PiRL: Participant-Invariant Representation Learning for Healthcare

Zhaoyang Cao
Department of Computational Applied Mathematics and Operations Research
Rice University
Houston, TX 77030
zc48@rice.edu

Han Yu, Huiyuan Yang, Akane Sano
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77030
{hy29, hy49, akane.sano}@rice.edu

Abstract

Due to individual heterogeneity, performance gaps are observed between generic (one-size-fits-all) models and person-specific models in data-driven health applications. However, in real-world applications, generic models are usually more favorable due to new-user-adaptation issues and system complexities, etc. To improve the performance of the generic model, we propose a representation learning framework that learns participant-invariant representations, named PiRL. The proposed framework utilizes maximum mean discrepancy (MMD) loss and domain-adversarial training to encourage the model to learn participant-invariant representations. Further, a triplet loss, which constrains the model for inter-class alignment of the representations, is utilized to optimize the learned representations for downstream health applications. We evaluated our frameworks on two public datasets related to physical and mental health, for detecting sleep apnea and stress, respectively. As preliminary results, we found the proposed approach shows around a 5% increase in accuracy compared to the baseline.

1 Introduction

Deep learning models have been developed using time-series data for solving health-related problems. For example, for heart diseases, Oh *et al.* proposed an automated system that combines a convolutional neural network (CNN) and long short-term memory network (LSTM) for the diagnosis of arrhythmia [15]. Erdenebayar *et al.* designed a deep neural network, recurrent neural networks, and gated-recurrent unit to distinguish apnea and hypopnea events using an electrocardiogram (ECG) signal [4]. Additionally, for mental health, Yu and Sano applied semi-supervised learning on leveraging unlabeled data to estimate the wearable-based momentary stress [24]. Radhika *et al.* proposed the frameworks that investigate the effectiveness of transfer learning and deep multimodal fusion on CNN stress detection models [17] [18].

As shown in the aforementioned studies, the deep learning methods have achieved some promising results in health applications. At the same time, researchers have observed that due to the heterogeneity among data including labels, person-specific models outperform generic models [6, 9, 12, 14, 20, 25]. For instance, Bsoul *et al.* showed that the accuracy of the subject-dependent sleep apnea classification model is 6% higher than that of the subject-independent model [3]. Moreover, Nath *et al.* [13]

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

showed a performance gap of 22.5% in accuracy between the subject-dependent and the generic LSTM models in emotion recognition. Although person-specific models have been widely proven to outperform the generic models in health applications [8, 10, 22], we cannot neglect its drawbacks. First, person-specific models cannot be easily extended to datasets from new populations the models have not seen yet [14]. Also, it is expensive to collect enormous datasets from individuals to build person-specific models.

Researchers have explored improving the performance of generic models by introducing person-specific information. For example, Radhika *et al.* used person specific information in the testing set during the feature extraction and selection [17, 19]. Bethge *et al.* utilized MMD loss to impose domain-invariant representations for emotion classification tasks[2], where each participant had his/her own private encoder with a classifier shared among all. Therefore, even though the performance improvements were observed in the generic models, problems still exist in the studies mentioned above because individual encoders result in high computational costs and difficulties in adapting to other subjects.

In this work, we aim to improve the performances of generic models without introducing person-specific information into the model, thus avoiding the aforementioned drawbacks of the person-specific methods. We propose a representation learning method to learn participant-invariant features, which alleviates the heterogeneous issues by integrating the maximum mean discrepancy (MMD) loss in representation learning to minimize the distribution shifts among features from different subjects. Additionally, we use domain classification loss from domain-adversarial training of neural networks (DANN) architecture, which aimed to blur the participant-distinguishable information among the learned representations, to make label predictor more robust to the target participant [2]. Further, the triplet loss, which aims to learn the label-distinguishable embedding, is used in part of the framework as another constraint to avoid trivial solutions in learned representations. We evaluate the proposed method using two public datasets, including downstream tasks: sleep apnea detection and stress detection. Our results suggest that the proposed method can help improve the model performance significantly compared to the baseline model with only an auto-encoder and a supervised learning model without any constraints.

2 Methodology

We propose a representation learning framework that aims to extract participant-invariant representations, named PiRL. The main PiRL framework is visualized in Figure 1 and detailed architecture is shown in Appendix A. We employ a 1D CNN-based auto-encoder structure as a deep feature extractor from input wearable data. On top of the auto-encoder, an MMD loss and domain classification loss is utilized to constrain the representations from distribution shifts to encourage the learning embedding to be participant-invariant. During training for downstream tasks, we optimize the model with a triplet loss for label-distinguishable representations. The following subsections will introduce the aforementioned components in detail.

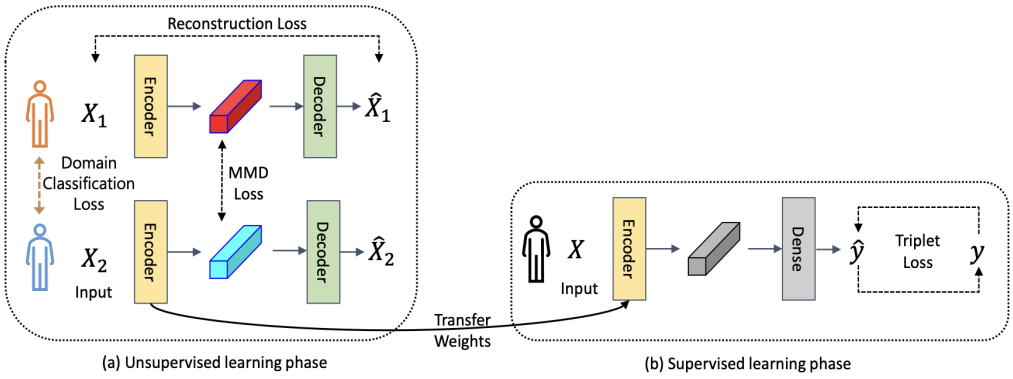

Figure 1: PiRL Network Architecture

2.1 Representation learning

We utilize a 1D CNN-based auto-encoder to extract learning representations from raw time-series sequences X. The encoder extracts latent representations from input sequences as e with a series of 1D CNN layers; whereas the decoder aims to output the reconstructed signal as $\hat{X}$ from e using up-sampling layers. The objective of the auto-encoder is:

$$\mathcal{L}_{ae} = \|X - \hat{X}\|_2^2 \tag{1}$$

2.1.1 MMD loss

To encourage the model to learn the participant-invariant representations, we constrain the model with an MMD loss function, which is widely used in eliminating distribution shifts among different groups of data [7].

Given training samples from two different subjects as X_i and X_j with the total number of subjects N, the MMD loss can be considered as follows:

$$\mathcal{L}_{mmd}(p, q, \mathcal{H}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sup_{f \in \mathcal{H}, ||f||_{\mathcal{H}} \leq 1} (\mathop{\mathbb{E}}_{p(X_i)} [f(X_i)] - \mathop{\mathbb{E}}_{q(X_j)} [f(X_j)])$$

where *p* and *q* are the distributions of variable X_i and X_j, and $\mathcal{H}$ is the Hilbert space. During the training process, the model is optimized to minimize the distribution distances between the pairs of subjects. Thus, the overall objective of unsupervised learning is:

$$\mathcal{L} = \mathcal{L}_{ae} + \lambda \cdot \mathcal{L}_{mmd} \tag{2}$$

where λ is the coefficient of MMD loss and set to 0.2.

2.1.2 Domain classification loss

A domain classifier from DANN architecture is a potent tool to make the label predictor more robust in target data by accomplishing the following adversarial tasks: minimizing the loss of label prediction and maximizing the loss of domain classification. Subsequently, with the addition of a gradient reversal layer before the domain classifier, the overall objective function is the sum of two minimizing problems [5]. In our case, we use an auto-encoder as a feature extractor, label predictor for supervised learning, and implemented a domain classifier in the training process together with reconstruction loss. In order to generate the domain classification, we treat each participant as an independent domain, with the number of domains equal to the number of participants. The corresponding domain classification loss is given below:

$$\mathcal{L}_{domain} = \|d - \hat{d}\|_2^2 \tag{3}$$

where d is the one-hot encoding of the source domain and $\hat{d}$ is the one-hot encoding computation of the target domain prediction output. As a result, the overall objective of unsupervised learning is:

$$\mathcal{L} = \mathcal{L}_{ae} + \lambda \cdot \mathcal{L}_{domain} \tag{4}$$

where λ is the coefficient of domain classification loss and set to -1.

2.2 Fine-tuning with triplet loss

The supervised learning models for downstream tasks are built on top of the learned representations. We design a fully connected dense layer as a classifier that follows the aforementioned CNN-based encoder that is pre-trained with the auto-encoder. Then the supervised learning structure is fine-tuned according to different downstream tasks. Furthermore, we also apply the triplet loss to optimize the representation of the training labels and avoid the trivial solutions learned in the pre-training procedure [21]. The triplet loss is given below:

$$\mathcal{L}_{triplet} = \max(d(a, p) - d(a, n) + margin, 0) \tag{5}$$

where *a* represents anchor sample data, *p* represents 'positive' sample data with the same label from the anchor, *n* represents 'negative' sample data with the opposite label against the anchor and the margin is a positive scalar. From the equation 5, we can see that the objective of the triplet loss is to decrease the distance between samples with the same labels and separate the distance between samples with different labels. During the supervised learning phase, the coefficient of triplet loss is set to 0.2.

3 Results and Discussion

We tested the proposed PiRL frameworks on two datasets, including CLAS and Apnea-ECG datasets for applications in mental health and physical health. Appendix B included detailed information on two datasets and experimental settings such as hyper-parameters. For each supervised prediction model, we pre-trained and fine-tuned the encoder at the beginning of each epoch. We compared the prediction accuracies of the PiRL models against the ones of the baseline (only an auto-encoder and a supervised learning model without any additional constraints) and person-specific models.

3.1 Stress Detection using Electrodermal Activity (EDA) with CLAS Dataset

Table 1 shows the prediction accuracy in all types of supervised learning models. The domain classification loss-based model did not show a significant increase of accuracy. The MMD loss-based model showed a higher accuracy of 66.5% compared to the baseline results (64.3 %). The prediction accuracy of the triplet loss only and the MMD + triplet loss models both exceeded 70%. The results illustrated that the model with triplet loss works better in stress prediction than the one with MMD loss-based models. To examine the statistical significance of the accuracy, we conducted an ANOVA (post-hoc: Tukey) test, and the corresponding results are also shown in Table 1. The prediction accuracy of the baseline framework was treated as the reference group. We found that both MMD and triplet loss based models have statistically higher accuracy than the baseline framework, but the triplet loss improved the model performance most obviously in stress detection. As expected, the person-specific models showed the highest accuracy (86.8%) but also highest standard deviations(0.189).

Table 1: Performance in stress detection on CLAS dataset. P-values are calculated by ANOVA (post-hoc: Tukey)

	Baseline	DANN	MMD	Triplet	MMD+Triplet	Person-Specific
Accuracy	64.3%	64.5%	66.5%	70.1%	70.6%	86.8%
SD	0.012	0.010	0.014	0.011	0.010	0.189
P-value < 0.01	-	×	✓	✓	✓	✓

3.2 Sleep Apnea Detection using ECG with Apnea-ECG Dataset

Table 2 shows the prediction accuracy of apnea detection using supervised learning models. The baseline model obtained a prediction accuracy of 75.2% with a standard deviation of 0.014. The accuracy of the domain classification loss-based model showed a slight numerical increase but not statistically significant difference. The accuracy of the remaining three PiRL frameworks reached over 79% and they were all statistically significantly higher than the baseline results with less standard deviation. The best framework for detecting apnea was the combination of MMD and triplet loss since it achieved the highest prediction accuracy. The prediction accuracy of the person-specific model was highest which exceed 95% and statistically higher than the baseline.

Table 2: Performance in sleep apnea detection on Apena-ECG dataset. P-values are calculated by ANOVA (post-hoc: Tukey)

	Baseline	DANN	MMD	Triplet	MMD+Triplet	Person-Specific
Accuracy	75.2%	75.7%	79.5%	79.1%	79.9%	95.7%
SD	0.014	0.013	0.010	0.013	0.009	0.011
P-value < 0.01	-	×	✓	✓	✓	✓

4 Latent Space Visualization

Visualization of the latent space is often helpful in demonstrating how the representations perform in a reduced-dimensional space. Hence, we used t-distributed stochastic neighbor embedding (t-SNE) to realize the latent space visualization in this study as it is commonly used for dimension reduction and visualization of high-dimensional datasets [23]. We plotted the distributions of learned representations

between the baseline and the MMD loss-based models in 2-dimensional coordinates to evaluate the effectiveness of the MMD loss function.

Figure 2 visualizes the latent space of invariant representations to further explore the influence of MMD loss on CLAS dataset. The left baseline figure is the distribution of representations extracted from the auto-encoder without any constrains. Similarly, the right figure is the distribution of representations with MMD loss. The range of the second dimensional components on the y axis showed a decrease from [-30,30] (baseline) to [2,4] (MMD) between participants, whereas the first component on the x axis ranged in [-20,20] in both plots. In general, as the representations of each participant spread out in distinct clusters, the latent space visualization of the baseline successfully revealed the heterogeneity across populations. However, for the participant-invariant representation learning, the data from different participants should be entangled and show no collapse in the feature space, but currently, representations on the right plot are still separable during the visualization [1]. Therefore, we still need to try some other approaches to better implement the latent space visualization.

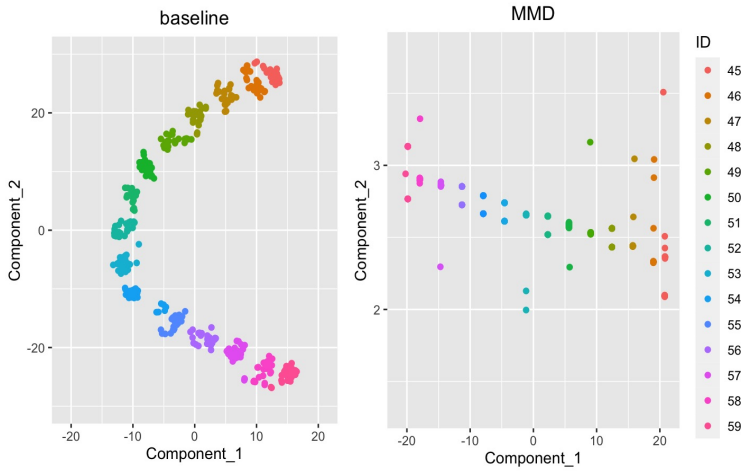

Figure 2: Latent Space Visualization Under MMD

5 Conclusions and Future Work

In this work, we proposed PiRL, which utilizes MMD and triplet loss for learning participant-invariant representations, to improve the performance of generic health detection models. We evaluated the performance and effectiveness of our framework using two public datasets for mental health and physical health. As preliminary results, we demonstrated that our proposed PiRL outperformed the baseline models and helped generic models achieve better performances. Performance improvement was not observed using DANN technique. The limitation of MMD loss is that it can only shorten the distribution distance in certain dimensions. In future work, we will investigate other approaches to further optimize the representations in data-driven health applications.

Acknowledgments

This work is supported by NSF #2047296 and #1840167.

References

[1] Aditya Kumar Akash, Vishnu Suresh Lokhande, Sathya N Ravi, and Vikas Singh. Learning invariant representations using inverse contrastive loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6582–6591, 2021.

[2] David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Ralf Mikut, Albrecht Schmidt, and Ozan Özdenizci. Domain-invariant representation learning from eeg

with private encoders. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1236–1240. IEEE, 2022.

[3] Majdi Bsoul, Hlaing Minn, and Lakshman Tamil. Apnea medassist: real-time sleep apnea monitor using single-lead ecg. *IEEE transactions on information technology in biomedicine*, 15(3):416–427, 2010.

[4] Urtnasan Erdenebayar, Yoon Ji Kim, Jong-Uk Park, Eun Yeon Joo, and Kyoung-Joung Lee. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Computer methods and programs in biomedicine*, 180:105001, 2019.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[6] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. Continuous stress detection using a wrist device: in laboratory and real life. In *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, pages 1185–1193, 2016.

[7] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[8] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[9] Lydia Kogler, Veronika I Müller, Amy Chang, Simon B Eickhoff, Peter T Fox, Ruben C Gur, and Birgit Derntl. Psychosocial versus physiological stress—meta-analyses on deactivations and activations of the neural correlates of stress reactions. *Neuroimage*, 119:235–251, 2015.

[10] Saskia Koldijk, Mark A Neerincx, and Wessel Kraaij. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on affective computing*, 9(2):227–239, 2016.

[11] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

[12] Yoshiki Nakashima, Jonghwa Kim, Simon Flutura, Andreas Seiderer, and Elisabeth André. Stress recognition in daily work. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 23–33. Springer, 2015.

[13] Debarshi Nath, Mrigank Singh, Divyashikha Sethia, Diksha Kalra, and S Indu. A comparative study of subject-dependent and subject-independent strategies for eeg-based emotion recognition using lstm network. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, pages 142–147, 2020.

[14] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. The effect of person-specific biometrics in improving generic stress predictive models. *arXiv preprint arXiv:1910.01770*, 2019.

[15] Shu Lih Oh, Eddie YK Ng, Ru San Tan, and U Rajendra Acharya. Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats. *Computers in biology and medicine*, 102:278–287, 2018.

[16] Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

[17] K Radhika and V Ramana Murthy Oruganti. Transfer learning for subject-independent stress detection using physiological signals. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6. IEEE, 2020.

[18] K Radhika and V Ramana Murthy Oruganti. Deep multimodal fusion for subject-independent stress detection. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 105–109. IEEE, 2021.

[19] Khandakar M Rashid and Joseph Louis. Times-series data augmentation and deep learning for construction equipment activity recognition. *Advanced Engineering Informatics*, 42:100944, 2019.

[20] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[22] Gaetano Valenza, Luca Citi, Antonio Lanatá, Enzo Pasquale Scilingo, and Riccardo Barbieri. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Scientific reports*, 4(1):1–13, 2014.

[23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[24] Han Yu and Akane Sano. Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild. *arXiv preprint arXiv:2202.12935*, 2022.

[25] Alexandros Zenonos, Aftab Khan, Georgios Kalogridis, Stefanos Vatsikas, Tim Lewis, and Mahesh Sooriyabandara. Healthyoffice: Mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2016.

A Model Description

A.1 Unsupervised Learning Phase

For the CLAS dataset, the encoder consists of 9 CNN layers, whereas the decoder consists of 8 up-sampling layers. For the Apnea-ECG dataset, the encoder consists of 12 CNN layers, whereas the decoder consists of 11 up-sampling layers.

A.2 Supervised Learning Phase

On both datasets, the encoder structure is same with the structure from unsupervised learning shape. In addition, the dense layer on two datasets both consist of 2 fully connected layers.

B Experimental Setting

B.1 Dataset

For this study, we investigated two datasets, described as follows:

B.1.1 CLAS Dataset

The CLAS dataset consists of recordings of physiological signals such as ECG, EDA, and Plethysmography (PPG) [11]. EDA time series data were chosen from 62 participants which 58 participants are included in this study. Labels for arousal, valence, and stress were included in the data and assigned based on the stimuli tags for interactive tasks. We were interested in stress detection using EDA data.

Table 3 summarizes the EDA data used for the experiments. Training set contains 746 time series sample data with label 0 ('non-stressed') and 247 time series sample data with label 1 ('stressed') from overall 45 participants. Testing set contains 269 sample data in non-stressed label and 90 sample data in stressed label form overall 13 participants.

Table 3: EDA data separation

	#Participants	#Non-stressed samples	#Stressed samples
Training Set	45	746	247
Testing Set	13	269	90
Total	58	1015	337

B.1.2 Apnea-ECG Dataset

The recordings of Apnea-ECG dataset includes continuous ECG signals and sets of annotations for apnea (respiratory signals). The length of the recordings ranges from slightly less than 7 hours to about 10 hours each. The Apnea-ECG dataset consisted of 70 participants was divided into a training set of 35 records (a01 through a20, b01 through b05, and c01 through c10), and a testing set of 35 records [16].

Table 4 summarizes the ECG data separation of the study. The labels of the sample are binary, which 0 represents normal breathing and 1 represents disordered breathing.

Table 4: Apnea-ECG data separation

	#Participants	#Normal breathing samples	#Disordered breathing samples
Training Set	35	10496	6514
Testing Set	35	10685	6548
Total	70	21181	13062

Table 5: Person-specific models data separation

	#Samples in EDA dataset	#Samples in Apnea-ECG dataset
Training Set (Tr1)	695	11907
Testing Set (Te1)	298	5103
Total	993	17010

B.1.3 Data Normalization

We used min-max normalization on the time series data for the CLAS and Apnea-ECG datasets to ensure that the normalized data had a similar scale. The formula of min-max normalization is given below:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$$

B.1.4 Person-specific Models

The performance of the person-specific models was calculated to compare against the baseline and the proposed PiRL frameworks. We used the original training set of CLAS dataset to obtain the performance of person-specific models. Specifically, for each participant, the original training set was divided into a new training set and testing set with a ratio of 70%:30%. The person-specific models were trained on the training sets and tested on the test sets to obtain the results. The final prediction accuracy of the person-specific models was recorded as the average accuracy of all participants.

B.2 Training Process

Our tasks are divided into two ways of evaluating representation learning. As for the representation learning, the objection function is mean squared error (MSE) loss between the reconstruction time series data and the original data. We need to obtain robust representations from the auto-encoder that is used in the following step. As for the supervised health condition detection, CNN-based encoder was pre-trained with the auto-encoder. Then the supervised learning structure is fine-tuned according to different downstream task

In the convolutional auto-encoder, we implemented 100 epochs per trail of training for both the baseline model and regularized model with MMD loss. We shuffled the data, set the batch size and learning rate to 32 and 0.001 at the start of each epoch in CLAS dataset, respectively. The batch size is 256 in Apnea-ECG dataset, which is the only difference between two datasets in training process. We specifically adjusted the weighted coefficient of the MMD loss to be 0.2 in the regularized model to prevent the dominance of the MMD loss. Additionally, the representation was set to be an 8×1 vector in order to guarantee that the representations we extracted are robust and contain sufficient personalized information.

In the supervised evaluation model, the parameters such as batch size and learning rate were maintained the same. These four supervised model types—baseline model, MMD loss only, triplet loss only, MMD and triplet loss were being taken into consideration. Noteworthy, the coefficient of the triplet loss was set to be 0.2 as well. Then we pre-trained the auto-encoder, initialized and saved all the representations and parameters in the encoder part. After that, we loaded the saved parameters to the supervised models and fine-tuned them. To get the predicted binary results of labels, we feeded the representations obtained from the encoder to the classifier, which consists of two fully connected layers. The final prediction accuracy was recorded as the mean of 10 trials of training.