Intra-Inter Subject Self-supervised Learning for Multivariate Cardiac Signals

Xiang Lan[1], **Dianwen Ng**[3], **Shenda Hong**[4, 5*], **Mengling Feng**[1, 2*]
[1]Saw Swee Hock School of Public Health, National University of Singapore, Singapore
[2]Institute of Data Science, National University of Singapore, Singapore
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[4]National Institute of Health Data Science, Peking University, Beijing, China
[5]Institute of Medical Technology, Health Science Center of Peking University, Beijing, China
{ephlanx, ephfm}@nus.edu.sg, dianwen001@e.ntu.edu.sg, hongshenda@pku.edu.cn

Abstract

Learning information-rich and generalizable representations effectively from unlabeled multivariate cardiac signals to identify abnormal heart rhythms (cardiac arrhythmias) is valuable in real-world clinical settings but often challenging due to its complex temporal dynamics. Cardiac arrhythmias can vary significantly in temporal patterns even for the same patient ($i.e.$, intra subject difference). Meanwhile, the same type of cardiac arrhythmia can show different temporal patterns among different patients due to different cardiac structures ($i.e.$, inter subject difference). In this paper, we address the challenges by proposing an **I**ntra-inter **S**ubject self-supervised **L**earning (ISL) model that is customized for multivariate cardiac signals. Our proposed ISL model integrates medical knowledge into self-supervision to effectively learn from intra-inter subject differences. In intra subject self-supervision, ISL model first extracts heartbeat-level features from each subject using a channel-wise attentional CNN-RNN encoder. Then a stationarity test module is employed to capture the temporal dependencies between heartbeats. In inter subject self-supervision, we design a set of data augmentations according to the clinical characteristics of cardiac signals and perform contrastive learning among subjects to learn distinctive representations for various types of patients. Extensive experiments on three real-world datasets were conducted. In a semi-supervised transfer learning scenario, our pre-trained ISL model leads about 10% improvement over supervised training when only 1% labeled data is available, suggesting strong generalizability and robustness of the model.

1 Introduction

Cardiovascular disease (CVD) is one of the primary causes of death globally. It has been reported with an estimation of 17.9 million deaths in 2019, representing 32% of the entire global deaths (WHO. 2019). In clinical practice, 12-lead electrocardiography (ECG) is widely adopted to screen overall heart conditions (Kligfield et al. 2007). Abnormal heart rhythms or heartbeats (cardiac arrhythmias) present in ECG may indicate anomaly cardiac functions that could result in severe CVD (Golany and Radinsky 2019). Therefore, early and accurate diagnosis of ECG plays an important role in preventing severe CVD and can improve treatment outcomes (Artis, Mark, and Moody 1991). Deep learning has been applied to improve the timeliness and accuracy of diagnosis of 12-lead ECG (Hong et al. 2020, 2019; Zhu et al. 2021). However, training deep learning models in a supervised manner often requires a large volume of high-quality labels to achieve strong generalization performance. Despite the daily collection of thousands of ECG data by medical institutions, these data are mostly unlabeled. Moreover, it requires intensive labeling on the 12-lead ECG before being able to train a decent model. The work is taxing and requires huge efforts from domain experts, which is costly and not feasible in the real-world clinical setting.

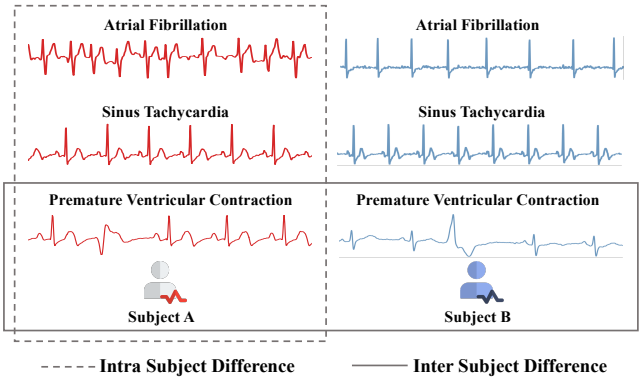

Figure 1: Example of the complex temporal dynamics of cardiac arrhythmias. The three different cardiac arrhythmias vary considerably in temporal patterns even if they all come from subject A ($i.e.$, intra subject difference). Also, because of different cardiac structures, the temporal features of Premature Ventricular Contraction between subject A and subject B are different ($i.e.$, inter subject difference).

Self-supervised learning is a rising field and provides a promising solution to process unlabeled ECG data, whose main idea is to leverage data's innate co-occurrence relationships as the self-supervision to mine useful representations (Liu et al. 2021). Nevertheless, due to the complex temporal dynamics, learning information-rich and generalizable representations from unlabeled multivariate cardiac signals to identify cardiac arrhythmias remains to be a challenge. As

*Corresponding Authors

illustrated in Figure 1, we can observe two categories of differences:

- **Intra Subject**: Different types of cardiac arrhythmias lead to significantly different temporal patterns in the cardia signals even for the same patient. For example, for Subject A in Figure 1, Sinus Tachycardia is associated to a pattern of a faster heart beats; Atrial Fibrillation then leads to a fast and at the same time irregular pattern; Premature Ventricular Contraction then results in abrupt changes to sinus rhythm causing structural changes in the cardic signals.
- **Inter Subject**: The same type of cardiac arrhythmia can show different temporal patterns within different patients because of variations in everyone's cardiac structures. For example, in Figure 1, Subject A has higher QRS voltages ($i.e.$, the narrow spikes in the signal) and larger fluctuation of T waves ($i.e.$, bumps after QRS) than Subject B, even if they are experiencing same cardiac arrhythmias ($e.g.$, Premature Ventricular Contraction). These differences could be due to Subject A has stronger cardiac muscles that produce higher electrical impulses.

In the literature, most of the previous methods were proposed for image data (Chen et al. 2020; Chen and He 2021), and not much focus has been put to the time series data. Thus neither the pretext tasks they used ($e.g.$, predicting image rotation angle) nor data augmentations they applied ($e.g.$, cropping the image into small patches) are suitable for time series such as cardiac signals.

To address these gaps, we propose an **I**ntra-inter **S**ubject self-supervised **L**earning (ISL) model that is customized for multivariate cardiac signals. Our ISL model is an end-to-end model that integrates two different self-supervision procedures: the intra subject self-supervision and the inter subject self-supervision. Both procedures incorporate medical domain knowledge. In **intra subject self-supervision**, based on the fact that our heart rhythm consists of heartbeats, we segment the cardiac signal of each subject into multiple equal length frames ($i.e.$, heartbeat-level time windows). Meanwhile, inspired by the experience of cardiologist in cardiac arrhythmia diagnosis, we design a CNN-RNN encoder with channel-wise attention to extract features from each frame ($i.e.$, heartbeat-level features). After that, we train the encoder to maximize the similarity between similar frames of each subject by leveraging a stationarity test module. In **inter subject self-supervision**, we first design a set of data augmentations according to the clinical characteristics of cardiac signals. Then we fuse heartbeat-level features extracted from the encoder to obtain subject's representations. Lastly, we perform subject-wise contrastive learning to learn distinctive representations.

The main contributions of this work are summarized in below.

- We present ISL, a novel self-supervision model for learning information-rich and generalizable representations from unlabeled multivariate cardiac signals.
- We are the first work that integrates medical knowledge into self-supervision to boost the performance of cardiac arrhythmias diagnosis, which has great value in real-world applications.
- We conducted extensive experiments on three public datasets to enable reproducibility. Experimental results demonstrate ISL outperforms current state-of-the-art methods in the downstream task under various scenarios.

2 Methods

2.1 Overview

In this section, we describe our approach in details. Figure 2 provides an overview of our proposed ISL model. We will elaborate on our methods of intra subject self-supervision and inter subject self-supervision in Section 2.2 and Section 2.3, respectively.

We represent the input multivariate cardiac signal as $\boldsymbol{X} \in \mathbb{R}^{H \times L}$, where H is the number of channels ($i.e.$, the number of leads in multi-lead ECG) and L is the length of the signal. In our experiments, H=12 and L=5,000 ($i.e.$, 10 seconds duration of heartbeats). We divide each signal into N equal length and non-overlapping frames, the $i-$th frame represented as $\boldsymbol{x}_i \in \mathbb{R}^{H \times l}$. We set N=10 and l=500, such that each frame is 1-second duration, which contains about one to two heartbeats.

Our goal is to train an encoder $F_{enc}(\theta_e)$ that projects $\boldsymbol{x}_i$ to an E dimensional latent representation $\boldsymbol{c}_i \in \mathbb{R}^{E \times 1}$ (Eq. 1) and finally obtain the full representation $\boldsymbol{z}_{\mathbf{x}} \in \mathbb{R}^{E \times 1}$ (Eq. 2) of a subject.

$$\boldsymbol{c}_i = F_{enc}(\boldsymbol{x}_i|\theta_e) \tag{1}$$

$$\boldsymbol{z}_{\mathbf{x}} = \mathrm{Fuse}(\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_N\}) \tag{2}$$

2.2 Intra Subject Self-supervision

As shown in Figure 1, the temporal patterns of cardiac arrhythmias can vary significantly even from the same subject. To capture such intra subject differences, we first design a channel-wise attentional CNN-RNN encoder to extract information-rich cross-channel features of the frames ($i.e.$, heartbeat-level features) from each subject (Eq. 1). Then we model the temporal dependencies between frames by utilizing a stationarity test module.

Multivariate Cardiac Signal Encoding. Cardiologists may only look at specific channels for possible abnormal cardiac functions in clinical practice, not all 12 leads. For example, Posterior Wall Myocardial Infarction (MI) is only presented in four chest leads of a 12-lead ECG. This suggests that different cardiac arrhythmias should have different importance weighting for each channel. Therefore, we use the 1-dimensional convolution with channel-wise attention (Hu, Shen, and Sun 2018) for the encoder of our model to extract information-rich cross-channel features. Formally, given an input frame $\boldsymbol{x}$, it is first fed into a 1-dimensional convolutional layer and then compacted using Global Average Pooling (GAP). The weight vector $\boldsymbol{u} \in \mathbb{R}^{E \times 1}$ obtained from GAP can be expressed as

$$\boldsymbol{u} = \mathrm{GAP}(\mathrm{Conv1d}(\boldsymbol{x})) \tag{3}$$

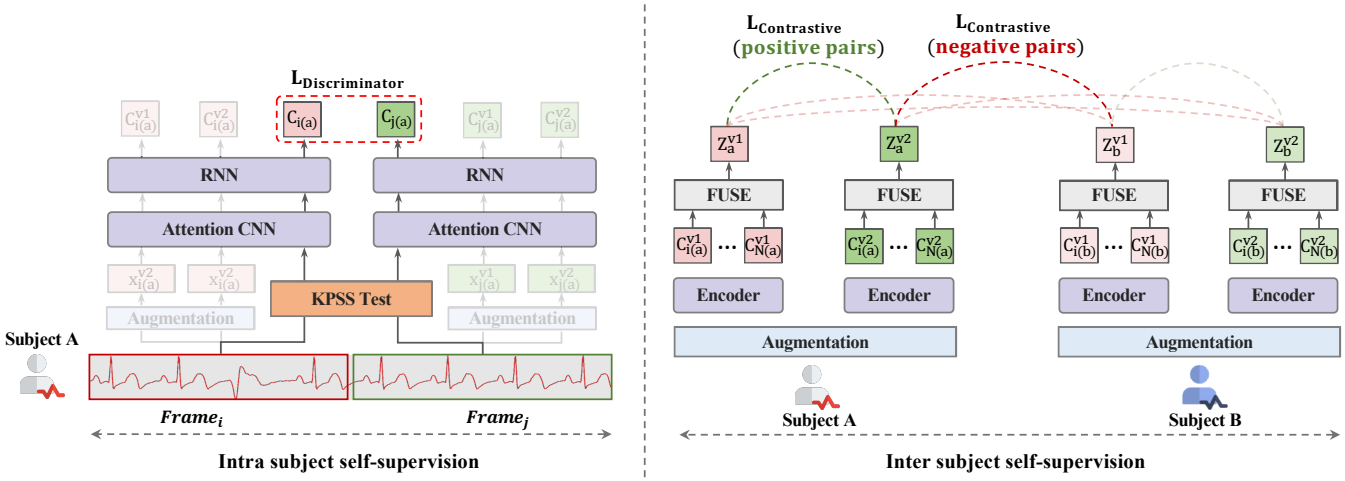

Figure 2: Overview of the ISL model which integrates two self-supervised learning procedures. **Intra subject self-supervision** aims to model the temporal dependencies within a cardiac signal. Take subject A as an example. The cardiac signal is first divided into N equal length frames ($i.e.$, heartbeat-level time windows). Then, ISL uses the KPSS test to determine the stationarity of neighboring frames. The abrupt change occurs in $frame_i$ leading to a non-stationary time series. As such, $frame_i$ and $frame_j$ are treated as dissimilar frames. Meanwhile, ISL utilizes a channel-wise attentional CNN-RNN encoder to extract heartbeat-level features from each frame ($e.g.$, $\boldsymbol{c}_{i(a)}$ and $\boldsymbol{c}_{j(a)}$). These features are then fed into a discriminator to predict the probability of frames being similar. **Inter subject self-supervision** aims to learn distinctive representations among subjects. ISL first creates two views for each cardiac signal using specially designed data augmentations. After that, the heartbeat-level features extracted from the encoder are fused to be subjects' representations. Lastly, ISL performs contrastive learning among subjects to learn distinctive representations. Representations from the same subject are treated as positive pairs ($e.g.$, $\boldsymbol{z}_a^{v1}$ and $\boldsymbol{z}_a^{v2}$), while representations from different subjects are negative pairs ($e.g.$, $\boldsymbol{z}_a^{v2}$ and $\boldsymbol{z}_b^{v1}$). Finally, ISL jointly minimizes the discriminator loss and contrastive loss.

Thereafter, the channel weights $\boldsymbol{s} \in \mathbb{R}^{E\times 1}$ are calculated by

$$\boldsymbol{s} = \text{Sigmoid}(W_2 \cdot \text{ReLU}(W_1 \cdot \boldsymbol{u})) \tag{4}$$

The output $\hat{\boldsymbol{x}} \in \mathbb{R}^{E\times l'}$ from the convolutional layer is then re-weighted to

$$\hat{\boldsymbol{x}} = \boldsymbol{s} \cdot \text{Conv1d}(\boldsymbol{x}) \tag{5}$$

Where $W_1 \in \mathbb{R}^{\frac{E}{r}\times E}$ and $W_2 \in \mathbb{R}^{E\times \frac{E}{r}}$ are learnable parameters (r is the reduction ratio). To better capture the context of a frame, we add a two-layer Recurrent Neural Networks (RNN) on top of the convolutional layers. The final representation $\boldsymbol{c} \in \mathbb{R}^{E\times 1}$ of an input frame is formulated as

$$\boldsymbol{c} = \text{RNN}(\hat{\boldsymbol{x}}) \tag{6}$$

Stationarity Test Module. Next, to capture some cardiac arrhythmias that can cause abrupt changes to the heart rhythm, the intra subject temporal dependencies between frames must be considered. For example, in Figure 2, an abrupt change occurs in $frame_i$ while $frame_j$ remains normal. Even though these two frames are from the same subject, they cannot be regarded as a positive pair since they are not semantically similar. Therefore, the representations between dissimilar frames should be discriminated.

For this purpose, we first assume that abrupt changes in the cardiac signal can lead to a non-stationary time series. Then, we apply the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (Kwiatkowski et al. 1992) to every neighboring frames to identify the occurrence of abrupt changes within the signal. The KPSS test is a statistical method for checking the stationarity of a time series using a null hypothesis that an observable time series is stationary.

In our approach, given a pair of neighboring frames $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$, we calculate the p-value for consecutive time series $(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$ using KPSS test. If the p-value from the test is below the threshold of 0.05, it means that the null hypothesis is rejected and suggests that the time series is non-stationary. We then use the test result as a pseudo label for each neighboring pair, and we train a discriminator $D(\theta_d)$ to predict the stationarity of $(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$ using the concatenated representation $(\boldsymbol{c}_i, \boldsymbol{c}_{i+1})$. The learning objective of the discriminator is to minimize the loss defined in Eq. 7.

$$\mathcal{L}_d = -y \cdot \log(\hat{y}) - (1-y) \cdot \log(1-\hat{y}) \tag{7}$$

Where y and θ_d is the pseudo label from the stationarity test and discriminator's parameters, respectively. $y = 1$ if $(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$ is stationary time series, $y = 0$ otherwise. $\hat{y} = D((\boldsymbol{c}_i, \boldsymbol{c}_{i+1})|\theta_d)$ is the estimated probability from the discriminator for a positive prediction. By doing so, the encoder is encouraged to generate distinguishable representations between similar and dissimilar frames, thus capture the intra subject temporal dependencies.

2.3 Inter Subject Self-supervision

Another challenge is that cardiac arrhythmias can show different temporal patterns among patients due to the different

cardiac structures or functions, even if the cardiac arrhythmias are coming from the same category. To address this challenge, we first fuse the heartbeat-level features (Eq. 1) to obtain a representation (Eq. 2) containing complete information of the cardiac signal ($i.e.$, heart rhythm-level feature). Then we perform contrastive learning among patients to effectively learn distinctive representations from inter subject differences.

Multivariate Cardiac Signal Augmentation. An important component in contrastive learning is data augmentation that provides different views of data. The augmented data should preserve the semantic meaning of the raw data while providing additional information that are targeted to maximize the agreement. However, most prevalent data augmentations are designed for image data that may not be suitable for cardiac signals. Therefore, in addition to regular transformations such as flipping the signal or reversing the magnitude of the signal, we apply explicitly three additional data augmentations for multivariate cardiac signals. (See Appendix A.4 for more details of each augmentation.)

- **Baseline filtering**: We apply Daubechies 5 (db5) wavelet with a decomposition level of 5 to the signal. Then we obtain the signal baseline by reconstructing the signal using the approximation coefficients array from the fifth level decomposition. We use the signal baseline to provide a view that contains morphological information.
- **Bandpass filtering**: We apply Finite Impulse Response (FIR) bandpass filter (Oppenheim, Willsky, and Nawab 1996) to decompose the signal into low, middle, and high-frequency bands. The low-frequency band (0.001-0.5 Hz) preserves the shape of the signal, while the high-frequency band (>50 Hz) is mostly noise. Therefore, we select the middle-frequency band (0.5-50 Hz) as the transformed signal to provide a view of denoised signal.
- **Channel-wise difference**: Inspired by (Mohsenvand, Izadi, and Maes 2020), we subtract each adjacent channel of the cardiac signal to obtain a new channel that represents the voltage difference between two channels. In this way, our model is encouraged to learn the relationships between channels.

Contrastive Learning. Contrastive learning aims to maximize the agreements of different yet relevant views from the same subject since they share the same underlying semantics and are regarded as positive pairs. Views from different subjects are treated as negative pairs, and the similarity between them should be minimized. We first generate two differently augmented data $\boldsymbol{X}^{v1} = g(\boldsymbol{X})$ and $\boldsymbol{X}^{v2} = g(\boldsymbol{X})$ for data $\boldsymbol{X}$, where the augmentations g are randomly selected from our augmentation set $G = \{g_1, g_2, \ldots, g_t\}$. $\boldsymbol{X}^{v1}$ and $\boldsymbol{X}^{v2}$ are then divided into N equal length frames $\{\boldsymbol{x}_1^{v1}, \boldsymbol{x}_2^{v1}, \ldots, \boldsymbol{x}_N^{v1}\}$ and $\{\boldsymbol{x}_1^{v2}, \boldsymbol{x}_2^{v2}, \ldots, \boldsymbol{x}_N^{v2}\}$. After that, the frames are fed into the encoder to extract heartbeat-level features $\{\boldsymbol{c}_1^{v1}, \boldsymbol{c}_2^{v1}, \ldots, \boldsymbol{c}_N^{v1}\}$ and $\{\boldsymbol{c}_1^{v2}, \boldsymbol{c}_2^{v2}, \ldots, \boldsymbol{c}_N^{v2}\}$ using Eq. 3, Eq. 4, Eq. 5, Eq. 6. Once we obtain the heartbeat-level features, we can fuse them to obtain the heart rhythm-level feature, which contains complete information to represent the subject. In our implementation, we simply use the aggregation of heartbeat-level features as the subject's representation, which can be denoted as $\boldsymbol{z}_{\mathbf{x}}^{v1} = \sum_{i=1}^{N} \boldsymbol{c}_i^{v1}$ and $\boldsymbol{z}_{\mathbf{x}}^{v2} = \sum_{i=1}^{N} \boldsymbol{c}_i^{v2}$. For a minibatch with B subjects, representations from the same subject are positive pairs. We treat the other $2B-1$ representations from different subjects as negative pairs. The learning objective is to minimize the contrastive loss as described in Eq. 8.

$$\mathcal{L}_c = -\sum_{m=1}^{B} \log \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{z}_m^{v1}, \boldsymbol{z}_m^{v2}\right)/\tau\right)}{\sum_{k=1}^{2B} \mathbb{1}_{[k \neq m]} \exp\left(\operatorname{sim}\left(\boldsymbol{z}_m^{v*}, \boldsymbol{z}_k^{v*}\right)/\tau\right)} \tag{8}$$

Where τ is the temperature parameter, $\mathbb{1}$ is an indicator function evaluating to 1 iff $k \neq m$. $\operatorname{sim}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}$ calculate the cosine similarity between representation $\boldsymbol{u}$ and $\boldsymbol{v}$.

Lastly, combined with intra-inter subject self-supervision, we jointly minimize the self-supervision loss defined in Eq. 7 and Eq. 8. The end-to-end training procedure of ISL is summarized in Algorithm 1.

Algorithm 1: Self-supervised training procedure of ISL

Input: Pre-training dataset $\mathcal{P} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_p\}$, augmentation set G, pretraining iterations Max_iter, number of frames N.
Parameter: ISL encoder $F_{enc}(\theta_e)$, ISL discriminator $D(\theta_d)$.
Output: Well-trained $F_{enc}(\hat{\theta}_e)$.

Initialize θ_e, θ_d
for $iter = 0$ to Max_iter: **do**
  Sample a mini-batch B from $\mathcal{P}$
  for $\boldsymbol{X}$ in B: **do**
    Random sample two augmentations: $(g_1, g_2) \in G$
    Create two views: $\boldsymbol{X}^{v1}$, $\boldsymbol{X}^{v2} \leftarrow g_1(\boldsymbol{X}), g_2(\boldsymbol{X})$
    Divide signal into equal length frames:
    $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \leftarrow \boldsymbol{X}$
    $\{\boldsymbol{x}_1^{v1}, \boldsymbol{x}_2^{v1}, \ldots, \boldsymbol{x}_N^{v1}\} \leftarrow \boldsymbol{X}^{v1}$
    $\{\boldsymbol{x}_1^{v2}, \boldsymbol{x}_2^{v2}, \ldots, \boldsymbol{x}_N^{v2}\} \leftarrow \boldsymbol{X}^{v2}$
    while $i \leq (N-1)$ **do**
      $p_i = \text{KPSS}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$
      $y_i = \begin{cases} 1, & \text{if } p_i \geq 0.05 \\ 0, & \text{if } p_i < 0.05 \end{cases}$
      $\hat{y}_i = D(F_{enc}((\boldsymbol{x}_i, \boldsymbol{x}_{i+1}), \theta_e), \theta_d)$
      Minimize the loss $\mathcal{L}_d$ in Eq. 7
    end while
    Get $\boldsymbol{c}_i^{v1}$, $\boldsymbol{c}_i^{v2}$ by Eq. 3, Eq. 4, Eq. 5, Eq. 6.
    Get representations of the subject:
    $\boldsymbol{z}_{\mathbf{x}}^{v1} = \sum_{i=1}^{N} \boldsymbol{c}_i^{v1}$, $\boldsymbol{z}_{\mathbf{x}}^{v2} = \sum_{i=1}^{N} \boldsymbol{c}_i^{v2}$
  end for
  Minimize the loss $\mathcal{L}_c$ in Eq. 8
end for
return $\hat{\theta}_e$

3 Experiment Setup

3.1 Datasets

To benchmark the performance of our proposed ISL model and to ensure reproducibility of our results, we pick three

Dataset	Train	Validation	Test	Categories
Chapman	6,352	2,113	2,123	4
CPSC	5,612	1,870	1,818	9
PTB-XL	13,104	4,361	4,370	71

Table 1: Number of samples and categories in each pre-processed dataset used in our experiments.

of the largest publicly available real-world ECG datasets for cardiac arrhythmias classification. We split each dataset into 60%, 20%, 20% in subject-wise for training, validation and testing. Table 1 shows description of each pre-processed dataset. More details of data pre-processing are provided in Appendix A.2.

Chapman. Chapman (Zheng et al. 2020) contains 12-lead ECG recordings of 10,646 patients with a sampling rate of 500 Hz. The dataset includes 11 common cardiac arrhythmias. Each recording length is 10 seconds, and we grouped these cardiac arrhythmias into four categories for a fair comparison with (Kiyasseh, Zhu, and Clifton 2021).

CPSC. CPSC (Liu et al. 2018) contains 12-lead ECG recordings of 6,877 patients with a sampling rate of 500 Hz. The dataset covers nine types of cardiac arrhythmias. The recording length is from 6 seconds to 60 seconds. In our setting, we truncated each record to the same length of 10 seconds.

PTB-XL. PTB-XL (Wagner et al. 2020) contains 21,837 12-lead ECG recordings from 18,885 patients with a sampling rate of 500 Hz. The recording length is 10 seconds and covers total 71 different cardiac arrhythmias.

3.2 Baselines

We compare our method with the following baselines. (1) **Random Init.:** Training a logistic regression model using features extracted from randomly initialized ISL encoder. (2) **Supervised:** Pre-training ISL in a supervised manner. (3) **CPC**(Oord, Li, and Vinyals 2018). (4) **BYOL**(Grill et al. 2020). (5) **SimCLR**(Chen et al. 2020). (6) **SSLECG**(Sarkar and Etemad 2020). (7) **CLOCS**(Kiyasseh, Zhu, and Clifton 2021).

Since CPC, BYOL, and SimCLR are termed as general self-supervised learning frameworks, for a fair comparison, we investigated the performance of these frameworks using the ISL encoder and the basic encoder used in their paper and report their best performance. Furthermore, we applied the same data augmentation as ISL to BYOL and SimCLR since their original augmentations were designed only for image data. The target decay rate of BYOL is set to 0.996. The temperature parameter of SimCLR is set to 0.1. For the implementation of SSLECG and CLOCS, we referred to their published codes during execution. In our comparison, the embedding dimension of all methods are set to E=256.

3.3 Downstream Evaluation Task

Accurate diagnosis of cardiac arrhythmias is of great importance in clinical workflows. Therefore, the downstream task is a multi-label classification of cardiac arrhythmias. Formally, given the subject representation $\boldsymbol{z}_{\mathbf{x}}$, we train a logistic regression model with parameters θ_{lr} to identify all cardiac arrhythmias presented in the cardiac signal $\boldsymbol{X}$ of a subject:

$$\text{Predictions} = \text{LogisticRegression}(\boldsymbol{z}_{\mathbf{x}}|\theta_{lr}) \quad (9)$$

We comprehensively evaluate our ISL model under three scenarios.

Linear Evaluation of Representations. We follow the standard linear evaluation scheme in (Devlin et al. 2019; Chen et al. 2020), where the model is pre-trained on the training set first, then a logistic regression model is trained on top of the frozen encoder to perform downstream task.

Transferability Evaluation of Representations. The transfer learning scenario aims to evaluate the robustness and generalizability of the learned representations. In this setting, the model is pre-trained on one dataset first, then fine-tuned on the other two datasets to perform downstream task, respectively.

Semi-supervised Learning Experiments. To simulate the real-world situation that some medical institutions may have limited labeled data, we investigate the performance of our pre-trained model in a label-scarce scenario. In this setting, the model is pre-trained on one dataset and fine-tuned with different percentages of labeled data on another dataset.

Dataset	Chapman	CPSC	PTB-XL
Supervised	0.990±0.001	0.888±0.012	0.781±0.010
Random Init.	0.825±0.006	0.588±0.007	0.549±0.019
CPC	0.844±0.019	0.711±0.030	0.628±0.027
BYOL	0.592±0.014	0.583±0.042	0.539±0.023
SimCLR	0.771±0.027	0.619±0.008	0.631±0.013
SSLECG	0.526±0.026	0.512±0.014	0.476±0.036
CLOCS	0.906±0.003	0.764±0.011	0.619±0.020
ISL(w/o Inter)	0.764±0.011	0.700±0.009	0.660±0.011
ISL(w/o Intra)	0.921±0.030	0.825±0.034	0.713±0.024
ISL	**0.965±0.008**	**0.854±0.012**	**0.722±0.012**

Table 2: Test AUROC of the linear evaluation.

3.4 Implementation Details

The model is optimized using Adam optimizer (Kingma and Ba 2015) with a learning rate of 3e-3 and weight decay of 4e-4. We use a hard-stop of 40 epochs and a batch size of 232 for both pre-training and downstream tasks, as the training loss does not further decrease. The experiments were conducted using PyTorch 1.8 (Paszke et al. 2019) on NVIDIA GeForce Tesla V100 GPU, we run each experiment 5 times with 5 different seeds, and we report the mean and standard deviation of the test Area Under the Receiver Operating Characteristic Curve (AUROC).

4 Experimental Results

4.1 Linear Evaluation

We conduct a linear evaluation to evaluate the quality of representations learned by ISL. In Table 2, ISL outperforms

Dataset	Chapman	CPSC	PTB-XL
ISL(*E*=32)	0.908±0.012	0.715±0.017	0.583±0.021
ISL(*E*=64)	0.930±0.015	0.801±0.014	0.621±0.018
ISL(*E*=128)	0.915±0.034	0.833±0.007	0.633±0.035

Table 3: Test AUROC of the linear evaluation with different embedding dimensions.

other state-of-the-art approaches on all three datasets, and the performance is comparable with supervised training. For example, ISL has a 9% and 10.3% improvement compared with CLOCS on CPSC and PTB-XL datasets, respectively. On the Chapman dataset, the performance gap between ISL and supervised training is only 2.5%. Furthermore, in Table 3, we show that our model performs consistently well even when we reduced the dimension of the representation. Our ISL still achieves a decent AUROC score when the representation's dimension decreased to 32. Lower embedding dimension means less computational consumption and faster pre-training. This property might be useful in real-world applications where some hospitals have limited computational power. Overall, the results of our linear evaluation imply that, in comparison to other state-of-the-art methods, representations learned by ISL are richer in information.

4.2 Transferability Evaluation

We evaluate the generalizability of representations learned by ISL in transfer learning scenarios. Table 4 shows the performance comparisons under six transfer learning settings. In general, ISL outperforms the other state-of-the-art approaches in five out of six settings. We find that ISL brings a 4% and 5% improvement on CPSC and PTB-XL, respectively, compared with supervised training in Table 2. This suggests that the representations learned by ISL are robust and can generalize to other data sources to improve the performance further. We find that BYOL and SimCLR did not perform well in some of the transfer settings. We hypothesize that this may be due to the augmentations used during training. It is likely that BYOL and SimCLR adopts better to the set of augmentations built for images as compared to the ISL's augmentations techniques. We also find CPC performs consistently well compared with BYOL and SimCLR, which could be due to CPC taking intra subject temporal dependencies into account, while BYOL and SimCLR did not. However, CPC ignores the inter subject differences, thus the learned representations are not as generalizable as ISL.

4.3 Semi-supervised Learning Experiments

To further study the effects of different percentages of labeled data, we conduct semi-supervised learning experiments. As shown in Figure 3, we find that ISL shows significant advantages in the label-scarce scenario. For example, the pre-trained ISL has about 10% improvement on the Chapman dataset compared to supervised training when the percentage of labeled data is less than 50%. We can observe that the pre-trained ISL model constantly performs better than supervised training under the semi-supervised learning

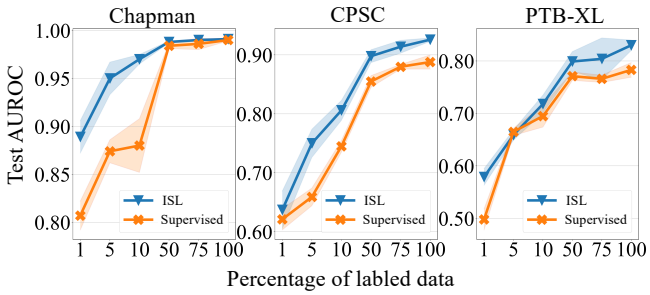

Figure 3: Test AUROC of semi-supervised learning experiments.

scenario on three datasets.

4.4 Ablation Analysis

We study the contribution of the two self-supervision procedures in an ablation analysis as presented in Table 2 and Table 4.

- **w/o Inter**: In this setting, we remove the inter subject contrastive learning, and perform the intra subject self-supervision only. The learning objective is to minimize the loss defined in Eq. 7.
- **w/o Intra**: In this setting, we remove the KPSS Test and the discriminator used in the intra subject self-supervision and perform inter subject contrastive learning only. The learning objective is to minimize the loss defined in Eq. 8.

The results reveal that, in both linear evaluation and transferability evaluation, inter subject self-supervision achieves a better performance than intra subject self-supervision, while intra subject self-supervision provides additional improvements to overall performance. Such results could be explained from two aspects. First, the intra subject self-supervision is based on a pretext task, which may limit the generalizability of the learned representations (Chen et al. 2020). Second, when used solely with the intra subject self-supervision, the performance could be affected by the sample size of cardiac arrhythmias that cause abrupt changes of heartbeats. For example, the number of subjects that have Premature Ventricular Contraction is only 9% of the total subjects in the CPSC dataset, resulting in imbalanced training examples for the discriminator. This issue could be alleviated by training on datasets with more non-stationary cardiac signal samples, such as the PTB-XL dataset. For example, in Table 2, the performance gap between the two self-supervision procedures on PTB-XL dataset is much narrower than on the Chapman and CPSC dataset.

5 Related works

5.1 Self-supervised Contrastive Learning

Self-supervised contrastive learning has recently shown great promises and achieved state-of-the-art performance in many tasks. The motivation of self-supervised contrastive learning is to excavate shared information from different data transformations. For this purpose, multiple views or

Pre-training Dataset	Chapman		CPSC		PTB-XL	
Downstream Dataset	CPSC	PTB-XL	Chapman	PTB-XL	Chapman	CPSC
Supervised	0.797 ± 0.013	0.723 ± 0.008	0.919 ± 0.009	0.757 ± 0.011	0.818 ± 0.050	0.777 ± 0.010
CPC	0.868 ± 0.003	0.824 ± 0.009	0.887 ± 0.004	$\mathbf{0.850 \pm 0.010}$	0.874 ± 0.010	0.884 ± 0.003
BYOL	0.865 ± 0.033	0.695 ± 0.093	0.958 ± 0.049	0.638 ± 0.044	0.885 ± 0.185	0.865 ± 0.064
SimCLR	0.579 ± 0.024	0.568 ± 0.031	0.834 ± 0.071	0.643 ± 0.005	0.574 ± 0.016	0.551 ± 0.034
SSLECG	0.922 ± 0.004	0.784 ± 0.019	0.977 ± 0.005	0.742 ± 0.052	0.977 ± 0.003	0.896 ± 0.023
CLOCS	0.843 ± 0.006	0.740 ± 0.007	0.957 ± 0.004	0.741 ± 0.004	0.948 ± 0.006	0.775 ± 0.003
ISL(w/o Inter)	0.895 ± 0.011	0.762 ± 0.022	0.987 ± 0.002	0.794 ± 0.009	0.989 ± 0.001	0.843 ± 0.106
ISL(w/o Intra)	0.914 ± 0.014	0.819 ± 0.024	0.989 ± 0.002	0.821 ± 0.017	0.989 ± 0.003	0.917 ± 0.011
ISL	$\mathbf{0.928 \pm 0.003}$	$\mathbf{0.831 \pm 0.009}$	$\mathbf{0.991 \pm 0.001}$	0.830 ± 0.010	$\mathbf{0.990 \pm 0.002}$	$\mathbf{0.926 \pm 0.003}$

Table 4: Test AUROC of the transferability evaluation.

data augmentations are created for a sample. The learning objective is to maximize the mutual information between different augmentations of the same sample in the latent space. In view of this methodology, (Tian, Krishnan, and Isola 2020) propose contrastive multiview coding to learn invariant representations between different views of the same scene. (He et al. 2020) present momentum contrast (MoCo) where the feature extractor is trained in a dictionary look-up manner. (Chen et al. 2020) present SimCLR that removed the memory bank. More recently, self-distillation based methods such as BYOL (Grill et al. 2020) and DINO (Caron et al. 2021) are proposed, where negative pairs are no longer necessary.

5.2 Self-supervision for Physiological Signals

Some works have studied self-supervision in time series data (Yue et al. 2021; Mehari and Strodthoff 2021; Spathis et al. 2020; Banville et al. 2019; Ma et al. 2019; Franceschi, Dieuleveut, and Jaggi 2019; Hyvarinen and Morioka 2016). While a few works have recently explored the effectiveness of self-supervision for physiological signals. (Mohsenvand, Izadi, and Maes 2020; Eldele et al. 2021) proposed self-supervised learning frameworks that tested with brain waves data. (Sarkar and Etemad 2020) proposed a multi-task ECG representation learning framework for emotion recognition, where the model is pre-trained by a pretext task of predicting six handcrafted data transformations. (Oord, Li, and Vinyals 2018) presented contrastive predictive coding (CPC) for speech recognition by predicting the near future state in a given utterance.

One work that is relevant to our method is TNC (Tonekaboni, Eytan, and Goldenberg 2021), a general unsupervised learning framework modeling the progression of temporal dynamics of time series. However, TNC pre-train the model by only predicting predefined neighboring relationships between time windows, thus having limited capacity in modeling complex cardiac signals. In comparison, ISL simultaneously models the intra subject temporal dependencies and inter subject differences so that the learned representations are more generalizable and robust. Moreover, TNC is mainly designed for univariate time series and is hard to generalize to 12-lead ECG data.

Another work relevant to our method is CLOCS (Kiyasseh, Zhu, and Clifton 2021), a contrastive learning model designed explicitly for multivariate cardiac signals. ISL differs CLOCS from the following aspects. First, CLOCS assumes that abrupt changes are unlikely to occur in heartbeats in few seconds. Thus all segments from the same cardiac signal share the same context in latent space. In other words, segments from the same subject are positive pairs. CLOCS also assumes that different channels of the same signal are positive pairs because they share the same underlying states. While as illustrated in Figure 1, we did not make such assumptions for ISL since cardiac arrhythmias' patterns are multifarious. Instead, we consider all possible cases where abnormalities could occur suddenly or in specific ECG leads so that the learned representations can be more comprehensive. Furthermore, CLOCS directly learns subject's representations from the entire cardiac signal to perform contrastive learning. In contrast, ISL first learns heartbeat-level features and later fuses them to represent the subject. In this way, ISL is able to capture full-scale temporal dynamics of cardiac signals.

6 Conclusion

In this paper, we propose a novel self-supervision model, ISL, for learning information-rich and generalizable representations from unlabeled multivariate cardiac signals to improve cardiac arrhythmias diagnosis. Our ISL model integrates intra subject self-supervision and inter subject self-supervision. The intra subject self-supervision procedure addresses the issue that temporal patterns differ considerably between cardiac arrhythmias, even from the same patient. The inter subject self-supervision procedure addresses the problem that the same type of cardiac arrhythmia shows different temporal patterns between patients due to different cardiac structures. Extensive experiments on three real-world datasets were conducted, the results over different evaluation scenarios show that the representation learned by ISL is information-rich and more generalizable than other state-of-the-art methods. Moreover, the results in label-scarce scenario suggest strong potential of ISL in real clinical applications.

7 Acknowledgments

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2020-055 and AISG-GC-2019-001-2A), and National Natural Science Foundation of China (No.62102008).

References

Artis, S.; Mark, R.; and Moody, G. 1991. Detection of atrial fibrillation using artificial neural networks. In *[1991] Proceedings Computers in Cardiology*, 173–176.

Banville, H.; Albuquerque, I.; Hyvärinen, A.; Moffat, G.; Engemann, D.-A.; and Gramfort, A. 2019. Self-supervised representation learning from electroencephalography signals. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15750–15758.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. *IJCAI*.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised Scalable Representation Learning for Multivariate Time Series. *Advances in Neural Information Processing Systems*, 32: 4650–4661.

Golany, T.; and Radinsky, K. 2019. Pgans: Personalized generative adversarial networks for ecg synthesis to improve patient-specific deep ecg classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 557–564.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hong, S.; Xiao, C.; Ma, T.; Li, H.; and Sun, J. 2019. MINA: Multilevel Knowledge-Guided Attention for Modeling Electrocardiography Signals. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5888–5894. International Joint Conferences on Artificial Intelligence Organization.

Hong, S.; Zhou, Y.; Shang, J.; Xiao, C.; and Sun, J. 2020. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 103801.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hyvarinen, A.; and Morioka, H. 2016. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29: 3765–3773.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kiyasseh, D.; Zhu, T.; and Clifton, D. A. 2021. CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5606–5615. PMLR.

Kligfield, P.; Gettes, L. S.; Bailey, J. J.; Childers, R.; Deal, B. J.; Hancock, E. W.; van Herpen, G.; Kors, J. A.; Macfarlane, P.; Mirvis, D. M.; Pahlm, O.; Rautaharju, P.; and Wagner, G. S. 2007. Recommendations for the Standardization and Interpretation of the Electrocardiogram. *Circulation*, 115(10): 1306–1324.

Kwiatkowski, D.; Phillips, P. C.; Schmidt, P.; and Shin, Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3): 159–178.

Liu, F.; Liu, C.; Zhao, L.; Zhang, X.; Wu, X.; Xu, X.; Liu, Y.; Ma, C.; Wei, S.; He, Z.; et al. 2018. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7): 1368–1373.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021. Self-supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Ma, Q.; Zheng, J.; Li, S.; and Cottrell, G. W. 2019. Learning representations for time series clustering. *Advances in neural information processing systems*, 32: 3781–3791.

Mehari, T.; and Strodthoff, N. 2021. Self-supervised representation learning from 12-lead ECG data. *arXiv preprint arXiv:2103.12676*.

Mohsenvand, M. N.; Izadi, M. R.; and Maes, P. 2020. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, 238–253. PMLR.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Oppenheim, A. V.; Willsky, A. S.; and Nawab, S. H. 1996. *Signals & Systems (2nd Ed.)*. USA: Prentice-Hall, Inc. ISBN 0138147574.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037.

Sarkar, P.; and Etemad, A. 2020. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*.

Spathis, D.; Perez-Pozuelo, I.; Brage, S.; Wareham, N. J.; and Mascolo, C. 2020. Learning Generalizable Physiological Representations from Large-scale Wearable Data. *arXiv preprint arXiv:2011.04601*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Multiview Coding. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 776–794. Cham: Springer International Publishing. ISBN 978-3-030-58621-8.

Tonekaboni, S.; Eytan, D.; and Goldenberg, A. 2021. Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding. In *International Conference on Learning Representations*.

Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Kreiseler, D.; Lunze, F. I.; Samek, W.; and Schaeffter, T. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1): 1–15.

WHO. 2019. Cardiovascular diseases (CVDs).

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; and Xu, B. 2021. Learning Timestamp-Level Representations for Time Series with Hierarchical Contrastive Loss. *arXiv preprint arXiv:2106.10466*.

Zheng, J.; Zhang, J.; Danioko, S.; Yao, H.; Guo, H.; and Rakovski, C. 2020. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1): 1–8.

Zhu, Z.; Lan, X.; Zhao, T.; Guo, Y.; Kojodjojo, P.; Xu, Z.; Liu, Z.; Liu, S.; Wang, H.; Sun, X.; and Feng, M. 2021. Identification of 27 abnormalities from multi-lead ECG signals: an ensembled SE_ResNet framework with Sign Loss function. *Physiological Measurement*, 42(6): 065008.