Learning Beyond Similarities: Incorporating Dissimilarities between Positive Pairs in Self-Supervised Time Series Learning

Adrian Atienza.
Department of Health Technology
Technical University of Denmark
adar@dtu.dk

Jakob Eyvind Bardram
Department of Health Technology
Technical University of Denmark
jakba@dtu.dk

Sadasivan Puthusserypady
Department of Health Technology
Technical University of Denmark
sapu@dtu.dk

Abstract

By identifying similarities between successive inputs, Self-supervised Learning (SSL) methods for time series analysis have demonstrated their effectiveness in encoding the inherent static characteristics of temporal data. However, an exclusive emphasis on similarities might result in representations that overlook the dynamic attributes critical for modeling cardiovascular diseases within a confined subject cohort. Introducing Distilled Encoding Beyond Similarities (DEBS), this paper pioneers an SSL approach that transcends mere similarities by integrating dissimilarities among positive pairs. The framework is applied to electrocardiogram (ECG) signals, leading to a notable enhancement of +10% in the detection accuracy of Atrial Fibrillation (AFib) across diverse subjects. DEBS underscores the potential of attaining a more refined representation by encoding the dynamic characteristics of time series data, tapping into dissimilarities during the optimization process. Broadly, the strategy delineated in this study holds the promise of unearthing novel avenues for advancing SSL methodologies tailored to temporal data.

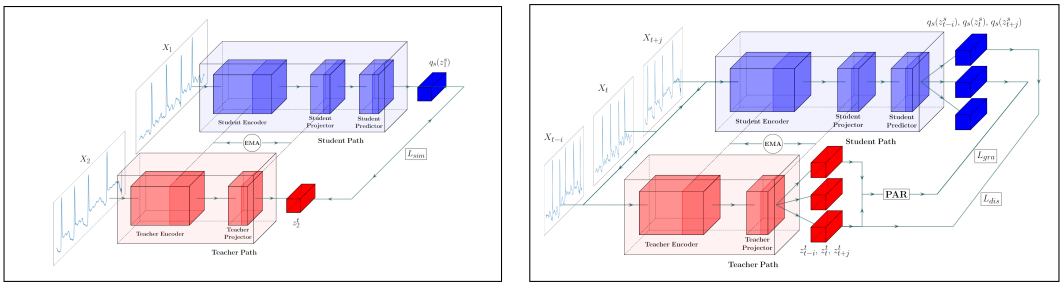
1 Introduction

This paper presents a novel Self-supervised Learning (SSL) method for time series analysis, namely the Distilled Encoding Beyond Similarities (DEBS)[1], with a specific focus on the analysis of physiological signals. The underlying concept of this approach is based on the categorization of signal characteristics into two types: (i) inherent static features that account for individual characteristics such as gender and age, and (ii) dynamic features that can reveal transitional states or events experienced by the subjects during the recording, such as heart arrhythmias in ECG.

As such, DEBS *not only captures what is common but also the changes between two positive pairs.* This is achieved by enhancing the dissimilarities between the positive pairs. We extend the ongoing research trajectory that considers the naturally obtained multiple views as an organic source of variance in order to avoid data augmentation. In addition, we hypothesize that (i) looking solely for similarities can lead the representations to neglect altogether the variance and, therefore, not to encode meaningful dynamic features contained in the data, and (ii) incorporating a focus on the

[1] Throughout this paper, the term distilled is used in its idiomatic sense rather than in the deep learning sense.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Similarity Path. X_1 and X_2 belong to the same subject. In practise, the two inputs are fed into both the teacher and student networks, and the resulting outputs are compared with each other.

(b) Dissimilarity Path. X_{t-i}, X_t and X_{t+i} belong to the same record. In practise, PAR is computed and compared with the X_t representation between networks.

Figure 1: DEBS Architecture. For the sake of clarity, some redundant connections have been omitted.

dissimilarities between positive pairs can result in the representation of the dynamic features and, consequently, an improvement in the performance of downstream tasks such as AFib detection.

In summary, the contributions of this paper are: (i) We introduce DEBS, the first SSL method that enforces the representations of positive pairs to be dissimilar as a part of the objective function. (ii) We show that by incorporating dissimilarities during the optimization process, it increases the AFib detection accuracy over 10%. (iii) We open a new approach in which dissimilarities between positive pairs should be considered for learning the dynamic features of the signals when handling temporal data, such as ECG.

2 Distilled Encoding Beyond Similarities (DEBS)

Traditional SSL studies require the creation of at least one version of the same input in order to train the model to learn an invariant representation with respect to the artificial variance introduced through the use of data augmentation. We consider it a bottleneck in the SSL domain, due to (i) data augmentation methods specifically designed for physiological signals are still an ongoing area of development. Recent studies [18, 13] indicate that the optimal approach to data augmentation in this context has yet to be determined. (ii) Despite achieving a consensus on effective data augmentation procedures, as seen in the field of Computer Vision, the choice of specific data augmentation techniques remains crucial for the success of the SSL method being employed [2, 5].

Instead of creating the new version of the signal utilizing two time series belonging to the same subject, we consider two time series belonging to the same record as positive pairs. Utilizing these organic multiple views and, thereby obviating the need for data augmentation raises an important implicit question. The object of interest evolves across time while exhibiting changes, i.e., the dynamic characteristics of temporal data. A SSL which only considers similarities between these multiple views, will neglect these dynamic characteristics, resulting in a loss of information within the representations and a consequent low performance in identifying events in downstream tasks. Therefore, *learning beyond similarities is essential for capturing the dynamic characteristics of the temporal data.* DEBS represents the first SSL technique to incorporate dissimilarities between positive pairs as part of the objective in addition to similarities during the training process, with the purpose of driving the representations to reflect what has changed within the signal and therefore, capturing the dynamic characteristics within the representation.

2.1 Description of DEBS

Non-Contrastive Method: The omission of the negative pairs enable the proposed SSL methods to surpass the conventional emphasis on similarity. By not considering dissimilarity among negative pairs to avoid mode collapse, dissimilarity can be incorporated among positive pairs. As Bootstrap Your Own Latent (BYOL) [7] framework, DEBS incorporates both a teacher network and a student network. While the student network is optimized using Stochastic Gradient Descent (SGD) with respect to the loss function, the teacher network serves as an exponential moving average (EMA) of the student one, effectively operating as its delayed version. This EMA updating rule is described as:

$$\xi \leftarrow \tau \cdot \xi + (1 - \tau) \cdot \theta, \tag{1}$$

where τ, ξ, and θ are the updating hyperparameter, the teacher weights, and the student weights.

In contrast to the BYOL method, DEBS integrates two projectors within both the student and teacher networks. Consequently, two predictors are also incorporated into the student network, deviating from using a single predictor. As a result, the encoder generates representations that traverse two distinct paths, namely the similarity path and the dissimilarity path, as termed in this work. The rationale behind this design is to enable the first path to capture static features inherent in the representation while the second path to capture the dynamic features.

Similarity path: The objective of the student network's predictor is to produce a representation that closely aligns with the one generated by the same path in the teacher network. It is illustrated in Figure 1a. The degree of similarity serves as one of the cost functions in the proposed method, termed the "Similarity Loss ($\mathcal{L}_{sim}$)", and it is described as the following:

$$\mathcal{L}_{sim}(\mathbf{z}_2^t, \mathbf{q}_s(\mathbf{z}_1^s)) = 1 - \frac{\mathbf{z}_2^t \cdot \mathbf{q}_s(\mathbf{z}_1^s)}{\max\left(\|\mathbf{z}_2^t\|_2 \cdot \|\mathbf{q}_s(\mathbf{z}_1^s)\|_2, \epsilon\right)}, \tag{2}$$

where $\mathbf{z}_2^t$ and $\mathbf{q}_s(\mathbf{z}_1^s)$ are the representation vector and the representation prediction for X_2 and X_1, computed by the teacher and the student network, respectively. ϵ has been set with a value of $1e - 8$ in our work. This similarity path is depicted in Figure 1a.

Dissimilarity path: In contrast to the similarity path, the dissimilarity path aims to predict representations that exhibit differences between two inputs. To achieve this, the "Dissimilarity Loss ($\mathcal{L}_{dis}$) (Eq.(4))" is introduced. It is a cost function specifically designed to guide the optimization process and encourage the model to generate dissimilar representations.

$$\mathcal{L}_{dis}(\mathbf{z}_{t+j}^t, \mathbf{q}_s(\mathbf{z}_{t-i}^s)) = 1 + \frac{\mathbf{z}_{t+j}^t \cdot \mathbf{q}_s(\mathbf{z}_{t-i}^s)}{\max\left(\left\|\mathbf{z}_{t+j}^t\right\|_2 \cdot \left\|\mathbf{q}_s(\mathbf{z}_{t-i}^s)\right\|_2, \epsilon\right)}, \tag{3}$$

where $\mathbf{z}_{t+j}^t$ and $\mathbf{q}_s(\mathbf{z}_{t-i}^s)$ are the representation vector and the representation prediction for X_{t+j} and X_{t-i}, computed by the teacher and the student network, respectively.

In addition to $\mathcal{L}_{dis}$, we introduce the "Gradual Loss ($\mathcal{L}_{gra}$)" as a part of the training objective. We consider that it is not only essential for the representations of two time points drawn from the same subject, X_{t-i} and X_{t+j}, to be dissimilar, but also for the representation of X_t to lie between them. In other words, if a subject's state evolves from X_{t-i} to X_{t+j}, the representation of X_t, i.e., z_t, should approximate an intermediate point between these two extremes. This ensures that the temporal evolution is properly captured within the representations. It is described as:

$$\mathcal{L}_{gra}(\mathbf{q}_s(\mathbf{z}_t^s), \mathcal{PAR}(\mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t)) = 1 - \frac{\mathbf{q}_s(\mathbf{z}_t^s) \cdot \mathcal{PAR}(\mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t)}{\max\left(\|\mathbf{q}_s(\mathbf{z}_t^s)\|_2 \cdot \left\|\mathcal{PAR}(\mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t)\right\|_2, \epsilon\right)}, \tag{4}$$

where Pondered Average Representation (PAR) is the approximation of $\mathbf{z}_t$, drawn from $\mathbf{z}_{t-i}$ and $\mathbf{z}_{t+j}$. We do not force $\mathbf{z}_t$ to be equally distant from both $\mathbf{z}_{t-i}$ and $\mathbf{z}_{t+j}$, therefore, we calculate PAR as,

$$\mathcal{PAR}(\mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t) = \frac{\mathbf{z}_{t-i}^t \cdot j + \mathbf{z}_{t+j}^t \cdot i}{i + j}. \tag{5}$$

The overall loss function that is minimized in the dissimilarity path ($\mathcal{L}_{dispath}$) is described as:

$$\begin{aligned}\mathcal{L}_{dispath}(\mathbf{q}_s(\mathbf{z}_t^s), \mathbf{q}_s(\mathbf{z}_{t-i}^s), \mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t) = {} & \alpha \cdot \mathcal{L}_{dis}(\mathbf{z}_{t+j}^t, \mathbf{q}_s(\mathbf{z}_{t-i}^s)) \\ & + \mathcal{L}_{gra}(\mathbf{q}_s(\mathbf{z}_t^s), \mathcal{PAR}(\mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t)),\end{aligned} \tag{6}$$

where α is the dissimilarity coefficient. This dissimilarity path is illustrated in Figure 1b.

3 Experimental evaluation

Comparison against SOTA methods: In this experiment, we evaluate DEBS against three different baselines: (i) PCLR [3], (ii) Mixing-Up [15] and (iii) TF-C [17]. For this evaluation, a Support Vector Classifier (SVC) [11] is fitted on top of the representations, based on samples obtained from the MIT-BIH Arrhythmia Database (MIT-ARR) database [10], and evaluated in two different databases (MIT-BIH Atrial Fibrillation Database (MIT-AFIB) [9] and Computing in Cardiology Challenge 2017 (CINC2017) [1]). All used datasets are publicly available in Physionet [6].

We have optimized the same model used in this work, under the same configuration (optimizer, data, batch size and number of iterations), except for the TF-C method, where their proposed model has been used. This is due to the fact that it requires the use of two encoders instead of one. Note that this model contains approximately 32 million parameters, which is 30x more than our proposed model. To ensure that the model converges, the latter has been optimized over 75K iterations, instead of the 25K iterations proposed in this work. We have saved the model after 25K, 50K and 75K iterations.

Table 1 shows that the proposed method clearly outperform all the baselines in the different databases.

Table 1: Comparison agains SOTA SSL Methods

Dataset	SSL Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
MIT AFIB	PCLR [3]	72.7	65.6	78.9
	Mixing-Up [15]	65.0	60.5	67.2
	TF-C (25K) [17]	72.2	65.0	78.5
	TF-C (50K) [17]	69.8	62.3	76.2
	TF_C (75K) [17]	71.3	65.9	76.4
	DEBS	**77.5**	**75.6**	**79.5**
CINC2017 (Training)	PCLR[3]	63.5	21.2	93.0
	Mixing-Up[15]	68.0	20.5	90.9
	TF-C (25K)[17]	62.4	20.5	92.8
	TF-C (50K)[17]	62.0	20.3	92.7
	TF-C (75K)[17]	62.4	20.6	92.9
	DEBS	**78.2**	**34.0**	**95.3**
CINC2017 (Validation)	PCLR [3]	70.6	43.3	89.2
	Mixing-Up [15]	66.0	36.0	82.8
	TF-C (25K) [17]	69.2	42.0	42.0
	TF-C (50K) [17]	65.5	38.0	87.0
	TF-C (75K) [17]	67.7	40.7	89.9
	DEBS	**81.8**	**59.0**	**92.7**

Discussion of the results: This study has demonstrated that by incorporating dissimilarities during the training process, we can enable both static and dynamic characteristics to be captured within the representation, as they are projected separately into distinct components.It leads to a significant difference in the proposed downstream task at the end of the training procedure. With these results we have shown substantial support for hypothesis: (i) looking solely for similarities can lead the representations to neglect altogether the variance and, therefore, not to encode meaningful dynamic features contained in the data, and (ii) incorporating a focus on the dissimilarities between positive pairs can result in the representation of the dynamic features and, consequently, an improvement in the performance of downstream tasks such as AFib identification.

4 Conclusion

In this paper, we have presented DEBS, the first SSL method that incorporates dissimilarities between positive pairs. We have shown that by incorporating this new objective function, the representations capture not only the static nature of the data but also the dynamic features. It leads to a significant improvement of over 10% when evaluated on AFib classification in dynamic ECG time series data.

References

[1] G. C. amd Chengyu Liu, B. Moody, L. wei H. Lehman, I. Silva, Q. Li, A. Johnson, and R. G. Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. *Compututational Cardioliology*, 2017.

[2] F. Bordes, R. Balestriero, and P. Vincent. Towards democratizing joint-embedding self-supervised learning, 2023.

[3] N. Diamant, E. Reinertsen, S. Song, A. D. Aguirre, C. M. Stultz, and P. Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):1–16, 02 2022. doi: 10.1371/journal.pcbi.1009862. URL `https://doi.org/10.1371/journal.pcbi.1009862`.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] L. Ericsson, H. Gouk, and T. M. Hospedales. Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks, 2022.

[6] A. Goldberger, L. Amaral, L. Glass, S. Havlin, J. Hausdorg, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng, H. Stanley, and P. Physiobank. Components of a new research resource for complex physiologic signals. *PhysioNet*, 101, 01 2000.

[7] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

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

[10] G. Moody and R. Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society*, 20:45–50, 06 2001. doi: 10.1109/51.932724.

[11] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

[12] S. Quan, B. Howard, C. Iber, J. Kiley, F. Nieto, G. O'Connor, D. Rapoport, S. Redline, J. Robbins, J. Samet, and Wahl. The sleep heart health study: Design, rationale, and methods. *Sleep*, 20:1077–85, 01 1998. doi: 10.1093/sleep/20.12.1077.

[13] A. Raghu, D. Shanmugam, E. Pomerantsev, J. Guttag, and C. M. Stultz. Data augmentation for electrocardiograms, 2022.

[14] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999. doi: 10.1162/089976699300016728.

[15] K. Wickstrøm, M. Kampffmeyer, K. Ø. Mikalsen, and R. Jenssen. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognition Letters*, 155:54–61, mar 2022. doi: 10.1016/j.patrec.2022.02.007. URL `https://doi.org/10.1016%2Fj.patrec.2022.02.007`.

[16] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline. The national sleep research resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*, pages 572–572, 08 2018. doi: 10.1145/3233547.3233725.

[17] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency, 2022.

[18] J. Zhu, J. Qiu, Z. Yang, D. Weber, M. A. Rosenberg, E. Liu, B. Li, and D. Zhao. Geoecg: Data augmentation via wasserstein geodesic perturbation for robust electrocardiogram prediction, 2022.

Appendix

This appendix comprises the following components: (i) Pseudocode delineating the proposed method, (ii) Logical reasoning elucidating the rationale behind DEBS, (iii) Implementation details designed to facilitate reproducibility, (iv) Additional evaluation, and (v) Ablation Study.

A Algorithms

Algorithm 1: Similarity Path

```
Input:
    D, K and N                          ▷ Set of time series, Number of iterations and Batch Size
    $\mathbf{f}_s, \mathbf{q}_s$ and $\theta$                ▷ Student Block, Student Predictor and Student Parameters
    $\mathbf{f}_t$, and $\xi$                       ▷ Teacher Block and Teacher Parameters
    $\mathcal{L}_{sim}$, opt and $\tau$             ▷ Similarity Loss Function, Optimizer and EMA update parameter

for $k \leftarrow 0$ to $K$ do
    $\mathcal{B} \leftarrow \{X_i^1, X_i^2 \in D\}_{i=0}^N$     ▷ Sample $N$-sized Batch. $X_i^1, X_i^2$ belongs to the same subject.
    for $X_i^1, X_i^2 \in \mathcal{B}$ do
        $z_1^s, z_2^s \leftarrow \mathbf{f}_s(X_i^1, X_i^2)$                    ▷ Student Block Projections
        $z_1^t, z_2^t \leftarrow \mathbf{f}_t(X_i^1, X_i^2)$                    ▷ Teacher Block Projections
        $\mathcal{L}_i^{sim} \leftarrow 0.5 \cdot (\mathcal{L}_{sim}(\mathbf{q}_s(z_1^s), z_2^t) + \mathcal{L}_{sim}(\mathbf{q}_s(z_2^s), z_1^t))$   ▷ Similarity Loss
    end
    $\partial\theta \leftarrow \sum_{i=0}^N \partial_\theta \mathcal{L}_i^{sim}$                  ▷ Compute loss gradients for $\theta$
    $\theta \leftarrow opt(\theta, \partial_\theta)$                       ▷ Update Student Parameters
    $\xi \leftarrow \tau \cdot \xi + (1 - \tau) \cdot \theta$                  ▷ Update Teacher Parameters
end
```

Algorithm 2: Dissimilarity Path

```
Input:
    D, K and N                          ▷ Set of time series, Number of iterations and Batch Size
    $\mathbf{f}_s, \mathbf{q}_s$ and $\theta$                ▷ Student Block, Student Predictor and Student Parameters
    $\mathbf{f}_t$, and $\xi$                       ▷ Teacher Block and Teacher Parameters
    opt and $\tau$                      ▷ Optimizer and EMA update parameter
    $\mathcal{PAR}$                             ▷ Ponderate Average Representation
    $\mathcal{L}_{dis}$ and $\mathcal{L}_{gra}$                ▷ Dissimilarity and Gradual Loss Function
    $w_{size}$ and $\alpha$                     ▷ Windows Size and Dissimilarity Coefficient

for $k \leftarrow 0$ to $K$ do
    $\mathcal{B} \leftarrow \{X_n^{t-i}, X_n^t, X_n^{t+j} \in D\}_{n=0}^N$        ▷ Sample $X_n^{t-i}, X_n^t, X_n^{t+j}$ from same record
    assert$(i + j \leq w_{size})$
    for $X_n^{t-i}, X_n^t, X_n^{t+j} \in \mathcal{B}$ do
        $z_{t-i}^s, z_t^s, z_{t+j}^s \leftarrow \mathbf{f}_s(X_n^{t-i}, X_n^t, X_n^{t+j})$          ▷ Student Block Projections
        $z_{t-i}^t, z_t^t, z_{t+j}^t \leftarrow \mathbf{f}_t(X_n^{t-i}, X_n^t, X_n^{t+j})$          ▷ Teacher Block Projections
        $\mathcal{L}_n^{gra} \leftarrow 0.5 \cdot (\mathcal{L}_{gra}(\mathbf{q}_s(\mathbf{z}_t^s), \mathcal{PAR}(\mathbf{z}_{t-i}^t, \mathbf{z}_{t+j}^t))$
                $+ \mathrm{L}_{gra}(\mathcal{PAR}(\mathbf{q}_s(\mathbf{z}_{t_i}^s), \mathbf{q}_s(\mathbf{z}_{t+j}^s)), \mathbf{z}_t^t))$          ▷ Gradual Loss
        $\mathcal{L}_n^{dis} \leftarrow \alpha \cdot 0.5 \cdot (\mathcal{L}_{dis}(\mathbf{q}_s(\mathbf{z}_{t-i}^s), \mathbf{z}_{t+j}^t) + \mathcal{L}_{dis}(\mathbf{q}_s(\mathbf{z}_{t+j}^s), \mathbf{z}_{t-i}^t)))$   ▷ Dissimilarity Loss
    end
    $\partial\theta \leftarrow \sum_{n=0}^N (\partial_\theta \mathcal{L}_n^{dis} + \partial_\theta \mathcal{L}_n^{gra})$          ▷ Compute loss gradients for $\theta$
    $\theta \leftarrow opt(\theta, \partial_\theta)$                       ▷ Update Student Parameters
    $\xi \leftarrow \tau \cdot \xi + (1 - \tau) \cdot \theta$                  ▷ Update Teacher Parameters
end
```

B Intuitions behind DEBS

Intuitions behind being a two-step procedure: Our belief is that understanding the changes in the input requires first understanding what remains constant. This principle drives DEBS to undergo two distinct learning phases. In the initial phase, the method focuses on reducing the variance by ensuring similarity among representations, i.e., encoding the static characteristics of the signals. Once these are adequately captured, DEBS proceeds to emphasize the remaining variance that encapsulates information about the dynamic nature of the data. By doing so, the method does not neglect but understands the remaining variance, ultimately capturing the dynamic characteristics of the temporal data.

Intuitions behind the dissimilarity coefficient (α): $\mathcal{L}_{dis}$ serves the purpose of enforcing distinctiveness among representations, thus capturing the dynamic features inherent in the signal. $\mathcal{L}_{dis}$ leverages the *Cosine Similarity* metric, which varies in the range of values $[-1, +1]$, being -1, a completely different representation, and $+1$, a completely equal representation. While this loss function is minimized when the first value is reached, it is not realistic to expect the representations to be entirely dissimilar. This is due to the fact that not only do static features remain constant throughout the signal, but also the dynamic characteristics need to maintain some level of relational information. Hence, DEBS deliberately introduces α as a regularization factor for lowering the weight of this objective.

Intuitions behind $\mathcal{L}_{gra}$: Our approach takes into consideration an additional factor: the need for the representations to capture the temporal context. It is not only essential for the representations of two-time points drawn from the same subject, $\mathbf{X}_{t-i}$ and $\mathbf{X}_{t+j}$, to be dissimilar, but also for the representation of $\mathbf{X}_t$ to lie between them. In other words, if a subject's state evolves from $\mathbf{X}_{t-i}$ to $\mathbf{X}_{t+j}$, the representation of $\mathbf{X}_t$, i.e., z_t, should approximate an intermediate point between these two extremes. This ensures that the temporal evolution is properly captured within the representations.

Intuitions behind the window size: An essential consideration in implementing the method is determining the appropriate spacing window size between $\mathbf{X}_{t-i}$ and $\mathbf{X}_{t+j}$, i.e., how much these inputs may be separated in time. This spacing window size must be large enough to accommodate signal changes, yet narrow enough to contain only a single one. If successive changes occur within this window, it can lead to conflicting directions in which these changes are reflected in the representations, thereby $\mathcal{PAR}(\mathbf{z}^t_{t-i}, \mathbf{z}^t_{t+j})$ may not be aligned with $\mathbf{q}_s(\mathbf{z}^s_t)$.

C Implementation details

Architecture: We use the same model as the one described in Subject-Based non Contrastive Learning (SBnCL). It is an adaptation of the Vision Transformer (ViT) [4] model for performing physiological signals. The input data is a time series of 1000 samples, which is split into patches of size 20. The model counts with 6 regular transformer blocks with 4 heads each. The model dimension is set to 128, for a total of 1,192,616 trainable parameters.

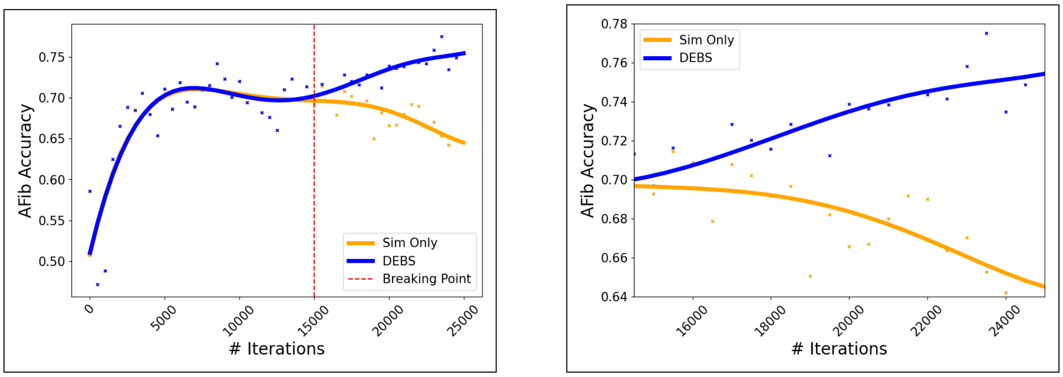
DEBS implementation: The projectors and predictors in our approach are implemented as a two-layer Multilayer Perceptron (MLP). These layers have a dimensionality of 256 and 64, respectively. Batch normalization and rectified linear unit (ReLU) operations are incorporated between the two layers of each structure. The model is trained with the Sleep Heart Health Study (SHHS) dataset [16, 12]. The EMA updating factor (τ) is set to 0.995. The window size is set to 2 minutes, and the $\mathcal{L}_{diss}$ coefficient is set to 0.1. The value of these last two hyper-parameters is presented in Section E.

Optimization: The training procedure consists of 25,000 iterations. After 15,000 iterations, dissimilarities are integrated into the objective function, while similarities are no longer taken into account. Before starting the second step, we update the teacher weights as a copy of the current student weights. The effect of this two-step procedure is discussed in Section E. We use a batch size of 256, and Adam [8] with a learning rate of $3e-4$ and a weight decay of $1.5e-6$ as the optimizer. The training procedure and the subsequent evaluations are performed on a local computer, with a Nvidia GeForce RTX 3070 GPU.

D Further Evaluation

D.1 The effect of incorporating dissimilarities:

To comprehend the impact of integrating the dissimilarity path, an analysis is conducted on the performance of the MIT-ARR → MIT-AFIB framework during the optimization process, with evaluations performed every 500 iterations. The results, illustrated in Figure 2, highlight that exclusive emphasis on similarities leads to a degradation in model performance. Conversely, the incorporation of dissimilarities contributes to a consistent enhancement in model performance, which leads to a difference of + 10% at the end of the training procedure.

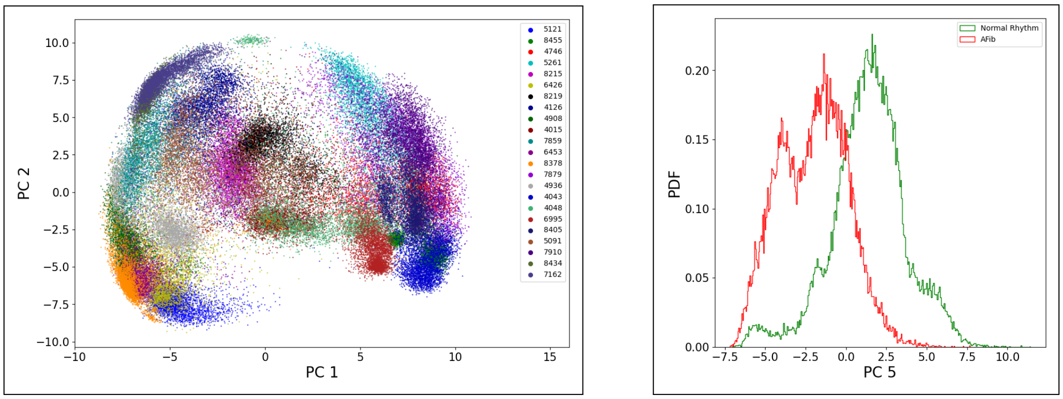


(a) Accuracy across training procedure

(b) Effect of incorporating dissimilarities

Figure 2: Atrial Fibrillation (AFib) classification accuracy across training procedures. For an easier track of the evolution across the iterations, a polynomial has been fitted according to the obtained metrics.

D.2 Principal Component Analysis (PCA) on the representations:



(a) Subject-based PCA Analysis

(b) Event-based PCA Analysis

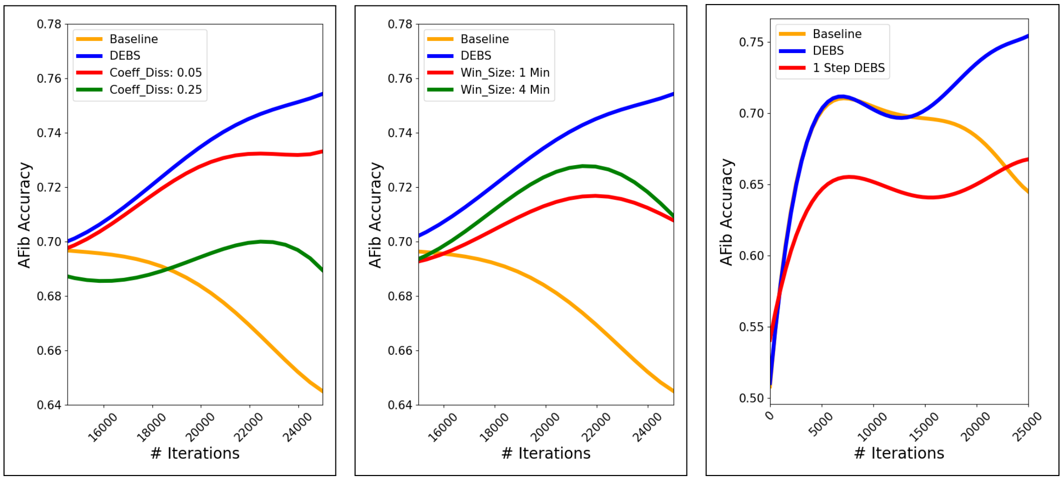
Figure 3: PCA Analysis on MIT-AFIB representations

To complement the previous evaluation, we performed a PCA [14] on the MIT-AFIB representations. We hypothesize that DEBS can drive the model to contain both the static characteristics and the dynamic characteristics of the temporal data within the representations. In this evaluation, the subject characteristics are considered static characteristics since they are constant during the record. Furthermore, we consider AFib as an event that can be characterized by the dynamic features of the time series data, since a subject can suffer it or not during the same record. For assessing this hypothesis, these two features of the signal should be projected in different PCA components. Figure

3 represents the results of this analysis. Figure 3a shows that the inputs belonging to the same subject obtain similar values for the first two components. Figure 3b demonstrates that the dynamic characteristics are also captured in the representations. The values of the 5^{th} component exhibit two distinct nearly normal distributions, corresponding to the events of AFib and Normal Rhythm. This behaviour is not seen when the dissimilarities are not incorporated.

E Ablations Study

In order to investigate the impact of various hyperparameters employed by DEBS. Specifically, we examine the influence of the window size, dissimilarity coefficient, and the choice between a one-step optimization process, in which similarities and dissimilarities are concurrently considered or a two-step optimization process, in which we first consider the similarities, and then the dissimilarities. Figure 4 illustrates the results, revealing that while all configurations exhibit improvements over the baseline (which considers only similarities), the proposed configuration (window size=2 min, dissimilarity coefficient=0.1, and two-step process) yields the most favorable outcomes.



(a) Effect of dissimilarity coefficient, α

(b) Effect of window size

(c) Effect of learning in two steps

Figure 4: Ablations Study. For an easier track of the evolution across the iterations, a polynomial has been fitted according to the obtained metrics.

F Limitations:

While we assert the potential applicability of the DEBS approach to physiological data in general, we acknowledge that our experiments have been limited to the analysis of ECG data. Nevertheless, we posit that DEBS can be adapted to handle diverse types of time series data by adjusting hyperparameters such as window size and dissimilarity coefficient.