enables FMs to be adaptable to a reduced number of input channels, ensuring resource efficiency and high performance. We also introduce Random Lead Selection, a novel augmentation to build more robust and channel-agnostic FMs. Our experiments on three datasets and five reduced-channel configurations demonstrate that Self-DANA significantly enhances resource efficiency while achieving superior or comparable performance to the literature alternative.

1 Introduction

Foundation Models (FM), as defined by [1], are large-scale machine learning models "*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*". A pivotal role in their development and diffusion was played by Self-Supervised Learning (SSL), a training paradigm where models learn useful general representations from unlabeled data by solving auxiliary (pretext) tasks.

Our work primarily focuses on the Electrocardiogram (ECG), a bio-signal that records the heart's electrical activity. Recently, a growing interest in SSL methods for ECG analysis [2, 3] has emerged, leading to the advent of the first ECG FMs [4–8]. Most of them rely on Self-Supervised Contrastive Learning (SSCL), which learns meaningful representations by solving a contrastive pretext task.

One of the key properties of FMs is transferability, i.e., the ability to generalize to a wide range of downstream scenarios. A key yet underexplored aspect in the ECG literature is FMs' adaptability from the standard 12-lead configuration to downstream scenarios involving a reduced number of channels (or *leads*). This aspect is particularly crucial in the wearables or portable devices domain, which often operates with a limited number of sensors or channels due to size, power consumption, and user comfort constraints.

Channel-agnostic learning in the ECG field has been explored in *PhysioNet/Computing in Cardiology Challenge 2021* [9, 10], although focused on end-to-end approaches rather than in the context of FMs. Most of the literature keeps the same channel configuration (usually 12-lead) for pre-train and fine-tune, requiring pre-training ad-hoc FMs with fewer leads in case of reduced-lead downstream tasks (e.g. [5]). Other techniques rely on the combination of 12 leads, so this approach would not even be possible [11, 12]. A better solution, instead, would be to have a single FM able to adapt to any downstream scenario with minimal fine-tuning. To date, the only channel-agnostic ECG FM in the literature is [4], which adopts the technique proposed by [13]. It includes Random Lead Masking

(RLM) during pre-training as contrastive learning augmentation and zero-padding to account for the dimensionality mismatch in the fine-tuning phase. However, the use of zero-padding makes this technique computationally inefficient, especially when only a few leads are available for the downstream task.
In this work, we propose Self-DANA, an easy-to-integrate solution to make SSL architectures adaptable to reduced lead configurations while ensuring resource efficiency and high performance. It combines a dimension-adaptive architecture with an ad-hoc augmentation for SSCL pre-training. This makes FMs adaptable and robust to reduced leads, while optimizing memory and computation.

2 Methods

2.1 Self-DANA

Self-DANA extends and optimizes for SSL frameworks the idea of Dimension Adaptive Neural Architecture (DANA) proposed in [14] for supervised approaches. As in [14], the first component is the introduction of the Dimension Adaptive Pooling (DAP) layer, to make the architecture adaptive to variable input dimensions. In SSL, it also makes pre-trained models adaptable to any reduced-lead configuration required in the downstream task. Differently to zero-padding, this technique uses only the available channels, avoiding including additional values that consume memory without contributing meaningful information. The second element of DANA is the Dimension Adaptive Training (DAT). We adapt it to SSL frameworks by introducing RLS, a new ad-hoc augmentation. Through contrastive-learning, it encourages the model to extract representations independent from the given channel combinations.

2.2 Framework and architecture

We exploited SimCLR [15] as SSL framework for pretraining. Positive pairs are generated with three different sequences of *augmentations* depending on the experiment.

- **Base augmentations**. We apply amplitude scaling, followed by one augmentation randomly chosen from Gaussian noise, crop and resize, time masking, and time warping.
- **Random Lead Masking (RLM)**. Introduced in [13], RLM consists of randomly masking a subset of channels by setting their values to 0, while keeping the rest unaltered.
- **Random Lead Selection (RLS)**. Our proposed augmentation consists of randomly selecting only a subset of the input channels. For both RLM and RLS, we randomly choose, with uniform probability, both the type (any of the 12 standard ECG leads) and the number of channels to mask or to select, independently for each branch.

As a *backbone*, we propose a memory-efficient and dimension-adaptive variant of the architecture exploited in [13] inspired by [16], which integrates a convolutional feature extractor and a transformer encoder. To make the feature encoder architecture independent of the number of input channels, we replaced the original 1D-convolutions with 2D-convolutions with kernel $(1, k)$ and stride $(1, s)$ (i.e., only temporal convolution), and we added a DAP layer on top of it. The latter reduces any input dimension (C, T) to $(1, 156)$ by applying an average pooling. The backbone is followed by a single linear layer (*projection head*) for pre-training and by a fully connected layer with sigmoid activation (*classification head*) for the downstream tasks. More details are reported in the Appendix.

2.3 Datasets

Pre-training We pre-trained all the models on a large collection of 12-lead ECGs, from seven open-access datasets: *Code-15%* [17, 18], *Off-test* [19, 20] and five datasets from *PhysioNet/Computing in Cardiology Challenge 2021* [9, 10, 21, 22]: (*Ningbo* [23], *Chapman-Shaoxing* [24], *INCART*, *CPSC* and *CPSC-extra* [25]. This amounts to a total of 406'117 recordings from 295'245 different subjects (855'424 5-second 12-lead ECG windows after segmentation).

Fine-tuning The main experiments have been conducted on the out-of-domain *Georgia* dataset [9, 10, 21, 22], for consistency with [13]. We additionally conducted performance comparison on two other out-of-domain datasets (i.e., never seen during pre-training): *PTB-XL* [21, 26, 27], a well-known benchmarking dataset, and *CinC2017* dataset [21, 28], recorded with a single-lead heart

monitor. Georgia and PTB-XL have been used to classify cardiac abnormalities (23 for Georgia and 22 for PTB-XL) in a multi-label setting (up to 7 for each ECG) from reduced-lead ECGs; CinC2017, instead, to address a binary atrial fibrillation detection task. Georgia includes 9'458 12-lead ECGs, PTB-XL 21'604 12-lead ECGs, and CinC2017 8'528 single-lead (lead I) ECGs. To evaluate model adaptability to various reduced-leads configurations, five reduced-lead datasets have been extracted from Georgia and PTB-XL, by selecting the following leads from the original datasets, as in [13]: 12-lead, 6-lead (I, II, III, aVF, aVL, aVR), 3-lead (I, II, V2), 2-lead (I, II), and 1-lead (I).

All datasets have been preprocessed using the same procedure: resampling at 500 Hz, segmentation into non-overlapping 5s windows, and filtering. More details are provided in the Appendix.

2.4 Experiments

Three main experiments have been conducted and repeated five times with different seeds. Pre-training and fine-tuning experimental setup is also graphically represented in the Appendix.

***(i)* DAP layer** We examined whether the DAP layer adoption, compared to zero-padding, enhances resource efficiency without sacrificing performance. We pre-trained a baseline model (**PT-base**) with all 12 leads and only the *base augmentations*. We then fine-tuned on the five reduced-leads datasets:

- **FT-base-ZP**: zero-padding is applied to match the required input number of channels (12)
- **FT-base-DAP**: the DAP layer is exploited to adapt to the available number of channels

***(ii)* Self-DANA** We evaluated whether the combination with the ad-hoc RLS augmentation is beneficial for the DAP layer approach. We also compared this setup with the RLM-based approach to ensure that the improved computational efficiency was not associated with a performance deterioration. We pre-trained **PT-RLM** applying the *base augmentations* and RLM and **PT-RLS** applying the *base augmentations* and RLS. We then fine-tuned them on the five reduced-leads datasets:

- **FT-RLM-ZP**: PT-RLM model has been fine-tuned on the downstream task, applying zero-padding to keep the number of channels always equal to 12, as in [13].
- **FT-RLM-DAP**: PT-RLM model has been fine-tuned on the downstream task without zero-padding, exploiting the DAP layer to keep the reduced number of channels.
- **Self-DANA**: PT-RLS model has been fine-tuned on the downstream task without zero-padding, exploiting the DAP layer to keep the reduced number of channels.

***(iii)* Channel-adaptive FM vs channel-specific supervised models** We investigated the benefit of fine-tuning a channel-adaptive FM on the desired task and lead configuration, rather than training different channel-specific models from scratch. Specifically, for each of the five reduced-lead configurations, we trained a dedicated fully supervised model using only the available channels.

Evaluation We assessed both classification performance and computational efficiency during fine-tuning. Performance on Georgia test set is primarily evaluated through CinC score, introduced for the challenge [9, 10] and used in [13]. Additionally, macro AUROC, macro F1 score, and weighted F1 score have been computed for Georgia and PTB-XL dataset, while AUROC, macro F1 score, weighted F1 score and accuracy for CinC2017. Computational efficiency is evaluated in terms of peak memory usage and training time normalized by the number of epochs, for both CPU and GPU.

3 Results

3.1 Performance

Table 1 includes the results of experiments *(i)*, *(ii)*, and *(iii)* on Georgia dataset. Additional evaluation with other metrics and with PTB-XL and CinC2017 dataset are provided in the Appendix.

***(i)* DAP layer** For all configurations, FT-base-DAP achieves comparable or slightly better performance than FT-base-ZP. It proves the feasibility of using the DAP layer as an alternative to zero-padding, without causing a performance drop.

Table 1: CinC score obtained in experiments *(i)*, *(ii)* and *(iii)* on the five Georgia reduced-leads test sets. Results are reported as *mean* $\pm$ *standard deviation*. Best results (highest CinC score) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.578 ± 0.006	0.573 ± 0.007	0.574 ± 0.005	0.581 ± 0.002	0.547 ± 0.003
FT-base-ZP	0.600 ± 0.005	0.589 ± 0.004	0.595 ± 0.009	0.590 ± 0.006	0.562 ± 0.004
FT-base-DAP	0.600 ± 0.005	0.595 ± 0.003	0.600 ± 0.005	0.598 ± 0.005	0.568 ± 0.005
FT-RLM-ZP	0.612 ± 0.006	0.606 ± 0.002	0.610 ± 0.003	0.611 ± 0.006	**0.585 ± 0.004**
FT-RLM-DAP	0.612 ± 0.006	0.603 ± 0.007	0.608 ± 0.003	0.609 ± 0.005	0.578 ± 0.003
Self-DANA	**0.619 ± 0.009**	**0.613 ± 0.004**	**0.617 ± 0.009**	**0.618 ± 0.008**	0.583 ± 0.004

Table 2: Peak CPU and GPU memory and average epoch training time for FT-RLM-ZP and Self-DANA during fine-tuning. Best results (lowest peak memory and lowest normalized time) in bold.

		12 leads	6 leads	3 leads	2 leads	1 lead
Peak CPU mem. (MB)	FT-RLM-ZP	**58.68**	58.68	58.68	58.68	58.68
	Self-DANA	**58.68**	**29.39**	**18.00**	**18.00**	**18.00**
Peak GPU mem. (GB)	FT-RLM-ZP	**23.19**	23.19	23.19	23.19	23.19
	Self-DANA	**23.19**	**18.84**	**16.66**	**15.93**	**15.21**
Avg CPU time (s)	FT-RLM-ZP	**137.71 ± 6.59**	123.50 ± 11.28	123.56 ± 10.52	119.39 ± 7.34	118.82 ± 8.92
	Self-DANA	140.43 ± 3.74	**116.37 ± 0.89**	**105.95 ± 0.76**	**102.72 ± 0.79**	**98.84 ± 0.88**
Avg GPU time (s)	FT-RLM-ZP	**139.58 ± 0.97**	139.81 ± 0.54	139.88 ± 0.37	139.68 ± 0.49	140.00 ± 0.41
	Self-DANA	140.13 ± 0.10	**122.08 ± 0.08**	**113.10 ± 0.16**	**110.17 ± 0.22**	**107.43 ± 0.12**

***(ii)* Self-DANA** Self-DANA consistently outperforms FT-base-DAP across all configurations, with RLS enhancing model robustness, even in the absence of lead configuration mismatch (12-lead). The combination of DAP layer and RLM (FT-RLM-DAP) achieves lower performance than Self-DANA and FT-RLM-ZP, suggesting that RLM needs to be combined with zero padding to better exploit its potential. Finally, Self-DANA is comparable or slightly superior to FT-RLM-ZP, further establishing our approach as a valid alternative to the RLM technique.
We provide reference CinC score values from the literature, even though test datasets and conditions were slightly different, and contrary to us, the other works included Georgia also in pre-training (in-domain). The performance obtained by [13] with SimCLR framework and RLM on Georgia and CPSC test sets combined are: 0.578 ± 0.015 for 12-lead, 0.497 ± 0.002 for 6-lead, 0.535 ± 0.015 for 3-lead, 0.484 ± 0.004 for 2-lead and 0.393 ± 0.012 for 1-lead configurations (mean $\pm$ 95% confidence interval). The winner [29] of the *PhysioNet/Computing in Cardiology Challenge 2021* obtained an average CinC score of 0.61 (mean between 12, 3, and 2 leads) on the (unavailable) Georgia test set.

***(iii)* Channel-adaptive FM vs channel-specific supervised models** Self-DANA consistently outperforms the supervised counterpart across all five configurations, further supporting that it is beneficial to exploit a channel-adaptive FM to obtain a model robust to reduced-lead lead configurations.

3.2 Resource efficiency

In Table 2, we compared Self-DANA and the reference FT-RLM-ZP in terms of resource efficiency. Both the peak GPU and CPU memory required by Self-DANA are considerably lower for all reduced-lead configurations and, as expected, decrease with the number of available channels. Self-DANA saves up to 34.41% GPU memory and up to 69.32% CPU memory, highlighting its superior efficiency. With Self-DANA, the average time per training epoch is lower, or comparable for the 12-lead case, and the standard deviations are considerably lower, indicating greater stability in the results.

4 Conclusions

Self-DANA offers an easy-to-integrate yet powerful solution for adapting FMs to reduced-lead scenarios, achieving performance comparable to the alternative proposed in the literature with significantly improved memory and computational efficiency. By combining the DAP layer, to avoid zero-padding, with our RLS augmentation, for enhanced robustness to reduced-leads configurations, Self-DANA stands out as a strong candidate for portable and wearable applications. Future works will assess its applicability to non-CL frameworks and other signal modalities and multimodal settings.

A Technical Appendices and Supplementary Material

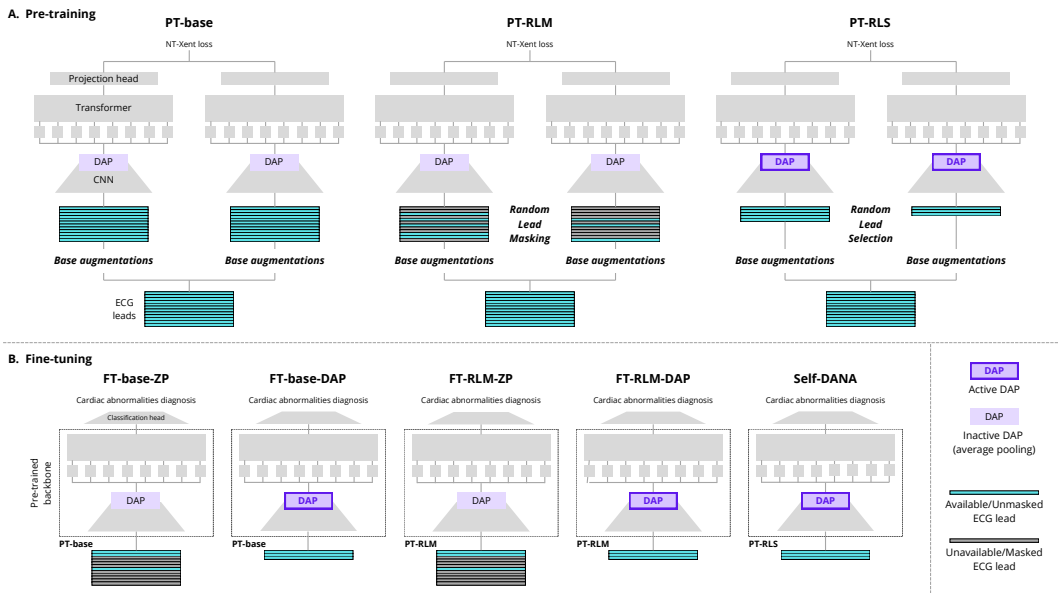

Figure 1: Overview of the experimental design, including the pre-training and fine-tuning procedures. The diagram illustrates how the Dimension Adaptive Pooling (DAP) layer and Random Lead Selection (RLS) components were integrated and evaluated across different experimental conditions.

A.1 Base augmentations

This section provides a description and parameters related to the five augmentations exploited in the *base augmentations* sequence. Amplitude scaling, Gaussian noise, crop and resize, time masking, and time warping have been selected as they are basic augmentation techniques usually exploited in the literature for ECG signals and are suitable for our context. To encourage the model to learn more robust representations, we apply first amplitude scaling, followed by one augmentation randomly chosen from the remaining four.

Amplitude scaling It multiplies the signal's amplitude by a random scaling factor s. This trains the model to be invariant to amplitude differences that may arise from patient-specific variations or electrode placement, encouraging the encoder to focus on morphological patterns and temporal structure rather than absolute signal magnitude. Based on the results of [30], we randomly select, for each ECG window, a scaling factor s in the range of 0.5 to 1.7, meaning that the ECG will be rescaled between 50% and 170% of its original amplitude.

Gaussian noise It simulates the noise inherently present in real-world settings, due to electrode movement, muscle artifacts, etc. The augmented ECG view is obtained by adding Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Based on the results of [30], we randomly select, for each ECG window, a standard deviation value σ in the range of 0.1 to 0.25, introducing minor variability while preserving the core structure of the signal.

Crop and resize It consists of cropping a random contiguous portion of the ECG signal and then resizing it back to the original length via interpolation. This transformation changes the temporal resolution but retains overall shape information. Following [31], a portion of the signal, with a randomly determined length between 50% and 100% of the original signal, is randomly cropped. The cropped segment is then resized to the target length using cubic spline interpolation.

Time masking It emulates signal dropout or corruption by setting a random portion of the signal to zero. This improves the model's robustness to missing or noisy data, forcing it to reconstruct or

interpret partial signals. Based on the results of [30, 31], we randomly select the masking ratio within the range of 0% to 50% of the signal length and then mask with zeros a random portion of the desired length.

Time warping It transforms the temporal structure of signals by stretching or squeezing random segments, helping the model handle variations in heart rate and temporal dynamics. A random number of segments, ranging from 4 to 9, is first selected. These segments are then randomly designated to be either stretched by a factor of 2 or squeezed by a factor of 0.5. To match the original length after these temporal modifications, resampling is performed using piecewise cubic Hermite interpolating polynomials (PCHIP). For the implementation, we refer to [32, 33].

A.2 Architectural details

Pre-training and fine-tuning experimental setup is graphically represented in Figure 1. The DAP layer has been included also in FT-base-ZP and FT-RLM-ZP, to ensure a fair comparison, but it is inactive, i.e., it is used as a classic average pooling layer, since the input channels are always 12, as in the pre-training phase.

Backbone A detailed scheme of the backbone architecture with hyperparameters is provided in Table 3, while a graphical overview is shown in Figure 2. The convolutional feature extractor includes four convolutional blocks, while the transformer encoder comprises a convolutional positional encoder and a sequence of 12 transformer-encoder blocks, each with 12 self-attention heads. The transformer encoder is identical to the one used in [13], while in the feature encoder, we replaced 1D convolutions with 2D convolutions and added the DAP layer on top. Moreover, since it was not needed with the SimCLR framework, the masking and quantization operations present in [13] have been omitted.

Pre-training For the pre-training phase, we exploited a projection head consisting of a fully connected layer with input dimension 768 and output dimension 256, followed by a 1D batch normalization layer applied to the 256 output neurons, as in [13]. The backbone comprises 90'367'616 parameters, and the projection head contributes an additional 197'376, resulting in a total parameter count of 90'564'992.
For the self-supervised pre-training with the SimCLR framework, we employed the original loss function, i.e., the *NT-Xent loss* function, defined as

$$\ell_{i,j} = -\log \frac{\exp\left(\frac{\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\frac{\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_k)}{\tau}\right)} \tag{1}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the projected representations of the positive pair, $\mathrm{sim}(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ is the cosine similarity, τ is a temperature parameter, and N is the batch size.
Specifically, we set $\tau = 0.5$ and $N = 128$.

Fine-tuning For the fine-tuning phase, we replaced the projection head with a classification head composed of a fully connected layer with input dimension 768 and output dimension 23 (number of classes of the downstream task) and a sigmoid activation function. The classification head accounts for a total of 796 parameters.
To address the cardiac abnormality classification task, where multiple labels can be associated with a single recording, we employed a binary cross-entropy loss function applied independently to each of the 23 output neurons, enabling the model to learn the presence or absence of each class.

Source code For the implementation of the encoder and projection head architecture, we refer to the source code provided in [34], reporting the architecture exploited in [13] and [4]. Unless otherwise specified in this Appendix, we kept the same hyperparameters as in the original work and code. The rest of the code structure, including the SimCLR framework, builds upon the selfEEG library [32, 33].

Table 3: Detailed backbone architecture with hyperparameters.

Layer	Hyperparameters	Output Shape
input	-	[B, C, 2500]
Convolutional feature encoder		
Conv2D	k=(1,2), s=(1,2)	[B, 256, C, 1250]
GroupNorm	n_groups=256, eps=1e-5	[B, 256, C, 1250]
GELU	-	[B, 256, C, 1250]
Conv2D	k=(1,2), s=(1,2)	[B, 256, C, 625]
GELU	-	[B, 256, C, 625]
Conv2D	k=(1,2), s=(1,2)	[B, 256, C, 312]
GELU	-	[B, 256, C, 312]
Conv2D	k=(1,2), s=(1,2)	[B, 256, C, 156]
GELU	-	[B, 256, C, 156]
Adaptive Avg Pool	out_dim=(1,156)	[B, 256, 1, 156]
Flatten	dim=2	[B, 256, 156]
Transpose	-	[B, 156, 256]
LayerNorm	norm_shape=256, eps=1e-05	[B, 156, 256]
Linear	in_dim=256, out_dim=768	[B, 156, 768]
Dropout	p=0.1	[B, 156, 768]
Convolutional positional encoder		
Conv1D	k=128, s=1, pad=64, groups=16	[B, 768, 157]
Padding	-	[B, 768, 156]
GELU	-	[B, 768, 156]
Transpose	-	[B, 156, 768]
Transformer encoder		
LayerNorm	norm_shape=768, eps=1e-05	[B, 156, 768]
Dropout	p=0.1	[B, 156, 768]
Transpose	-	[156, B, 768]
Transformer-encoder blocks (x 12)		
MultiHead Attention	embed_dim=768, n_heads=12, dropout=0.1, q=k=v=768	[156, B, 768]
Dropout	p=0.1	[156, B, 768]
LayerNorm	norm_shape=768, eps=1e-05	[156, B, 768]
Linear	in_dim=768, out_dim=3072	[156, B, 3072]
GELU	-	[156, B, 3072]
Linear	in_dim=3072, out_dim=768	[156, B, 768]
Dropout	p=0.1	[156, B, 768]
LayerNorm	norm_shape=768, eps=1e-05	[156, B, 768]
Transpose	-	[B, 156, 768]

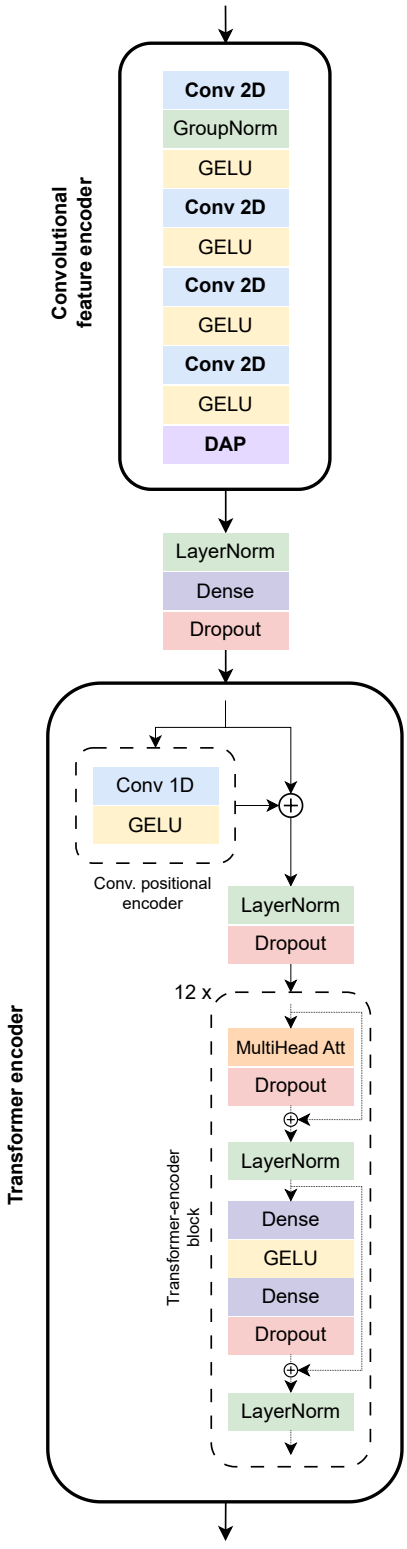

Figure 2: Graphical representation of backbone architecture.

A.3 Datasets

This section supplies further information about the datasets exploited in pre-training and fine-tuning experiments. Then, the datasets' analyses in terms of distributions of recordings and labels across the split sets, the preprocessing applied, and the exact splits subdivision are also provided.

CPSC and CPSC-extra *CPSC and CPSC-extra* datasets, from *PhysioNet/Computing in Cardiology Challenge 2021*, derive from the China Physiological Signal Challenge in 2018 (CPSC2018). It consists of two sets of 6,877 (male: 3,699; female: 3,178) and 3,453 (male: 1,843; female: 1,610) of 12-ECG recordings lasting from 6 seconds to 60 seconds. Each recording is sampled at 500 Hz.

INCART *St Petersburg INCART 12-lead Arrhythmia Database (INCART)* dataset, from *PhysioNet/Computing in Cardiology Challenge 2021*, consists of 74 annotated recordings extracted from 32 Holter records. Each record is 30 minutes long and contains 12 standard leads, each sampled at 257 Hz.

Chapman-Shaoxing and Ningbo *Chapman University, Shaoxing People's Hospital (Chapman-Shaoxing)* dataset and *Ningbo First Hospital (Ningbo)* dataset, from *PhysioNet/Computing in Cardiology Challenge 2021*, contain a total of 45,152 ECGs (all shared as training data). Each recording is 10 second long with a sampling frequency of 500 Hz.

Code-15% *CODE-15%* dataset is a subset of the larger CODE dataset, created through stratified sampling to include 15% of the patients. It consists of 345,779 annotated 12-lead ECG exams from 233,770 patients. The data was collected by the Telehealth Network of Minas Gerais (TNMG), a public telehealth system serving most municipalities in Minas Gerais, Brazil, between 2010 and 2016.

Off-test *Offline Test Set of ECG Multi-label Classification (Off-test* dataset is the offline test set for the study "Practical arrhythmias detection algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset." It includes 7,000 12-lead wearable ECG recordings, each 15 seconds long with a sampling frequency of 500 Hz. The data covers 60 rhythm classes, all reviewed and diagnosed by cardiologists and spanning a wide range of normal and abnormal heart conditions.

Georgia *Georgia* dataset, from *PhysioNet/Computing in Cardiology Challenge 2021*, represents a unique demographic of the Southeastern United States. It includes 20,672 ECG recordings, with 10,344 used for training, 5,167 for validation, and 5,161 for testing. Each ECG lasts between 5 and 10 seconds and is sampled at a frequency of 500 Hz.
We used just the training set of this dataset since it is the only one made available. Each recording is associated with multiple (up to 7) cardiac abnormalities, for a total of 833 different label combinations.

PTB-XL *PTB-XL* dataset contains 21'799 clinical 12-lead ECG recordings, each 10 seconds long, from 18'869 patients aged 0 to 95 years (median age 62), with a balanced gender distribution. The data was collected using Schiller AG devices between 1989 and 1996 and with a sampling frequency of 500 Hz. Each ECG was annotated by up to two cardiologists with a report string, which was converted into a standardized set of 71 standardized SCP-ECG diagnostic, form, and rhythm statements, and validated by technical experts. Along with the raw waveform data, the dataset includes patient metadata such as age, sex, weight, and height. The dataset is valuable for its broad representation of cardiac conditions, including many healthy controls, and comprehensive annotations with unique ECG and patient identifiers.
In our work, we exploited the recommended split provided with the dataset. Each recording is associated with multiple (up to 7) cardiac abnormalities, for a total of 682 different label combinations.

CinC2017 *The 2017 PhysioNet/CinC Challenge (CinC2017)* dataset consists of single-lead ECGs recorded with the AliveCor device, which were provided by the company for use in the challenge. The training set comprises 8'528 recordings with durations ranging from 9 seconds to over 60 seconds, while the test set includes 3'658 recordings of similar lengths. All recordings were sampled at 300 Hz and underwent bandpass filtering as part of the device's preprocessing. The dataset includes four label categories: normal sinus rhythm, atrial fibrillation (AF), other rhythms, and noisy signals. In our study, we used only the training set of the CinC2017 dataset, splitting it into train, validation,

and test sets. For the classification task, we framed it as a binary problem: distinguishing AF from non-AF recordings (normal sinus rhythm, other rhythms, and noisy recordings).

Preprocessing All pre-training and fine-tuning datasets have been prepared using the same procedure: resampling at 500 Hz, segmentation into non-overlapping 5s windows, and pre-processing. First, resampling was done, when needed, to a frequency of 500 Hz. Then, each recording has been segmented into non-overlapping 5s-windows. Finally, each window has been pre-processed by removing the mean and applying a 5th-order moving average filter and a Butterworth 4th-order band-pass filter with cutoff frequencies of 0.5 and 40 Hz.

Splits Pre-training datasets have been split into training and validation sets according to an 80:20 ratio and source dataset stratification. Fine-tuning datasets, instead, have been split into training, validation, and test sets according to an 80:10:10 ratio and label stratification. All the splits have been performed on a subject basis to ensure all the recordings of a subject fall in the same split set.

A.4 Experimental settings

Training details During the pre-training phase, we employed the Adam optimizer with an exponential decay setting an initial learning rate $lr = 5e - 5$ and a decay factor $\gamma = 0.97$. We trained our models for a maximum of 100 epochs, applying early stopping based on validation loss with a patience of 10 and a minimum improvement threshold of $1e - 5$. During the fine-tuning phase, we employed the Adam optimizer with an exponential decay setting an initial learning rate $lr = 1e - 5$ and a decay factor $\gamma = 0.97$. We trained our models for a maximum of 50 epochs, applying early stopping based on validation loss with a patience of 10 and a minimum improvement threshold of $1e - 3$. We repeated the fine-tuning five times with different seeds $(0, 1, 2, 3, 4)$. In experiments *i* and *ii*, full fine-tuning was carried out to address the downstream tasks.
The experiments have been performed on a cluster with an L40S GPU with 46068 MiB of RAM.

Data sampling strategies For batch generation, we exploited two different sampling strategies for the pre-training and fine-tuning phases. During pre-training, we employed a batch size of 128 5s-segments. We populate the batches by sampling the segments from the whole pre-training set with uniform distribution. During fine-tuning, we employed a batch size of 128 5s-segments. To balance the labels distribution in the batch, we first select one of the 23 cardiac abnormalities with random uniform sampling (*class sampling*); we then randomly chose a subject (*subject sampling*) and a recording (*recording sampling*) containing the desired label; and, finally, we randomly select a 5s-segment from that recording. An epoch is concluded after $Nb = Tr/N$ batches, where Tr is equal to the number of 5s segments in the train set, and N is the batch size.

A.5 Supplementary results

Datasets analysis Table 4 and 5 report the number of subjects, recordings, and 5s windows for each split set of pre-training and fine-tuning datasets, respectively. The label distribution in the Georgia and PTB-XL multi-label datasets and its corresponding splits is presented in two forms: by counting the occurrence of each label across recordings (see Table 6 and Figure 3 for Georgia and Table 8 and Figure 4 for PTB-XL), and by analyzing the different unique label combinations assigned to the recordings (see Table 7 for Georgia and Table 9 for PTB-XL). For CinC2017, instead, the label distribution is reported in Table 10 and Figure 5.

Table 4: Number of subjects (*sbj*), recordings (*rec*), and 5s windows (*win*) in training and validation sets of the pre-training datasets

	Train			Validation		
	sbj	*rec*	*win*	*sbj*	*rec*	*win*
Code-15%	186131	274850	549700	46601	68712	137424
Ningbo	27923	27923	55846	6981	6981	13962
INCART	25	59	21240	7	15	5400
Off-test	5600	5600	16800	1400	1400	4200
Chapman-Shaoxing	8198	8198	16396	2049	2049	4098
CPSC	5493	5493	16138	1384	1384	4024
CPSC-extra	2741	2741	8175	712	712	2021
Total	236111	324864	684295	59134	81253	171129

Table 5: Number of subjects (*sbj*), recordings (*rec*), and 5s windows (*win*) in training, validation and test sets of the fine-tuning datasets

	Train			Validation			Test		
	sbj	*rec*	*win*	*sbj*	*rec*	*win*	*sbj*	*rec*	*win*
Georgia	6745	6745	13450	1147	1147	2292	1566	1566	3130
PTB-XL	14960	17284	34568	1900	2160	4320	1870	2160	4320
CinC2017	7608	7608	43649	464	464	5472	456	456	5472

Table 6: Number of 5s windows in which each label is present in train, validation, test, and whole Georgia dataset. A description of the label acronym is also provided.

Label	Description	All	Train	Validation	Test
TAb	T wave abnormal	4605	3074	598	933
NSR	Normal sinus rhythm	3487	2787	350	350
SB	Sinus bradycardia	3345	2331	414	600
LQT	Prolonged QT interval	2774	1825	355	594
STach	Sinus tachycardia	2522	1728	336	458
LAD	Left axis deviation	1878	1098	280	500
TInv	T wave inversion	1620	889	238	493
IAVB	First degree av block	1535	891	224	420
PAC	Premature atrial contraction	1277	720	186	371
AF	Atrial fibrillation	1138	676	158	304
RBBB	Right bundle branch block	1110	647	170	293
QAb	Q wave abnormal	927	527	118	282
SA	Sinus arrhythmia	909	559	118	232
IRBBB	Incomplete right bundle branch block	809	392	128	289
LQRSV	Low QRS voltages	747	463	108	176
PVC	Premature ventricular contractions	713	326	111	276
LBBB	Left bundle branch block	462	250	74	138
NSIVCB	Nonspecific intraventricular conduction disorder	406	154	64	188
AFL	Atrial flutter	372	178	56	138
LAnFB	Left anterior fascicular block	360	174	56	130
BBB	Bundle branch block	231	120	40	71
RAD	Right axis deviation	163	49	29	85
Brady	Bradycardia	12	6	2	4

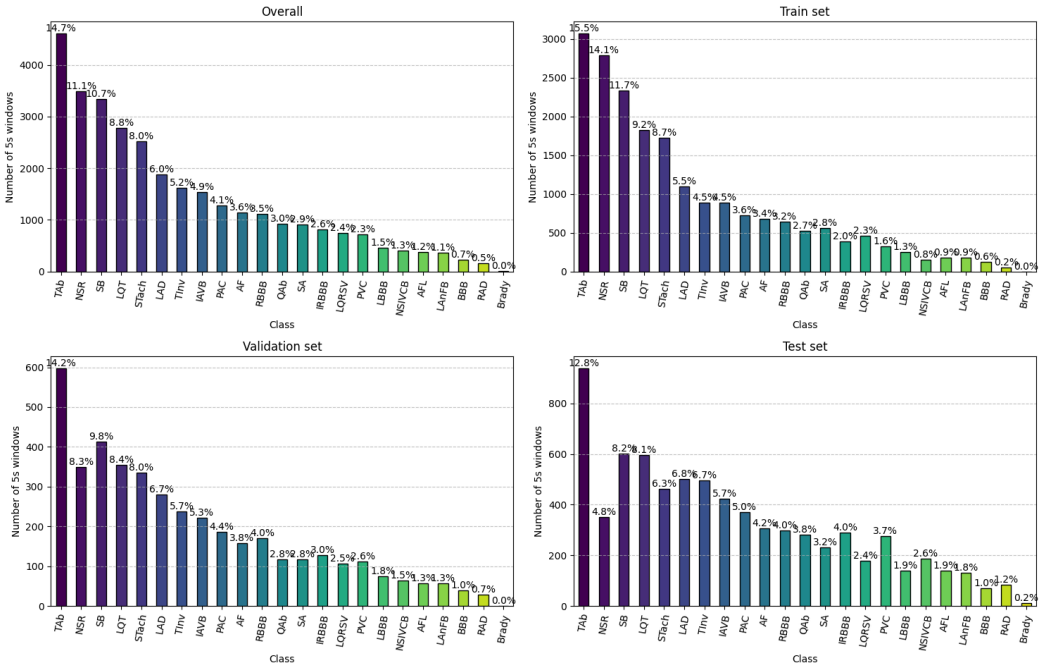

Figure 3: Label distribution of the Georgia dataset and in its train, validation, and test subsets. Labels' acronyms are described in Table 6.

Table 7: Number of 5s windows for each set of labels in training, validation, test, and whole Georgia dataset. Only the first and the last 25 more frequent label combinations (out of 833) are reported here.

Set of labels	**All**	**Train**	**Validation**	**Test**
NSR	3487	2787	350	350
SB	1433	1145	144	144
STach	1016	812	102	102
TAb	893	713	90	90
LQT	808	644	82	82
SA	386	306	40	40
LQT, TAb	384	304	40	40
LAD	340	272	34	34
PAC	322	254	34	34
TInv	288	228	30	30
AF	285	225	30	30
IAVB	258	206	26	26
STach, TAb	258	206	26	26
TAb, QAb	226	178	24	24
RBBB	224	176	24	24
LQRSV	207	163	22	22
IAVB, SB	203	159	22	22
TAb, TInv	187	147	20	20
TAb, SB	178	142	18	18
PVC	134	106	14	14
SA, SB	134	106	14	14
IRBBB	128	100	14	14
AF, TAb	126	98	14	14
LAD, SB	125	97	14	14
LAnFB	96	76	10	10
...				
LAD, STach, IAVB, BBB, LBBB, PAC	2	0	0	2
AF, TInv, IRBBB	2	0	0	2
PAC, SA, SB	2	0	0	2
QAb, TAb, SB, LQT, SA	2	0	0	2
LAnFB, TAb, SA	2	0	0	2
IAVB, SA, PAC	2	0	0	2
NSIVCB, AFL, SB	2	0	0	2
LAnFB, LQT, SB	2	0	0	2
AF, TAb, PVC, QAb	2	0	0	2
NSIVCB, LAD, LQT, SB	2	0	0	2
AF, TAb, IRBBB, RBBB, LQT	2	0	0	2
PAC, TAb, SA, QAb	2	0	0	2
LAD, TAb, SB, LQT, PAC	2	0	0	2
RAD, TAb, QAb	2	0	0	2
LAD, TAb, IRBBB, TInv, AFL	2	0	0	2
LAD, TAb, TInv, STach	2	0	0	2
LAD, TAb, IAVB, TInv, LQT, PAC	2	0	0	2
LAD, STach, PVC, IRBBB	2	0	0	2
RAD, SA, TInv	2	0	0	2
QAb, TInv, SB	2	0	0	2
AF, TAb, TInv, AFL	2	0	0	2
AF, TAb, PAC	2	0	0	2
AF, TAb, IRBBB, RBBB, SB	2	0	0	2
LAD, IAVB, BBB, LBBB, PVC	2	0	0	2
BBB, RBBB, PAC	1	0	0	1
RAD, TAb, TInv, IRBBB	1	0	0	1

Table 8: Number of 5s windows in which each label is present in train, validation, test, and whole PTB-XL dataset. A description of the label acronym is also provided.

Label	Description	All	Train	Validation	Test
NSR	Normal sinus rhythm	36184	28894	3658	3632
LAD	Left axis deviation	10292	8278	984	1030
TAb	T wave abnormal	4690	3722	460	508
LAnFB	Left anterior fascicular block	3252	2594	338	320
AF	Atrial fibrillation	3028	2454	278	296
IRBBB	Incomplete right bundle branch block	2236	1810	216	210
STach	Sinus tachycardia	1652	1304	172	176
IAVB	First degree av block	1594	1286	146	162
NSIVCB	Nonspecific intraventricular conduction disorder	1578	1248	168	162
SA	Sinus arrhythmia	1544	1262	148	134
SB	Sinus bradycardia	1274	1038	108	128
PAC	Premature atrial contraction	1110	888	130	92
QAb	Q wave abnormal	1096	852	134	110
RBBB	Right bundle branch block	1084	860	104	120
LBBB	Left bundle branch block	1072	886	86	100
RAD	Right axis deviation	686	550	74	62
LPR	Prolonged PR interval	680	542	80	58
PR	Pacing rhythm	592	496	48	48
TInv	T wave inversion	588	456	54	78
LQRSV	Low QRS voltages	364	294	32	38
LQT	Prolonged QT interval	236	188	22	26
AFL	Atrial flutter	146	120	18	8

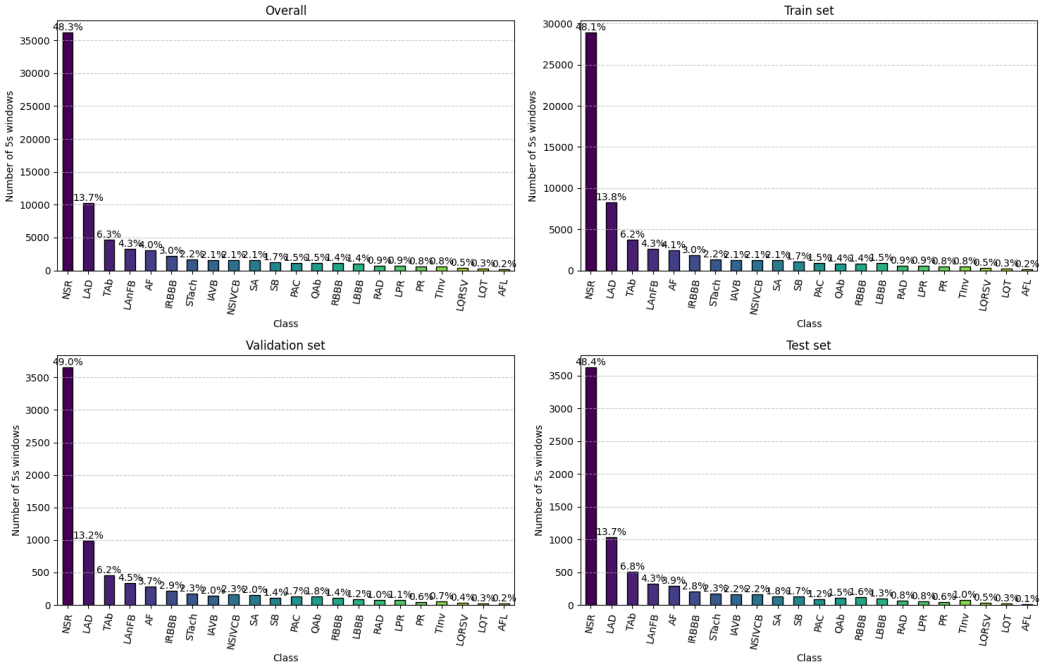

Figure 4: Label distribution of the PTB-XL dataset and in its train, validation, and test subsets. Labels' acronyms are described in Table 8.

Table 9: Number of 5s windows for each set of labels in training, validation, test, and whole PTB-XL dataset. Only the first and the last 25 more frequent label combinations (out of 682) are reported here.

Set of labels	**All**	**Train**	**Validation**	**Test**
NSR	18532	14818	1830	1884
LAD, NSR	3820	3022	394	404
NSR, TAb	2222	1730	244	248
LAD, NSR, LAnFB	1336	1080	142	114
AF	1030	836	98	96
IRBBB, NSR	928	740	104	84
NSR, SA	696	564	58	74
PR	552	468	40	44
NSR, SB	526	420	58	48
NSIVCB, NSR	470	354	68	48
TAb, AF	400	334	34	32
LAD, AF	390	310	38	42
LAD, TAb, NSR	364	294	30	40
LAD, LBBB, NSR	346	286	32	28
NSR, STach	338	262	46	30
NSR, PAC	334	252	48	34
NSR, QAb	334	254	46	34
NSR, IAVB	308	270	24	14
STach	288	226	28	34
NSR, LBBB	248	202	24	22
IRBBB, LAD, NSR	236	194	22	20
NSR, RBBB	218	172	26	20
NSIVCB, LAD, NSR	216	180	16	20
NSR, RAD	214	162	30	22
SA	210	170	24	16
...				
AF, QAb, LAnFB	2	2	0	0
IRBBB, TAb, SB	2	2	0	0
NSR, LQRSV, IAVB	2	0	0	2
QAb, LPR, NSR, LAD, LAnFB	2	0	2	0
LAD, PAC, RBBB	2	2	0	0
AF, LAD, QAb, LQRSV	2	2	0	0
NSR, TAb, SB	2	2	0	0
LQRSV, RAD, STach	2	2	0	0
NSR, TAb, IAVB, IRBBB, LAD	2	2	0	0
LAD, PAC, LQT	2	2	0	0
RBBB, STach, IAVB	2	2	0	0
AF, RBBB, RAD, IAVB	2	2	0	0
IRBBB, PR, RAD	2	0	0	2
NSIVCB, RAD, SB	2	2	0	0
NSR, SB, IRBBB, LAD, LAnFB	2	2	0	0
NSR, TAb, QAb, IAVB	2	2	0	0
IRBBB, SB, IAVB	2	2	0	0
LAnFB, LAD, NSR, SA	2	2	0	0
PR, SB	2	0	0	2
NSIVCB, LAD, IAVB, LAnFB	2	2	0	0
RBBB, IAVB	2	0	2	0
IRBBB, LAD, AF, LQRSV	2	2	0	0
LAD, QAb, STach	2	2	0	0
LAD, QAb, SB, LAnFB	2	2	0	0
LBBB, AF, IAVB	2	0	2	0

Table 10: Number of 5s windows in which each label is present in train, validation, test, and whole CinC2017 dataset. A description of the label acronym is also provided.

Label	Description	All	Train	Validation	Test
not AF	Not atrial fibrillation	49760	39800	4980	4980
AF	Atrial fibrillation	4833	3849	492	492

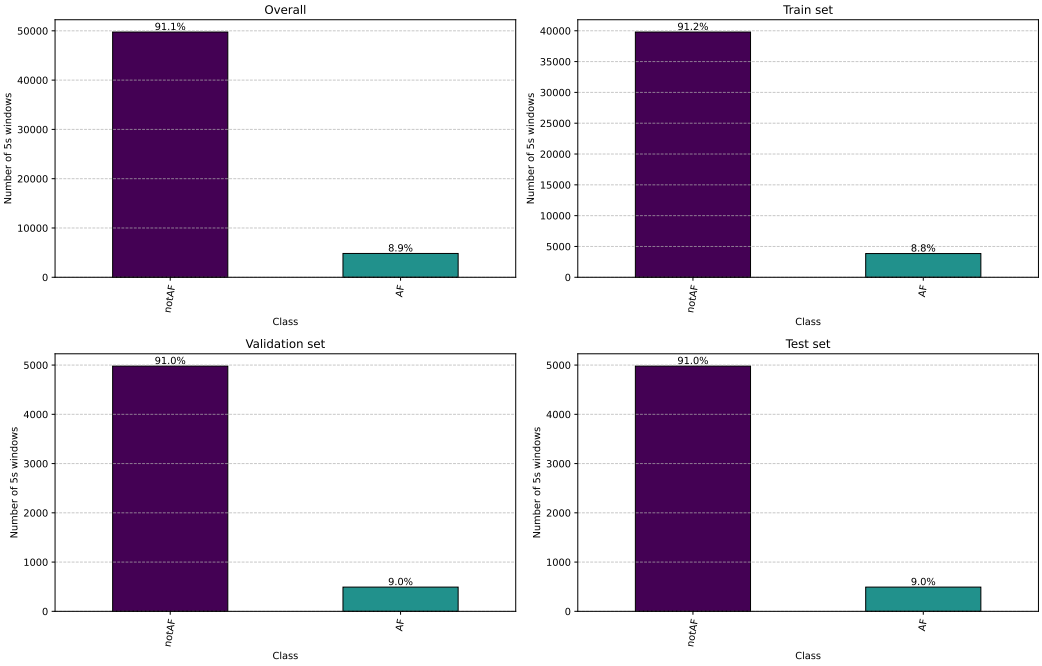

Figure 5: Label distribution of the CinC2017 dataset and in its train, validation, and test subsets. Labels' acronyms are described in Table 10.

Learning curves Figure 6 and 7 report the validation and training loss obtained, respectively, during the pre-training and fine-tuning of all the models and configurations considered in our experiments.

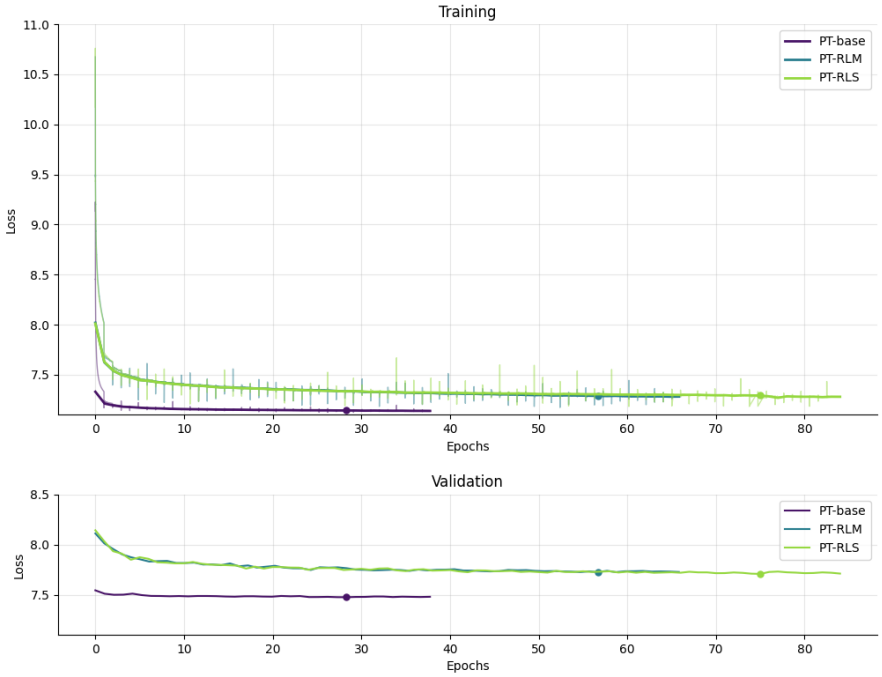

Figure 6: Training and validation loss curves during pre-training for the three models (PT-base, PT-RLM and PT-RLS).

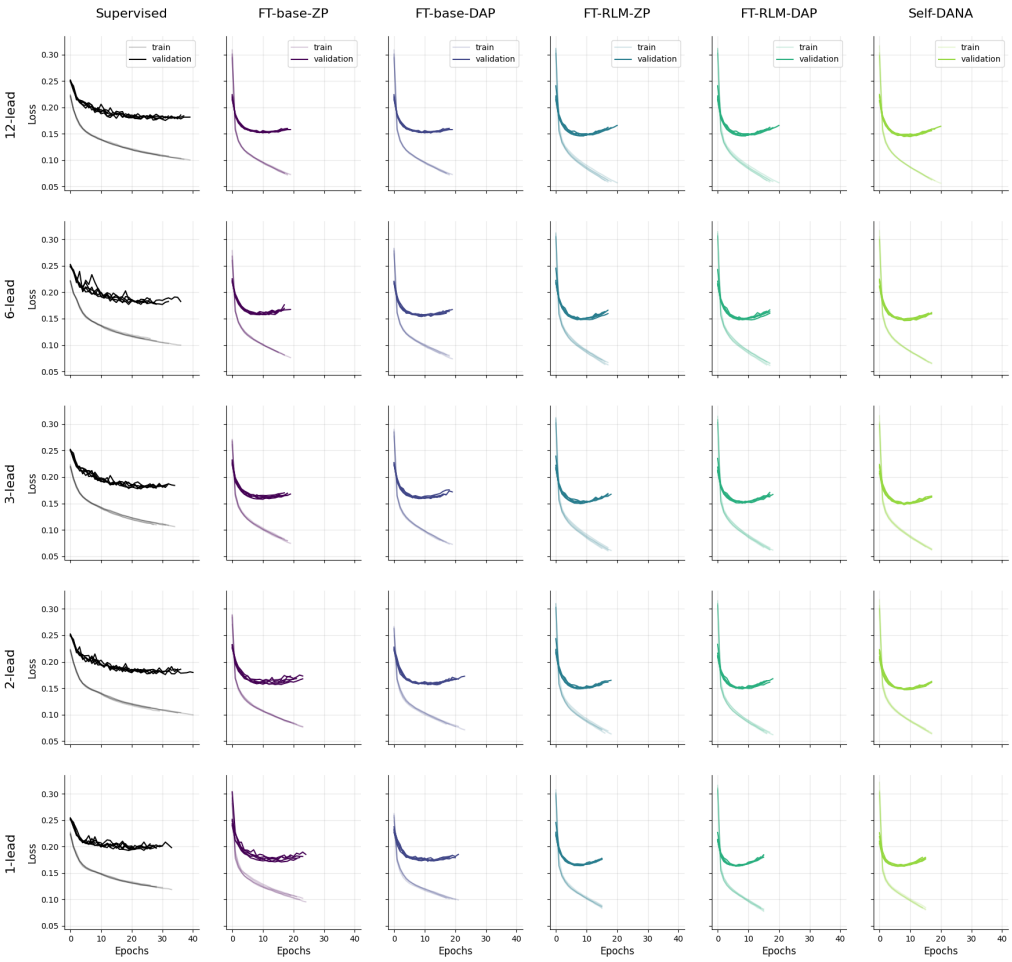

Figure 7: Training and validation loss curves during fine-tuning for all the models and reduced-lead configurations of our experiments.

Other single-lead models Table 11 reports the performance obtained by repeating our experiments with each of the 12 standard ECG leads in a single-lead setting. Self-DANA achieves the best performance for all 12 single-lead configurations, except for I, III, and aVF, for which it is comparable to the best results.

Table 11: CinC score obtained in experiments *(i)*, *(ii)* and *(iii)* on the Georgia test sets of the 12 single-lead configurations. Results are reported as *mean* ± *standard deviation*. Best results (highest CinC score) in bold.

	I	II	III	aVR	aVL	aVF
Supervised	0.547 ±0.003	0.563 ±0.003	0.510 ±0.008	0.569 ±0.006	0.516 ±0.003	0.536 ±0.003
FT-base-ZP	0.562 ±0.004	0.577 ±0.005	0.539 ±0.008	0.579 ±0.001	0.541 ±0.005	0.554 ±0.006
FT-base-DAP	0.568 ±0.005	0.582 ±0.003	0.545 ±0.004	0.582 ±0.003	0.544 ±0.002	0.558 ±0.003
FT-RLM-ZP	**0.585 ±0.004**	0.599 ±0.008	0.560 ±0.004	0.597 ±0.008	0.562 ±0.004	**0.575 ±0.005**
FT-RLM-DAP	0.578 ±0.003	0.596 ±0.010	**0.567 ±0.008**	0.589 ±0.010	0.557 ±0.005	**0.575 ±0.005**
Self-DANA	0.583 ±0.004	**0.600 ±0.005**	0.564 ±0.005	**0.607 ±0.008**	**0.567 ±0.004**	0.574 ±0.005

	v1	v2	v3	v4	v5	v6
Supervised	0.531 ±0.006	0.533 ±0.007	0.541 ±0.005	0.551 ±0.009	0.548 ±0.006	0.545 ±0.007
FT-base-ZP	0.549 ±0.004	0.550 ±0.002	0.559 ±0.003	0.567 ±0.007	0.569 ±0.004	0.572 ±0.003
FT-base-DAP	0.550 ±0.002	0.550 ±0.002	0.560 ±0.008	0.572 ±0.004	0.579 ±0.002	0.575 ±0.002
FT-RLM-ZP	0.568 ±0.004	**0.573 ±0.003**	0.578 ±0.009	0.589 ±0.010	0.590 ±0.006	0.585 ±0.006
FT-RLM-DAP	0.571 ±0.001	0.569 ±0.003	0.573 ±0.006	0.579 ±0.007	0.590 ±0.002	0.584 ±0.006
Self-DANA	**0.572 ±0.011**	**0.573 ±0.004**	**0.584 ±0.005**	**0.596 ±0.005**	**0.597 ±0.004**	**0.595 ±0.003**

Other metrics Table 12, 13, and 14 report the performance obtained on the Georgia dataset in terms of macro AUROC, macro F1 score, and weighted F1 score, respectively. Self-DANA always achieves the best performance for all configurations, or performance comparable to the best results, confirming our findings.

Table 12: Macro AUROC obtained in experiments *(i)*, *(ii)* and *(iii)* on the five Georgia reduced-leads test sets. Results are reported as *mean* ± *standard deviation*. Best results (highest macro AUROC) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.808 ±0.010	0.814 ±0.012	0.812 ±0.006	0.832 ±0.007	0.775 ±0.010
FT-base-ZP	0.828 ±0.003	0.813 ±0.003	0.830 ±0.006	0.831 ±0.009	0.800 ±0.006
FT-base-DAP	0.828 ±0.003	0.823 ±0.004	0.829 ±0.007	0.842 ±0.004	0.810 ±0.001
FT-RLM-ZP	0.843 ±0.008	0.836 ±0.008	0.843 ±0.007	0.849 ±0.011	0.812 ±0.006
FT-RLM-DAP	0.843 ±0.008	0.835 ±0.009	0.840 ±0.008	0.848 ±0.010	0.812 ±0.006
Self-DANA	**0.844 ±0.004**	**0.840 ±0.005**	**0.844 ±0.004**	**0.853 ±0.004**	**0.813 ±0.008**

Table 13: Macro F1 score obtained in experiments *(i)*, *(ii)* and *(iii)* on the five Georgia reduced-leads test sets. Results are reported as *mean* ± *standard deviation*. Best results (highest macro F1 score) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.394 ±0.007	0.389 ±0.009	0.386 ±0.002	0.396 ±0.008	0.358 ±0.005
FT-base-ZP	0.427 ±0.004	0.423 ±0.009	0.415 ±0.007	0.421 ±0.004	0.389 ±0.007
FT-base-DAP	0.427 ±0.004	0.427 ±0.002	0.422 ±0.005	0.430 ±0.003	0.397 ±0.007
FT-RLM-ZP	0.450 ±0.008	0.447 ±0.006	0.448 ±0.002	0.448 ±0.004	0.418 ±0.004
FT-RLM-DAP	0.450 ±0.008	0.450 ±0.007	0.445 ±0.005	0.447 ±0.005	0.412 ±0.005
Self-DANA	**0.446 ±0.007**	**0.448 ±0.004**	**0.448 ±0.001**	**0.449 ±0.006**	**0.419 ±0.003**

PTB-XL dataset Table 15, 16, and 17 report the performance obtained on the PTB-XL dataset in terms of macro AUROC, macro F1 score, and weighted F1 score, respectively. For all configurations, Self-DANA achieves the best performance or performance comparable to the best results, confirming the findings obtained with the Georgia dataset.

Table 14: Weighted F1 score obtained in experiments *(i)*, *(ii)* and *(iii)* on the five Georgia reduced-leads test sets. Results are reported as *mean* ± *standard deviation*. Best results (highest weighted F1 score) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.498 ±0.006	0.494 ±0.008	0.486 ±0.002	0.494 ±0.007	0.448 ±0.004
FT-base-ZP	0.540 ±0.003	0.530 ±0.001	0.521 ±0.007	0.523 ±0.004	0.483 ±0.004
FT-base-DAP	0.540 ±0.003	0.537 ±0.004	0.527 ±0.006	0.532 ±0.003	0.489 ±0.003
FT-RLM-ZP	**0.560 ±0.005**	0.558 ±0.004	0.552 ±0.002	0.556 ±0.003	**0.520 ±0.004**
FT-RLM-DAP	**0.560 ±0.005**	**0.559 ±0.005**	0.551 ±0.004	0.553 ±0.004	0.513 ±0.004
Self-DANA	0.557 ±0.006	**0.559 ±0.002**	**0.554 ±0.002**	**0.557 ±0.004**	0.519 ±0.001

Table 15: Macro AUROC obtained in experiments *(i)*, *(ii)* and *(iii)* on the five PTB-XL reduced-leads test sets. Results are reported as *mean* ± *standard deviation*. Best results (highest macro AUROC) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.859 ±0.006	0.859 ±0.002	0.853 ±0.003	0.859 ±0.002	0.825 ±0.001
FT-base-ZP	0.923 ±0.003	0.908 ±0.004	0.913 ±0.004	0.906 ±0.007	0.869 ±0.007
FT-base-DAP	0.923 ±0.003	0.912 ±0.002	0.916 ±0.004	0.912 ±0.003	0.872 ±0.004
FT-RLM-ZP	0.926 ±0.001	0.918 ±0.004	0.922 ±0.004	**0.919 ±0.004**	0.884 ±0.003
FT-RLM-DAP	0.927 ±0.001	0.918 ±0.004	**0.924 ±0.006**	0.918 ±0.003	**0.888 ±0.006**
Self-DANA	**0.929 ±0.004**	**0.920 ±0.008**	0.923 ±0.005	**0.919 ±0.005**	0.884 ±0.006

Table 16: Macro F1 score obtained in experiments *(i)*, *(ii)* and *(iii)* on the five PTB-XL reduced-leads test sets. Results are reported as *mean* ± *standard deviation*. Best results (highest macro F1 score) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.390 ±0.012	0.387 ±0.009	0.377 ±0.005	0.373 ±0.013	0.338 ±0.008
FT-base-ZP	0.437 ±0.007	0.420 ±0.007	0.415 ±0.008	0.412 ±0.006	0.356 ±0.009
FT-base-DAP	0.438 ±0.008	0.418 ±0.012	0.422 ±0.011	0.419 ±0.012	0.372 ±0.011
FT-RLM-ZP	0.438 ±0.003	**0.430 ±0.009**	**0.426 ±0.007**	0.425 ±0.008	0.390 ±0.008
FT-RLM-DAP	0.437 ±0.002	0.424 ±0.006	**0.426 ±0.011**	0.418 ±0.008	**0.391 ±0.012**
Self-DANA	**0.441 ±0.007**	0.429 ±0.017	0.419 ±0.007	**0.431 ±0.010**	0.389 ±0.011

Table 17: Weighted F1 score obtained in experiments *(i)*, *(ii)* and *(iii)* on the five PTB-XL reduced-leads test sets. Results are reported as *mean* $\pm$ *standard deviation*. Best results (highest weighted F1 score) in bold.

	12 leads	6 leads	3 leads	2 leads	1 lead
Supervised	0.711 ±0.007	0.720 ±0.004	0.698 ±0.002	0.702 ±0.006	0.655 ±0.005
FT-base-ZP	0.748 ±0.005	0.743 ±0.004	0.730 ±0.005	0.731 ±0.004	0.671 ±0.004
FT-base-DAP	**0.749 ±0.005**	0.741 ±0.005	0.732 ±0.004	0.734 ±0.005	0.677 ±0.005
FT-RLM-ZP	0.748 ±0.003	**0.749 ±0.004**	0.734 ±0.002	0.740 ±0.004	**0.691 ±0.004**
FT-RLM-DAP	0.747 ±0.003	0.746 ±0.003	**0.736 ±0.005**	0.735 ±0.002	**0.691 ±0.004**
Self-DANA	0.748 ±0.004	**0.749 ±0.007**	0.733 ±0.002	**0.742 ±0.006**	0.690 ±0.002

CinC2017 dataset Table 18, 19, 20, and 21 report the performance obtained on the CinC2017 dataset in terms of AUROC, macro F1 score, weighted F1 score, and accuracy, respectively. Self-DANA achieves performance comparable to or superior to the best results, confirming the findings obtained with the Georgia dataset.

Table 18: AUROC obtained in experiments *(i)*, *(ii)* and *(iii)* on CinC2017 test set. Results are reported as *mean* ± *standard deviation*. Best results (highest AUROC) in bold.

	Lead I
Supervised	0.871 ±0.032
FT-base-ZP	0.916 ±0.008
FT-base-DAP	0.907 ±0.003
FT-RLM-ZP	0.957 ±0.012
FT-RLM-DAP	**0.966 ±0.002**
Self-DANA	0.964 ±0.005

Table 19: Macro F1 score obtained in experiments *(i)*, *(ii)* and *(iii)* on CinC2017 test set. Results are reported as *mean* ± *standard deviation*. Best results (highest macro F1 score) in bold.

	Lead I
Supervised	0.663 ±0.046
FT-base-ZP	0.733 ±0.021
FT-base-DAP	0.724 ±0.016
FT-RLM-ZP	0.805 ±0.015
FT-RLM-DAP	**0.825 ±0.014**
Self-DANA	0.824 ±0.022

Table 20: Weighted F1 score obtained in experiments *(i)*, *(ii)* and *(iii)* on CinC2017 test set. Results are reported as *mean* $\pm$ *standard deviation*. Best results (highest weighted F1 score) in bold.

	Lead I
Supervised	0.848 ±0.032
FT-base-ZP	0.890 ±0.012
FT-base-DAP	0.885 ±0.011
FT-RLM-ZP	0.926 ±0.007
FT-RLM-DAP	**0.934 ±0.006**
Self-DANA	**0.934 ±0.010**

Table 21: Accuracy obtained in experiments *(i)*, *(ii)* and *(iii)* on CinC2017 test set. Results are reported as *mean* $\pm$ *standard deviation*. Best results (highest accuracy) in bold.

	Lead I
Supervised	0.814 ±0.044
FT-base-ZP	0.870 ±0.016
FT-base-DAP	0.863 ±0.015
FT-RLM-ZP	0.916 ±0.008
FT-RLM-DAP	**0.927 ±0.008**
Self-DANA	0.926 ±0.013

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL `https://arxiv.org/abs/2108.07258`.

[2] Federico Del Pup and Manfredo Atzori. Applications of self-supervised learning to biomedical signals: A survey. *IEEE Access*, 11:144180–144203, 2023. ISSN 21693536. doi: 10.1109/ACCESS.2023.3344531.

[3] Yu Han, Xiaofeng Liu, Xiang Zhang, and Cheng Ding. Foundation models in electrocardiogram: A review, 2024. URL `https://arxiv.org/abs/2410.19877`.

[4] Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model, 2024. URL `https://arxiv.org/abs/2408.05178`.

[5] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *12th International Conference on Learning Representations, ICLR 2024*, 12 2023. URL `https://arxiv.org/pdf/2312.05409`.

[6] Han Yu, Peikun Guo, and Akane Sano. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text, 2024. URL `https://arxiv.org/abs/2405.19366`.

[7] George Mathew, Daniel Barbosa, John Prince, and Subramaniam Venkatraman. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *npj Cardiovascular Health 2024 1:1*, 1:1–13, 10 2024. ISSN 2948-2836. doi: 10.1038/s44325-024-00027-5. URL `https://www.nature.com/articles/s44325-024-00027-5`.

[8] Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Hayit Greenspan, Jagat Narula, Benjamin Glicksberg, and Girish N. Nadkarni. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6:1–8, 12 2023. ISSN 23986352. doi: 10.1038/S41746-023-00840-9. URL `https://www.nature.com/articles/s41746-023-00840-9`.

[9] Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D Clifford. Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021. In *Computing in Cardiology*, volume 48, pages 1–4, 2021.

[10] Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D Clifford. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. *Physiological Measurement*, 2022.

[11] Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. *Proceedings of Machine Learning Research*, 158:156–167, 4 2021. ISSN 26403498. URL https://arxiv.org/pdf/2106.04452.

[12] Wenhan Liu, Zhoutong Li, Huaicheng Zhang, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. Dense lead contrast for self-supervised representation learning of multilead electrocardiograms. *Information Sciences*, 634:189–205, 7 2023. ISSN 0020-0255. doi: 10.1016/J.INS.2023.03.099. URL https://www.sciencedirect.com/science/article/pii/S002002552300422X.

[13] Jungwoo Oh, Hyunseung Chung, Dong-Gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. *Proceedings of Machine Learning Research*, 174:2022, 2022. URL https://github.com/Jwoo5/fairseq-signals.

[14] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Dana: Dimension-adaptive neural architecture for multivariate sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5:120, 8 2020. doi: 10.1145/3478074. URL http://dx.doi.org/10.1145/3478074.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3:1575–1585, 2 2020. URL https://arxiv.org/pdf/2002.05709.

[16] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 2020-December, 6 2020. ISSN 10495258. URL https://arxiv.org/pdf/2006.11477.

[17] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M.M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P.S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Meira Wagner, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature Communications*, 11:1–9, 12 2020. ISSN 20411723. doi: 10.1038/S41467-020-15432-4. URL https://www.nature.com/articles/s41467-020-15432-4.

[18] Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. Code-15 URL https://zenodo.org/records/4916206. v1.0.0; Accessed: 2024-01-17; CC BY 4.0 license.

[19] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications 2023 14:1*, 14:1–13, 6 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39472-8. URL https://www.nature.com/articles/s41467-023-39472-8.

[20] Offline test set of ecg multi-label classfication, 2023. URL https://www.scidb.cn/en/detail?dataSetId=58c4a92d5a01414390a78160d335380d. v1; Accessed: 2024-05-02; MIT license.

[21] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[22] Matthew A Reyna, Nadi Sadr, Annie Gu, Erick A Perez Alday, Chengyu Liu, Salman Seyedi, Amit J Shah, and Gari D Clifford. Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021. PhysioNet, 2022. URL https://physionet.org/content/challenge-2021/1.0.3/. v1.0.3; Accessed: 2024-02-19; CC BY 4.0 license.

[23] Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, Guohua Fu, Hai Yao, Dongbo Li, Hangyuan Guo, and Cyril Rakovski. Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10, 02 2020. doi: 10.1038/s41598-020-59821-7.

[24] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7, 02 2020. doi: 10.1038/s41597-020-0386-x.

[25] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, Eddie Ng, and Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8:1368–1373, 2018. doi: 10.1166/jmihi.2018.2442. URL http://www.icbeb.org/Challenge.html.

[26] Patrick Wagner, Nils Strodthoff, Ralf Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data 2020 7:1*, 7:1–15, 5 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6. URL https://www.nature.com/articles/s41597-020-0495-6.

[27] Patrick Wagner, Nils Strodthoff, Ralf Bousseljot, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3). 2022. doi: 10.13026/kfzx-aw45. URL https://doi.org/10.13026/kfzx-aw45.

[28] Gari D Clifford, Chengyu Liu, Benjamin Moody, Li-Wei H Lehman, Ikaro Silva, Qiao Li, A E Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. doi: 10.22489/CinC.2017.065-469. URL http://physionet.org/.

[29] Petr Nejedly, Adam Ivora, Radovan Smisek, Ivo Viscor, Zuzana Koscova, Pavel Jurak, and Filip Plesinger. Classification of ecg using ensemble of residual cnns with attention mechanism. *Computing in Cardiology*, 48, 2021. doi: 10.22489/CinC.2021.014.

[30] Sahar Soltanieh, Ali Etemad, and Javad Hashemi. Analysis of augmentations for contrastive ecg representation learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2022. doi: 10.1109/IJCNN55064.2022.9892600.

[31] Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141:105114, 2 2022. ISSN 0010-4825. doi: 10.1016/J.COMPBIOMED.2021.105114. URL https://www.sciencedirect.com/science/article/pii/S0010482521009082.

[32] Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Paolo Emilio Mazzon, and Manfredo Atzori. selfeeg. GitHub repository, 2024. URL https://github.com/MedMaxLab/selfEEG. v0.2.0; Accessed: 2024-10-09; MIT license.

[33] Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Paolo Emilio Mazzon, and Manfredo Atzori. Selfeeg: A python library for self-supervised learning in electroencephalography. *Journal of Open Source Software*, 9(95):6224, 2024. doi: 10.21105/joss.06224. URL https://doi.org/10.21105/joss.06224.

[34] fairseq-signals. GitHub repository. URL https://github.com/Jwoo5/fairseq-signals. Accessed: 2024-10-09; MIT license.