Parallel-Learning of Invariant and Tempo-variant Attributes of Single-Lead Cardiac Signals: PLITA

Adrian Atienza, Jakob E. Bardram, Sadasivan Puthusserypady

Technical University of Denmark
{adar, jakba, sapu}@dtu.dk

Abstract

Wearable sensing devices, such as Holter monitors, will play a crucial role in the future of digital health. Unsupervised learning frameworks such as Self-Supervised Learning (SSL) are essential to map these single-lead electrocardiogram (ECG) signals with their anticipated clinical outcomes. These signals are characterized by a tempo-variant component whose patterns evolve through the recording and an invariant component with patterns that remain unchanged. However, existing SSL methods only drive the model to encode the invariant attributes, leading the model to neglect tempo-variant information which reflects subject-state changes through time. In this paper, we present Parallel-Learning of Invariant and Tempo-variant Attributes (PLITA), a novel SSL method designed for capturing both invariant and tempo-variant ECG attributes. The latter are captured by mandating closer representations in space for closer inputs on time. We evaluate both the capability of the method to learn the attributes of these two distinct kinds, as well as PLITA's performance compared to existing SSL methods for ECG analysis. PLITA performs significantly better in the set-ups where tempo-variant attributes play a major role.

Introduction

The wearable sensing field has seen remarkable advancements in recent years. By enabling real-time data collection on physiological parameters, these devices will play a crucial role in the future of digital health. Among these wearable sensors, the Holter monitor captures the cardiac machinery as single-lead ECG signals. Leveraging the information that is accommodated in these signals has the potential of providing outstanding benefits: (i) Facilitating the early identification of irregular heart rhythms, such as Atrial Fibrillation (AFib), (ii) Simplifying the diagnostic process and minimizing the necessity for comprehensive testing (Himmelreich et al. 2019), and (iii) Enabling users to engage proactively in tracking their heart health by offering instant access to health data and insights (Abdou and Krishnan 2022). Generic models are mandated to map these data with their anticipated clinical outcomes. These models should compute informative single-lead ECG representations applicable to several downstream tasks and be optimized using large volumes of unlabelled data. This makes Self-Supervised Learning (SSL) framework particularly well-suited for addressing this clinical challenge.

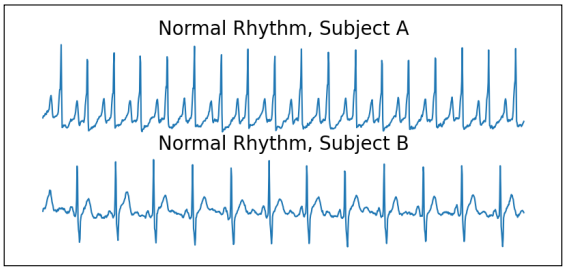


Figure 1: A pair ECG strips from distinct subjects are shown. The signal morphology accommodates a strong stationary component with visible differences between the subjects.

To ensure blood reaches the entire body, each heart periodically repeats a sequence of actions, i.e., contractions, relaxations, and repolarizations, involving its different chambers with clockwise precision. This unique execution of actions results in signals that exhibit a strong stationary component which is distinctive among hearts, as shown in Figure 1. These stationary attributes accommodate meaningful information such as the subject's gender (Attia et al. 2019b) or tendency to cardiac arrhythmia (Attia et al. 2019a). In parallel, the functioning of the heart evolves, since it has to adapt to the individual's temporary needs or it may fall into arrhythmias. This evolution, also captured in the recordings, adds a non-stationary component to the data. Therefore, single-lead ECG signals are characterized by two components of distinct kinds. The stationary component remains unchanging over time. Conversely, the non-stationary component is time-sensitive and its patterns evolve through the recording. This paper will refer to these components as invariant and tempo-variant attributes, respectively.

Current SSL techniques designed for single-lead ECG processing (Diamant et al. 2022; Kiyasseh, Zhu, and Clifton 2021; Wickstrøm et al. 2022) enforce the model to understand and encode the signal invariant patterns following a Contrastive Learning (Chen et al. 2020) approach. Despite each having its uniqueness, they all

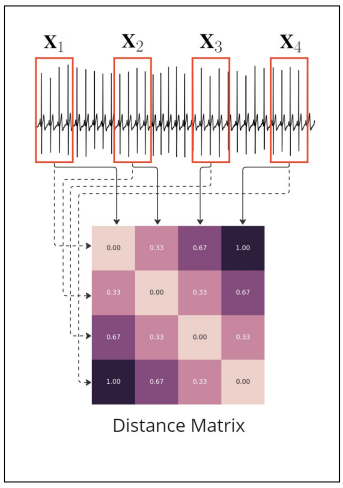


Figure 2: Considering time-sorted inputs equally spaced in time ($X_1 \dots X_4$), representations of nearby inputs in time are expected to be closer than time-distant ones.

consider non-overlapping signal strips from the same subject as positive pairs and enforce similar representations between them. These studies demonstrate that by exploiting the non-stationarity nature of the data, the invariant information of the Single-Lead ECG signals are better captured within the representations rather than creating two versions of the same input using data augmentation techniques. This common strategy is aligned with other time series SSL works, (Qian et al. 2021; Jing et al. 2019; Zhang and Crandall 2021), where non-overlapping frames from the same video are considered as positive pairs during the training procedure.

However, solely focusing on driving the model to capture the invariant attributes only covers a part of the whole picture. In other words, simply mandating similar representations from inputs belonging to the same recording will lead to the model neglecting the tempo-variant attributes, and thereby these changes over time. It leads to a loss of meaningful information that is particularly valuable in specific scenarios aimed at identifying occasional cardiovascular events that occur at irregular intervals throughout the recording, such as detecting AFib or classifying sleep stages.

To address this drawback, this paper presents Parallel-Learning of Invariant and Tempo-variant Attributes (PLITA), a novel SSL method designed to represent both the invariant and the tempo-variant attributes of single-lead ECG signals. The proposed PLITA approach is consistent with methodologies like Split Invariant-Equivariant (SIE) (Garrido, Najman, and Lecun 2023), where the model is designed to capture attributes of two different kinds. PLITA compels its learning model to integrate tempo-variant attributes by ensuring that the representations adhere to a coherent principle: representations of temporally proximate inputs should be closer than those of temporally distant inputs. This concept is depicted in Figure 2 and is encapsulated in the innovative "Tempo-variant Loss Function" ($\mathcal{L}_{tv}$), which forms an integral part of the training objective.

In this study, we hypothesize that: (i) Tempo-variant attributes contain significant information that is distinct from the information conveyed by invariant ones, (ii) Simply using the tempo-variant attributes as a source of natural variance limits the potential of the representation in several downstream tasks, (iii) These attributes can also be incorporated within the representations by the proposed $\mathcal{L}_{tv}$ loss function, and (iv) By encoding these attributes, the model performance improves significantly in set-ups where the tempo-variant attributes play an important role. To assess these hypotheses, we have conducted three experiments that require the invariant or/and the tempo-variant attributes to be encoded within the representations:

1. **AFib Classification.** AFib episodes can be sporadic in time. Moreover, the susceptibility of an individual to this disease is indicated in the baseline signal (Attia et al. 2019a). Therefore, both the invariant and tempo-variant attributes will play a role in this task.
2. **Sleep Stages Classification.** Various sleep stages occur throughout the sleep cycle, regardless of the individual. Consequently, it is essential for the model to capture the tempo-variant attributes to successfully perform this task.
3. **Gender Identification task.** Gender-related information will be persistent within the data throughout the recording. Therefore, driving the model to encode the invariant attributes will be required in this task.

We have evaluated the performance of PLITA against the state-of-the-art (SOTA) methods designed for sinle-lead ECG processing. The findings indicate a marked enhancement in performance on downstream tasks where tempo-variant characteristics are influential. Furthermore, the model demonstrates robust results in gender classification, confirming also the presence of invariant features within the representations. Additionally, we assess whether both tempo-variant and invariant features are reliably captured in the representations and if they fulfill their intended function in downstream tasks.

In summary, the contributions of this paper are:

- We demonstrate through empirical results that the tempo-variant attributes of cardiac signals are not merely a source of natural variation for efficient invariant attribute learning, as assessed in previous studies, but must also be integrated into the representations.
- We introduce PLITA, a novel SSL method that takes apart from existing ECG processing SSL methods by driving the model to encode the tempo-variant attributes via the novel $\mathcal{L}_{tv}$ function.
- We establish a new approach for harnessing tempo-variant features. We hypothesize that this may motivate future SSL work not only applied to ECG analysis but also broadly to other time series data.

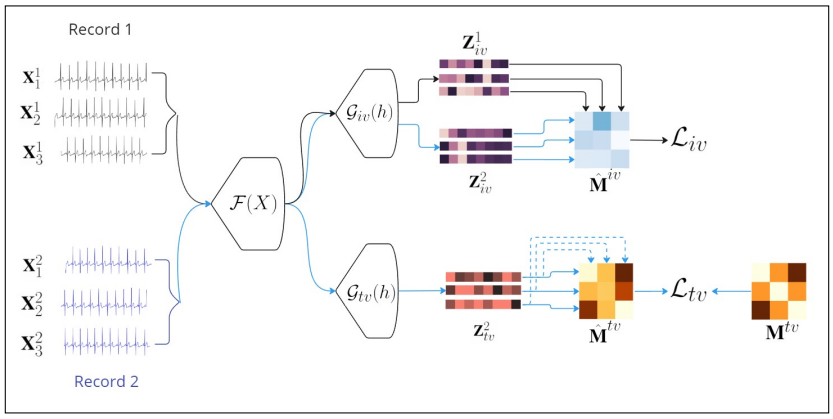


Figure 3: PLITA illustrated. Built on top of BYOL, PLITA includes both a student and a teacher network. For the sake of clarity, the teacher network is not included in the illustration. The losses are computed between representation triplets from both networks that process data equally. While $\mathcal{L}_{iv}$ is computed between a set of N time series representations belonging to different records (displayed in black and blue colors), $\mathcal{L}_{tv}$ is computed between representations belonging to the same record. All inputs belong to the same subject. The encoder ($\mathcal{F}(X)$) is saved at the end of the training procedure and used for downstream tasks.

Related Work

Tempo-Variant attributes in Time Series

Time series data represents the evolution of an object of interest over time. Existing methods (Qian et al. 2021; Jing et al. 2019; Zhang and Crandall 2021) leverage naturally occurring variations in consecutive frames to define positive pairs, avoiding heavy reliance on data augmentation techniques. They provide evidence that this approach enhances the model's effectiveness in handling downstream tasks. Another example of how the availability of a sequence of inputs can improve model training is Siamese Masked Autoencoders work (Gupta et al. 2023), which extends Masked Autoencoders (He et al. 2021) by optimizing the model to reconstruct the subsequent frames instead of the actual one.

SSL in Single-Lead ECG Signal Processing

Existing SSL methods tailored for single-lead ECG data are aligned with SSL methods for video processing. They leverage the data shifts across time for enhancing the learning of the invariant attributes. They all utilize the Contrastive Learning (Chen et al. 2020) as a common framework, considering non-overlapping inputs as positive pairs. The details of the positive-pair selection strategy set these methods apart: (i) The Mixing-Up method (Wickstrøm et al. 2022) introduces a more tailored data augmentation product of two time series from the same recording. (ii) Contrastive Learning of Cardiac Signals Across Space (CLOCS) (Kiyasseh, Zhu, and Clifton 2021) utilizes two consecutive ECG time strips as positive pairs, and (iii) Patient Contrastive Learning (PCLR) (Diamant et al. 2022) which considers two time strips from the same subject but different recordings. PLITA inherits the PCLR strategy by defining positive pairs from distinct recordings to ensure the representation of invariant attributes, as it outperforms the other methods (See Evaluation).

While the previous studies only focus on the invariant attributes, Intra-inter Subject Self-Supervised Learning (ISL) (Lan et al. 2022) mandates representing alterations between consecutive beats. PLITA contemplates the tempo-variant information spotted among temporally sparse inputs, not just between consecutive beats. The proposed method incorporates a novel loss function that drives the model to encode this information by comparing these delayed inputs.

SSL in 12-Lead ECG Signal Processing

The most recent work on the ECG field focuses on 12-Lead signals. Having multiple leads opens up the spatial dimension and thus the range of possibilities when designing methods to process this kind of data (Na et al. 2024; Wang et al. 2023). However, it is not possible to adapt them to single-lead ECG processing. Part of their potential is based precisely on exploiting the spatial dimension, which is not available in the wearable sensing field in which this study is placed. Therefore, they are not included as baselines during the evaluation of PLITA.

Parallel-Learning of Invariant and Tempo-variant Attributes (PLITA)

The aim of PLITA is to simultaneously drive the model to recognize both the invariant and tempo-variant attributes and encode them in the representations. Its workflow is illustrated in Figure 3. Here, *N* inputs equally delayed in time within a window size *W* are sampled from two records belonging to the same subject. The inputs are given to the learning model, displayed as $\mathcal{F}(X)$. The model computes the representations (denoted as *h*), which are passed through the invariant and tempo-variant projectors (denoted as $\mathcal{G}_{iv}$ and $\mathcal{G}_{tv}$, respectively). This workflow is mimicked by the teacher network. The invariant distance matrix, denoted as $\hat{\mathbf{M}}^{iv}$, is calculated

between the invariant projections of both recordings ($\mathbf{z}^1_{iv}$ and $\mathbf{z}^2_{iv}$,). The invariant loss function, $\mathcal{L}_{iv}$, minimizes the values of $\hat{\mathbf{M}}^{iv}$. Parallel to this, the tempo-variant distance matrix, $\hat{\mathbf{M}}^{tv}$, is calculated between the tempo-variant projections from the same recordings ($\mathbf{z}^2_{tv}$). This matrix is compared with the ideal tempo-variant distance matrix ($\mathbf{M}^{tv}$) by the tempo-variant loss function, $\mathcal{L}_{tv}$. The student encoder ($\mathcal{F}(X)$) is the module used when addressing the downstream tasks. The other components are discarded after the training.

Projecting the Invariant and Tempo-variant into Distinct Spaces

PLITA is matched with other existing methods such as SIE, which are designed to integrate various types of attributes into a singular representation. Analogously, PLITA incorporates two projectors ($\mathcal{G}_{iv}$ and $\mathcal{G}_{tv}$) to project the representations into two different spaces and avoid some conflicting goals during the training procedure. These arise from the fact that while $\mathcal{L}_{iv}$ minimizes the distance between the representations, $\mathcal{L}_{tv}$ encourages these distances between time-sorted inputs to occur and follow a spatial order.

The equivariant attributes defined in SIE's method are not related to the tempo-variant ones, nor how they are considered in each respective method. PLITA also differs from SIE in not splitting the representations nor incorporating the $\mathcal{L}_{reg}$ term. Variations in baseline ECGs are crucial for identifying spontaneous episodes of cardiovascular diseases (Attia et al. 2019a). We contend that a holistic consideration of the representation will yield more precise tempo-variant projections. Moreover, the divergent objectives of $\mathcal{L}_{iv}$ and $\mathcal{L}_{tv}$ are anticipated to be adequate in preventing mode collapse between the two projections. Yet the implications of these two decisions have been explored (Refer to Appendix).

Capturing Invariant and Tempo-variant Attributes

This section provides a detailed description of the main technical contribution of this work, which is PLITA's proposed parallel learning for driving the model to capture both the invariant and the tempo-variant attributes of cardiac signals.

Capturing Invariant Attributes: Similarly to PCLR method, PLITA drives the model to encode invariant attributes by employing inputs from different recordings from the same subject. Nevertheless, PLITA approaches this objective in a Non-Contrastive Learning fashion, being built on top of Bootstrap Your Own Latent (BYOL) framework. Both decisions are supported by superior performance in the invariant downstream task of gender classification (see Evaluation). We use the "cosine similarity" for calculating the distance matrix between the invariant projections, denoted as $\hat{\mathbf{M}}^{iv}$ and it defined as the following:

$$\hat{\mathbf{M}}^{iv}_{i,j} = 1 - \frac{(\zeta^1_{iv})^i \cdot \mathcal{Q}_{iv}((\mathbf{z}^2_{iv})^j)}{\max\left(\|(\zeta^1_{iv})^i\|_2 \cdot \|\mathcal{Q}_{iv}((\mathbf{z}^2_{iv})^j)\|_2, \epsilon\right)}, \quad (1)$$

where $(\zeta^1_{iv})^i$ and $\mathcal{Q}_{iv}((\mathbf{z}^2_{iv})^j)$ are the invariant outputs of the teacher and student networks respectively, for the inputs with index i and j drawn from the records 1 and 2. Note that the BYOL framework features both a teacher and a student network. These parallel networks and the student projector are not illustrated in Figure 3 for the sake of clarity. Yet in practice, we calculate each loss function by comparing the output of the student prediction, ($\mathcal{Q}(\mathbf{z})$) with the subsequent output of the teacher projector (ζ). This logic also applies to the $\mathcal{L}_{iv}$ introduced below.

Making the invariant projections similar implies minimizing the values of $\hat{\mathbf{M}}^{iv}_{i,j}$, therefore we define the $\mathcal{L}_{iv}$ as;

$$\mathcal{L}_{iv} = \frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N}\hat{\mathbf{M}}^{iv}_{i,j}. \quad (2)$$

Capturing Tempo-variant Attributes: PLITA requires the model to compute representations in a spatial order aligned with the chronological sequence of time. In other words, the closer the inputs are on time, the closer the representations should be in space. This desired behavior is modeled by the so-called "Ideal Tempo-variant Distances Matrix" ($\mathbf{M}^{tv}$), which is defined as;

$$\mathbf{M}^{tv}_{i,j} = \frac{|i-j|}{N-1}, \quad (3)$$

where i and j are the indices of the inputs sorted in time, and N is the number of inputs considered in each window size. Although we consider the "cosine similarity" as the distance metric between two representations from the same recording, other choices for distance metrics have been evaluated (Refer to Appendix). These pair-wise distances are captured by $\hat{\mathbf{M}}^{tv}$, which is computed as;

$$\hat{\mathbf{M}}^{tv}_{i,j} = 1 - \frac{(\zeta^2_{tv})^i \cdot \mathcal{Q}_{tv}((\mathbf{z}^2_{tv})^j)}{\max\left(\|(\zeta^2_{tv})^i\|_2 \cdot \|\mathcal{Q}_{tv}((\mathbf{z}^2_{tv})^j)\|_2, \epsilon\right)}, \quad (4)$$

where $(\zeta^2_{tv})^i$ and $\mathcal{Q}_{tv}((\mathbf{z}^2_{iv})^j)$ are the tempo-variant outputs of the teacher and student networks respectively, for the inputs with index i and j drawn from the same record. $\hat{\mathbf{M}}^{tv}$ is scaled before calculating the $\mathcal{L}_{iv}$ as the following;

$$\hat{\mathbf{M}}'^{tv} = a + \frac{(\hat{\mathbf{M}}^{tv} - \min(\hat{\mathbf{M}}^{tv}))(b-a)}{\max(\hat{\mathbf{M}}^{tv}) - \min(\hat{\mathbf{M}}^{tv})}, \quad (5)$$

where $a = 1/(N-1)$ and $b = 1$. By scaling $\hat{\mathbf{M}}^{tv}$, we do not only ensure that the values of $\hat{\mathbf{M}}'^{tv}$ and $\mathbf{M}_{tv}$ lie within the same range, but also we alleviate the constraints imposed by $\mathcal{L}_{tv}$. PLITA just mandates the representations to follow a tempo-spatial order without imposing a constant distance for every set of inputs. This constant distance would be a problem since the same magnitude of variance can not be expected for each set in the batch. The final $\mathcal{L}_{tv}$ that enforces the tempo-variant attributes to be represented into the representations is defined as;

$$\mathcal{L}_{tv} = \frac{1}{N(N-1)}\sum_{i}^{N}\sum_{j\neq i}^{N}\left((\mathbf{M}^{tv})_{i,j} - \hat{\mathbf{M}}'^{tv}_{i,j}\right)^2. \quad (6)$$

Note that PLITA does not take into account the diagonal terms, since the invariant features are expected to be modeled by the $\mathcal{L}_{iv}$ loss function. The evaluation of the tempo-variant loss term's integration is presented in (See Ablation).

Evaluation Task	AFIB Classification				Sleep Stage Classification				Gender Identification			
Pre-Train Dataset	**SHHS**		**Icentia**		**SHHS**		**Icentia**		**SHHS**		**Icentia**	
Method / Metric	*Accu. (%)*	*F1 Score*	*Accu. (%)*	*F1 Score*	*Accu. (%)*	*AUC*	*Accu. (%)*	*AUC*	*Accu. (%)*	*AUC*	*Accu. (%)*	*AUC*
PCLR	76.4	73.7	73.7	73.6	71.6	0.75	72.3	0.77	76.4	**0.84**	66.5	**0.74**
Mix-Up	73.4	72.3	62.9	57.3	73.8	0.79	72.6	0.75	70.4	0.76	64.2	0.69
CLOCS	75.7	73.8	73.6	72.7	73.2	0.78	72.0	0.74	70.4	0.76	65.3	0.7
BYOL	76.6	74.8	75.3	72.5	72.9	0.77	73.3	0.78	**76.7**	0.83	**66.7**	0.72
Ti-MAE	72.1	70.9	52.9	60.0	69.3	0.61	69.3	0.66	60.0	0.6	60.0	0.61
Siam Auto	76.5	73.0	53.3	70.1	72.9	0.74	69.3	0.68	73.3	0.8	55.4	0.47
PLITA	**80.7**	**78.4**	**80.0**	**78.2**	**75.3**	**0.81**	**74.8**	**0.8**	76.5	0.83	66.5	0.72

Table 1: Evaluation Results for the three downstream tasks, using the pretrained model trained from both Icentia and SHHS datasets. The bold type indicates the best-performing method for each metric.

Implementation Details

To ensure the replication of the method, we meticulously outline the hyperparameter settings and the model architecture.

Model Architecture: The Vision Transformer (ViT) (Dosovitskiy et al. 2021) model is used for processing the single-lead ECG signals. The input consists of a one dimensional 10-second signal sampled at 100 Hz. The patch size is set to 20. The model counts with 6 regular transformer blocks with 4 heads each and a dimension of 128.

PLITA Implementation and Optimization: The window size W is set to 10 seconds. N is set to 4, so 4 inputs are drawn from each window. The effect of both W and N is discussed in Section . Although we do not incorporate any data augmentation, the effect of it is discussed in the Appendix. The projectors and predictors are implemented as a two-layer Multilayer Perceptron (MLP) with a dimensionality of 512 and 256, respectively. The exponential moving average (EMA) updating factor (τ) is set to 0.995. The training procedure consists of 35,000 iterations. We use a batch size of 256, Adam (Kingma and Ba 2017) with a learning rate of $3e-4$, and a weight decay of $1.5e-6$ as the optimizer. The training procedure and the evaluations are performed on a desktop computer, with a Nvidia GeForce RTX 3070 GPU.

Evaluation

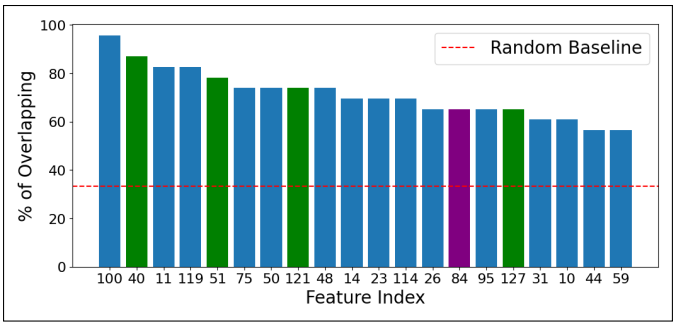
In this section, we evaluate the performance of PLITA compared with the most relevant existing SSL methods for single-lead ECG processing. We also have conducted a study on representations to assess whether the model's learned representations effectively separate the invariant and tempo-variant attributes. Overall, the evaluation involves four distinct databases: MIT-BIH Atrial Fibrillation Database (MIT-AFIB) (Moody and Mark 1983), MIT-BIH Polysomnographic Database (MIT-PSG), (Ichimaru and Moody 1999), Physionet Challenge 2017 (Cinc2017) (Clifford et al. 2017) and Sleep Heart Health Study (SHHS) (Zhang et al. 2018). All databases are publicly available in Physionet (Goldberger et al. 2000) and National Sleep Research Resource (NSRR).

Comparison against state-of-the-art (SOTA)

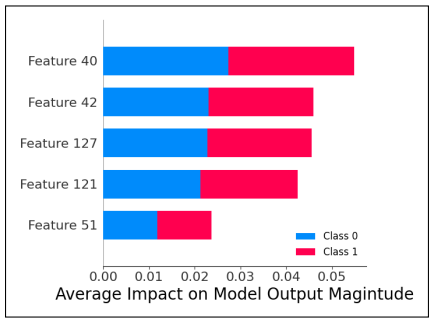
The performance of the proposed PLITA method has been compared against the three most relevant energy-based SOTA methods, namely, (i) PCLR (Diamant et al. 2022), (ii) CLOCS (Kiyasseh, Zhu, and Clifton 2021), and (iii) Mixing-Up (Wickstrøm et al. 2022). Reconstruction methods such as (iv) Ti-MAE (Li et al. 2023) or (v) Siamese Masked Autoencoderss (Gupta et al. 2023) (For more details about this latter implementation, refer to Appendix). Finally, We have also included (vi) the BYOL method tailored by ECG processing by following the PCLR strategy for selecting the positive pairs. To guarantee an equitable assessment, we have optimized the identical model employed in this study, maintaining consistent settings such as the optimizer, data, batch size, and iteration count. Each method has been trained in two distinct datasets, SHHS (Zhang et al. 2018; Quan et al. 1998) and Icentia (Tan et al. 2019), that are composed of long-term single-lead ECG recordings.

AFib Classification: To assess the ability of the method to generalize different classes within the same record, given a limited number of labelled records, we have conducted a Leave-One-Out (LOO) cross-validation across the 23 MIT-AFIB subjects. This ensures no subject-overlapping between the training and validation sets which would significantly simplify AFib identification. A Support Vector Classificatier (SVC) (Platt 2000) is fitted on top of the representations. Table 1 reflects the results, where it can be seen that PLITA significantly outperforms the other methods.

Sleep Stage Detection: We have used the MIT-PSG database in order to assess the capability of the representations to discriminate between Sleep and Wake classes. Since the golden standard classification is performed every 30 seconds and our model has been optimized for processing 10-second signals, a Gated Recurrent Unit (GRU) (Cho et al. 2014) layer is fitted on top of the representations during 5 epochs for processing 30 seconds of data sequentially, in non-overlapping 10-second chunks. The pre-trained model is kept frozen. We have carried out a LOO cross-evaluation for the 18 records contained in the dataset. The outcomes of this analysis are detailed in Table 1. It is evident that PLITA achieves a notably higher level of performance.

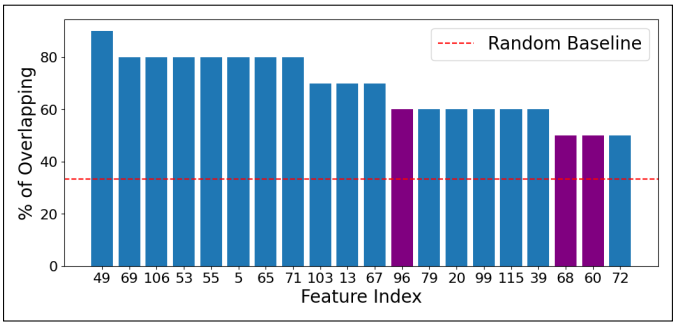


(a) Disentangling Study for invariant features in MIT-AFIB dataset.

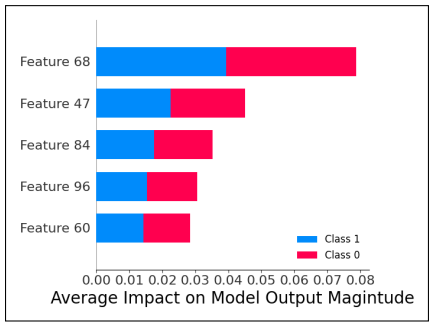


(b) SHAP Analysis in gender classification.

Figure 4: The features that play an important role in the gender classification task (displayed in Figure 4b), are highlighted in green in Figure 4a. The feature that accounts for the AFib classification task (Figure 5b) is displayed in purple.



(a) Disentangling Study for tempo-variant features in MIT-AFIB.



(b) SHAP Analysis in AFib classification.

Figure 5: The informative features in the AFib classification are displayed in Figure 5b and highlighted in purple in Figure 5a.

Gender Classification: We conducted a five-fold cross-validation over 1500 randomly-selected inputs from distinct subjects from the SHHS database. A SVC is fitted on top of the representations. Table 1 shows that despite not achieving the best performance, PLITA reaches competitive results.

Results conclude that training with Icentia data tends to yield worse results, possibly due to increased noise in the data. This decline in performance is more pronounced for reconstruction-based methods. PLITA performs best in tasks that involve encoding tempo-variant attributes. The comparable results achieved in gender classification were also expected due to the only influence of the invariant attributes in this task and the identical manner of BYOL, PCLR and PLITA to drive the model to encode them.

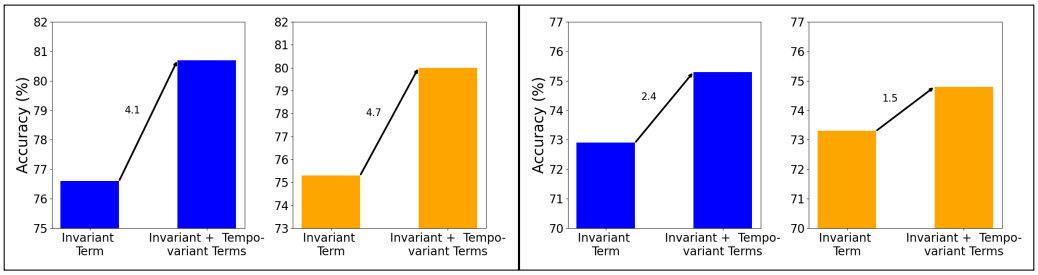
Representation Study

The purpose of this two-phase experimental analysis is to verify that: (i) PLITA prompts the model to encapsulate both invariant and tempo-variant attributes into a suite of discrete features that are consistent across different recordings, and (ii) The collection of invariant features is crucial for the gender classification task, as anticipated. Additionally, the tempo-variant attributes are highly relevant for AFib classification (for more details on the AFib experiment, see Appendix).

Disentangling Study: In order to verify that the two non overlapping sets of features that express the invariant and tempo-variant attributes across the different records are consistent, we have computed the representations of the 23 recordings from the MIT-AFIB dataset. Each feature value is normalized to ensure uniform value ranges. Since it can be assumed that invariant and tempo-variant features will have low and high intra-record variance respectively, this variance is used as a measure of discretization. We cluster the 33% of the features with the least variance as the invariant features while the 33% of them with the highest variance are clustered as the tempo-variant ones. Finally, we tallied the occurrences of each feature in both clusters for each recording.

Figure 4a shows the 20 features that appear most often as invariant features for each recording. The ratio of appearances goes from 95.7% to 56.2% so we can give it a statistical value since in a random baseline, the number of appearances would be around 33.3%. This serves as evidence that the features which represent the invariant attributes of the ECG signals are consistent across recordings. In a similar manner, Figure 5a represents the 20 tempo-variant features, with a ratio of appearances from 90% to 50%. Therefore, it is assessed that the tempo-variant features are also consistent.

SHapley Additive exPlanations (SHAP) Analysis: We have conducted a SHAP analysis (Lundberg and Lee 2017)



(a) AFib Classification. (b) Sleep Stage Classification

Figure 6: Effect of incorporating the tempo-variant attributes in distinct tasks. Blue and Orange bars represent that the model has been pre-trained on SHHS and Icentia respectively.

for the AFib and gender classification tasks. Since the Disentangling study has been carried out in the MIT-AFIB database, we used the AFib and SR instances from Cinc2017, adhering to the training and testing set proposed on it. For the gender task, we used the same dataset used in the evaluation. Figure 4b and Figure 5b reflect how four and three of the five most important features are included among the invariant and tempo-variant sets, respectively. This result is of statistical relevance, since in a random baseline only $6e{-}3$ features would be included in each of the 20-size set of features.

Notably, the third most important feature for AFib detection exhibits an invariant nature (Feature 84). Although at first sight, this may seem contradictory, this aligns PLITA with the findings of other studies (Attia et al. 2019a) which claim that invariant attributes present in ECG baselines enable discretizing the subjects that are susceptible to suffer episodes.

Discussion of the Results

Throughout this comprehensive evaluation, it has been established that PLITA successfully captures both invariant and tempo-variant features within a unified representation. Moreover, this capability allows PLITA to achieve markedly enhanced results in a variety of downstream tasks, as detailed in Table 1. These findings provide robust evidence in favor of the hypotheses posited by this study: (i) Tempo-variant features hold valuable information that is distinct from that of invariant features, (ii) Merely utilizing tempo-variant features as a source of natural variability restricts the representation's effectiveness in numerous downstream tasks, (iii) These features can be integrated into the representations through the proposed $\mathcal{L}_{tv}$ loss function, and (iv) The inclusion of these features leads to a significant improvement in model performance in scenarios where tempo-variant features are crucial.

Ablation and Sensitivity Studies

We have studied both the effect of incorporating the novel $\mathcal{L}_{tv}$ loss function as well as the role of the hyperparameters when computing it. Figure 6 demonstrates that the incorporation of $\mathcal{L}_{tv}$ leads to a significant and positive impact on the tempo-variant related tasks.

The impact of the hyperparameters introduced, i.e the number of inputs from each recording (N) and the window size (W) have also been evaluated. Table 2 indicates that the selected configuration, highlighted in bold, achieves the best performance, but also that all configurations yield superior results compared with existing methods.

Downstream Task		**AFIB Classification**		**Sleep Stage Classification**	
Pre-Train Dataset		**SHHS**	**Icentia**	**SHHS**	**Icentia**
N	**W (Seconds)**	*Accu. (%)*	*Accu. (%)*	*Accu. (%)*	*Accu. (%)*
3	120	79.5	77.6	74.3	73.7
4	**120**	**80.7**	**80.0**	**75.3**	**74.8**
5	120	77.3	78.6	73.2	73.9
4	90	76.6	77.8	74.1	74.5
4	150	76.9	79.4	73.3	72.9

Table 2: Sensitivity Study.

Conclusions

This research provides strong evidence that merely using the tempo-variant attributes of ECG signals as a source of natural variability is insufficient. To overcome this, we introduce PLITA, a novel SSL technique for ECG analysis. By incorporating the $\mathcal{L}_{tv}$ loss function into the training objective, the model is directed to efficiently encode these tempo-variant attributes. This significantly enhances the model's ability to excel in various tasks where such attributes are crucial.

Limitations: PLITA has only been evaluated on a single architecture (ViT). In addition, only BYOL has been utilized to capture the invariant features, leaving aside other non-contrastive frameworks such as Self-Distillation with no Labels (DINO) (Caron et al. 2021) or Variance-Invariance-Covariance Regularization (VIC-REG) (Bardes, Ponce, and LeCun 2022). However, the incorporation of $\mathcal{L}_{tv}$, is agnostic to any of these two we can hypothesize that a similar performance improvement will be obtained for any combination.

Broader Impact: We believe that the incorporation tempo-variant information as a training objective will inspire not only future ECG but also in general time series SSL methods.

References

Abdou, A.; and Krishnan, S. 2022. Horizons in Single-Lead ECG Analysis From Devices to Data. *Frontiers in Signal Processing*, 2.

Attia, Z.; Noseworthy, P.; Lopez-Jimenez, F.; Asirvatham, S.; Deshmukh, A.; Gersh, B.; Carter, R.; Yao, X.; Rabinstein, A.; Erickson, B.; Kapa, S.; and Friedman, P. 2019a. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394.

Attia, Z. I.; Friedman, P. A.; Noseworthy, P. A.; Lopez-Jimenez, F.; Ladewig, D. J.; Satam, G.; Pellikka, P. A.; Munger, T. M.; Asirvatham, S. J.; Scott, C. G.; Carter, R. E.; and Kapa, S. 2019b. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. *Circulation: Arrhythmia and Electrophysiology*, 12(9): e007284.

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. arXiv:2105.04906.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv:1406.1078.

Clifford, G. D.; Liu, C.; Moody, B.; Lehman, L.-w. H.; Silva, I.; Li, Q.; Johnson, A. E.; and Mark, R. G. 2017. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, 1–4.

Diamant, N.; Reinertsen, E.; Song, S.; Aguirre, A. D.; Stultz, C. M.; and Batra, P. 2022. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2): 1–16.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Garrido, Q.; Najman, L.; and Lecun, Y. 2023. Self-supervised learning of Split Invariant Equivariant representations. arXiv:2302.10283.

Goldberger, A.; Amaral, L.; Glass, L.; Havlin, S.; Hausdorg, J.; Ivanov, P.; Mark, R.; Mietus, J.; Moody, G.; Peng, C.-K.; Stanley, H.; and Physiobank, P. 2000. Components of a new research resource for complex physiologic signals. *PhysioNet*, 101.

Gupta, A.; Wu, J.; Deng, J.; and Fei-Fei, L. 2023. Siamese Masked Autoencoders. arXiv:2305.14344.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377.

Himmelreich, J. C.; Karregat, E. P.; Lucassen, W. A.; van Weert, H. C.; de Groot, J. R.; Handoko, M. L.; Nijveldt, R.; and Harskamp, R. E. 2019. Diagnostic Accuracy of a Smartphone-Operated, Single-Lead Electrocardiography Device for Detection of Rhythm and Conduction Abnormalities in Primary Care. *The Annals of Family Medicine*, 17(5): 403–411.

Ichimaru, Y.; and Moody, G. B. 1999. Development of the polysomnographic database on CD-ROM. *Psychiatry and Clinical Neurosciences*, 53.

Jing, L.; Yang, X.; Liu, J.; and Tian, Y. 2019. Self-Supervised Spatiotemporal Feature Learning via Video Rotation Prediction. arXiv:1811.11387.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kiyasseh, D.; Zhu, T.; and Clifton, D. A. 2021. CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients. arXiv:2005.13249.

Lan, X.; Ng, D.; Hong, S.; and Feng, M. 2022. Intra-Inter Subject Self-Supervised Learning for Multivariate Cardiac Signals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4): 4532–4540.

Li, Z.; Rao, Z.; Pan, L.; Wang, P.; and Xu, Z. 2023. Ti-MAE: Self-Supervised Masked Time Series Autoencoders. arXiv:2301.08871.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Moody, G.; and Mark, R. 1983. A new method for detecting atrial fibrillation using R-R intervals. *Computers in Cardiology*, 227–230.

Na, Y.; Park, M.; Tae, Y.; and Joo, S. 2024. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. In *International Conference on Learning Representations*.

Platt, J. 2000. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv. Large Margin Classif.*, 10.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; and Cui, Y. 2021. Spatiotemporal Contrastive Video Representation Learning. arXiv:2008.03800.

Quan, S.; Howard, B.; Iber, C.; Kiley, J.; Nieto, F.; O'Connor, G.; Rapoport, D.; Redline, S.; Robbins, J.; Samet, J.; and Wahl, 998. The Sleep Heart Health Study: Design, Rationale, and Methods. *Sleep*, 20: 1077–85.

Tan, S.; Androz, G.; Chamseddine, A.; Fecteau, P.; Courville, A.; Bengio, Y.; and Cohen, J. P. 2019. Icentia11K: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery. arXiv:1910.09570.

Wang, N.; Feng, P.; Ge, Z.; Zhou, Y.; Zhou, B.; and Wang, Z. 2023. Adversarial Spatiotemporal Contrastive Learning

for Electrocardiogram Signals. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Wickstrøm, K.; Kampffmeyer, M.; Mikalsen, K. Ø.; and Jenssen, R. 2022. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognition Letters*, 155: 54–61.

Zhang, G.-Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; and Redline, S. 2018. The National Sleep Research Resource: Towards a Sleep Data Commons. *Journal of the American Medical Informatics Association*, 572–572.

Zhang, Z.; and Crandall, D. 2021. Hierarchically Decoupled Spatial-Temporal Contrast for Self-supervised Video Representation Learning. arXiv:2011.11261.