ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414752

SELF-SUPERVISED LEARNING FOR SLEEP STAGE CLASSIFICATION WITH PREDICTIVE AND DISCRIMINATIVE CONTRASTIVE CODING

Qinfeng Xiao[⋆†], *Jing Wang*[⋆⋆†], *Jianan Ye*[⋆†], *Hongjun Zhang*[⋆†], *Yuyan Bu*[⋆], *Yiqiong Zhang*[⋆], *Hao Wu*[‡]

[⋆] School of Computer and Information Technology, Beijing Jiaotong University, China
[†] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
[‡] School of Science, Beijing Jiaotong University, China

ABSTRACT

The purpose of this paper is to learn efficient representations from raw electroencephalogram (EEG) signals for sleep stage classification via self-supervised learning (SSL). Although supervised methods have gained favorable performance, they heavily rely on manually labeled datasets. Recently, SSL arrives comparable performance with fully supervised methods despite limited labeled data by extracting high-level semantic representations. To alleviate the severe reliance of labels, we propose SleepDPC, a novel sleep stage classification algorithm based on SSL. By incorporating two dedicated predictive and discriminative learning principles, SleepDPC discovers underlying semantics from raw EEG signals in a more efficient manner. We thoroughly evaluate the performance of our proposed method on two publicly available datasets. The experimental results show that our method not only learns meaningful representations but also produces superior performance versus various competing methods despite limited access of labeled data.

Index Terms— Self-supervised Learning, Representation Learning, Electroencephalography, Sleep Stage Classification

1. INTRODUCTION

Sleep stage classification plays an essential role in sleep assessment and disease diagnosis. Typically, sleep experts identify sleep stages via electrical activity recording, which is called polysomnogram (PSG). A PSG includes an electroencephalogram (EEG), electrocardiogram (ECG) and other biomedical recordings. Specifically, multiple EEG leads were widely used to assess sleep quality. These signals were segmented into 30-s segments (called an epoch) and then be classified into different sleep stages by sleep experts according to sleep manuals such as Rechtschaffen and Kales (R&K) [1] and American Academy of Sleep Medicine (AASM) [2]. Although the manuals provide valuable rules for sleep stage classification, this task is still time-consuming and subjective. Therefore, an automatic sleep staging approach is required.

*Contact author.

To date, there have been many studies of sleep staging algorithm based on fully supervised models. Shallow models including SVM [3] and deep models including neural networks [4, 5] have made great progress on it. However, those supervised methods face exacting challenges. First, the labeling work is costly and laborious in terms of specialist experience and manual work in biomedical research. Although labeled data are scarce, unlabeled data are still abundant in most situations. Second, ground truth labels annotated by sleep experts can also be contradictory, which exerts a bad influence on label-relied learning tasks. Finally, the extracted representations by supervised models are not generalized and thus can not directly be applied to other biomedical applications. In other words, the data itself (million bytes) can provide much richer information than human annotations (a few bytes).

To tackle the above challenges, we propose a new framework *Sleep Discriminative and Predictive Coding* (SleepDPC), for self-supervised learning (SSL) based sleep stage classification. Specifically, we task SleepDPC to predict future representations of epochs recursively and to distinguish epochs from different epoch sequences. Both the two learning principles can be formulated as a multi-way classification problem that enforces higher similarities between matched representations while repels unmatched representations. The contributions of our work are three-fold: First, the proposed SleepDPC framework is a pioneer exploration to apply SSL on sleep stage classification. Second, the proposed two learning principles, Predictive Contrastive Coding (PCC) and Discriminative Contrastive Coding (DCC), enable SleepDPC to extract high-level semantics (such as underlying rhythms and patterns) from raw EEG signals. Third, we thoroughly evaluate the learned representations on two publicly available datasets. Experimental results show that our method arrives competitive performance over baseline approaches with different magnitudes of labeled data.

2. RELATED WORK

Fully supervised methods for automatic sleep stage classification can be categorized into two categories: traditional

machine learning models and deep neural networks. The first class applies traditional classification models such as Support Vector Machines (SVM) on dedicated features extracted by experts [3, 6]. Recently, the second class, deep neural networks, is widely used in sleep stage classification [4, 5]. Those deep supervised models bring considerable relief from manually feature engineering and achieve higher accuracy. However, they still suffer from expensive labeling work. Thus, learning generalized semantic representations in an unsupervised manner is desired.

Self-supervised learning, a popular branch of unsupervised learning, has produced an impressive performance in many research fields such as computer vision, video processing and natural language processing [7, 8, 9, 10, 11, 12, 13], showing its great power of extracting high-level semantic features. However, only a few studies have discovered the power of SSL applying to biomedical signals. In [14], a convolutional neural network is trained to distinguish different transformations of ECG signals. The learned representations are further used to perform the emotion recognition task. [15] explored the performance of three SSL learning principles on two downstream tasks. However, the pretext tasks used in [15] are so crude and the performance is far not comparable with supervised methods.

3. PROPOSED METHOD

3.1. Learning Framework

SleepDPC aims at extracting efficient representations from raw EEG signals. To achieve this, SleepDPC learns representations by two principles. The first learning principle is predictive, which means to predict a slowly varying representation based on recent past observations [12, 11]. Following the Slow Feature Analysis [16], we assume that an appropriate representation should slowly evolves over time (e.g. nearby epochs are more likely belongs to a same stage). While another learning principle tries to discriminate samples (epochs) from different segments (a batch of continuous epochs).

Figure 1 demonstrates the framework of our proposed method. In detail, EEG signals from different patients are first partitioned into non-overlapping segments, with each segment $\mathbf{X} \in \mathbb{R}^{S\times C\times T}$ containing an equal number of epochs, where S, C and T represent the number of epochs, the number of EEG channels and the timesteps of an epoch respectively. Each epoch $\boldsymbol{x}_t \in \mathbb{R}^{1\times C\times T}$, the gray box in Figure.1, is compressed into a local representation $\boldsymbol{z}_t \in \mathbb{R}^{1\times F_Z}$ with a parameterized encoder $f_\theta(\cdot)$ (can be implemented by a convolutional neural network or other parameterized functions), i.e. $\boldsymbol{z}_t = f_\theta(\boldsymbol{x}_t)$, where F_Z is the dimension of a local representation.

Then an aggregator $g_\psi(\cdot)$ (can be implemented by any autoregressive model, e.g. GRU and LSTM) aggregates c consecutive local representations $\{\boldsymbol{z}_{t'}\}_{t'\leq t}$ to produce a context representation $\boldsymbol{c}_t \in \mathbb{R}^{1\times F_C}$ (the pink box in Figure.1), i.e. $\boldsymbol{c}_t = g_\psi(\boldsymbol{z}_t|\boldsymbol{z}_{t-c}, \cdots, \boldsymbol{z}_{t-1})$, where F_C is the dimension of a context representation.

To ensure that the learned local representations and the context representation capture strong semantics, one can infer future local representations from the context representation. Given recent past local representations $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t$ and the context representation $\boldsymbol{c}_t$, the next local representation $\hat{\boldsymbol{z}}_{t+1}$ is predicted by a predictor $\phi(\cdot)$ feeding the context representation $\boldsymbol{c}_t$, i.e. $\hat{\boldsymbol{z}}_{t+1} = \phi(\boldsymbol{c}_t)$:

After that, the predicted $\hat{\boldsymbol{z}}_{t+1}$ can be further passed into the aggregator $g_\psi(\cdot)$ to get $\boldsymbol{c}_{t+1}$ gathering information of $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_{t+1}$. By sequentially obtaining context representations, we can predict local representations $\{\hat{\boldsymbol{z}}_{t'}\}_{t'>t}$ sequentially.

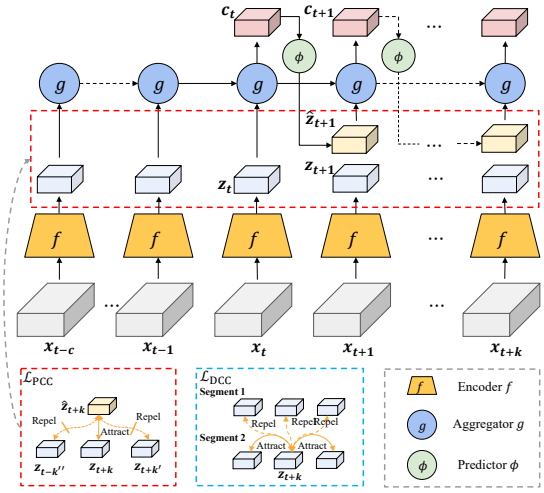

Fig. 1. The framework of our proposed model. The red dashed box depicts the first learning principle in SleepDPC: PCC $\mathcal{L}_{\text{PCC}}$, which is explained in Section 3.2.1. The blue dashed box describes another learning principle: DCC $\mathcal{L}_{\text{DCC}}$ (more details can be found in Section 3.2.2).

3.2. Objective Function

We build the learning principles of SleepDPC based on contrastive learning, which forces the similarity scores of positive pairs to be higher than corresponding negative pairs. The idea of contrastive learning paradigm is first introduced in [17], and further applied in unsupervised representation learning [7, 11]. The learning objectives of SleepDPC are two-fold: PCC and DCC.

3.2.1. Predictive Contrastive Coding

PCC evaluates the learned representations by "how well the representation can be distinguished from a related positive representation and unrelated negative representations". Considering each segment in a mini-batch, we predict p future lo-

cal representations based on c past observations where $p+c=S$. For each predicted representation $\hat{\boldsymbol{z}}_t$, the only related positive representation is $\boldsymbol{z}_t$. Any other representations $\{\boldsymbol{z}_j\}_{j\neq t}$ in the mini-batch are considered as "unrelated". The probability of distinguishability of $\boldsymbol{z}_t$ is defined as the similarity of $\boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_t$ dividing a normalization factor:

$$p(\boldsymbol{z}_t|\hat{\boldsymbol{z}}_t) = \frac{\exp(\hat{\boldsymbol{z}}_t^\top \cdot \boldsymbol{z}_t)}{\exp(\hat{\boldsymbol{z}}_t^\top \cdot \boldsymbol{z}_t) + \sum_j \exp(\hat{\boldsymbol{z}}_j^\top \cdot \boldsymbol{z}_j)}, \quad (1)$$

where the similarity of $\boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_t$ is a simple inner product of $\boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_t$. Summing logarithmic probabilities over the whole sequence arrives the PCC Loss $\mathcal{L}_{\text{PCC}}$:

$$\mathcal{L}_{\text{PCC}} = -\sum_t \log p(\boldsymbol{z}_t|\hat{\boldsymbol{z}}_t) \quad (2)$$

$$= -\sum_t \left[\log \frac{\exp(\hat{\boldsymbol{z}}_t^\top \cdot \boldsymbol{z}_t)}{\exp(\hat{\boldsymbol{z}}_t^\top \cdot \boldsymbol{z}_t) + \sum_j \exp(\hat{\boldsymbol{z}}_j^\top \cdot \boldsymbol{z}_j)}\right], \quad (3)$$

where $\boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_t$ represent the matched representation pair, and $\boldsymbol{z}_j$'s are unmatched representations. Cause the relationship with mutual information, the loss function is also referred as InfoNCE Loss [11].

3.2.2. Discriminative Contrastive Coding

DCC further evaluates the learned representations by "how well the representations can help discriminating segments". Since epochs in different segments of a mini-batch are temporally distant, they correspond to different classes. Meanwhile, epochs in the same segment are temporally close, thus they belong to the same class. The learning principle is formulated as a pseudo classification task, which estimates the probability of $\boldsymbol{z}_t$ being classified as τ-th class by:

$$p(c_t = \tau|\boldsymbol{z}_t) = \frac{\sum_{k,c_k=\tau} \exp\left(\boldsymbol{z}_t^\top \cdot \boldsymbol{z}_k\right)}{\sum_i \sum_j \exp\left(\boldsymbol{z}_i^\top \cdot \boldsymbol{z}_j\right)}, \quad (4)$$

where c_t represents the pseudo class label of $\boldsymbol{z}_t$. The learning objective is to minimize the negative log-likelihood over the mini-batch:

$$\mathcal{L}_{\text{DCC}} = -\sum_t \log p(c_t = \tau|\boldsymbol{z}_t) \quad (5)$$

$$= -\sum_t \left[\log \frac{\sum_{k,c_k=\tau} \exp\left(\boldsymbol{z}_t^\top \cdot \boldsymbol{z}_k\right)}{\sum_i \sum_j \exp\left(\boldsymbol{z}_i^\top \cdot \boldsymbol{z}_j\right)}\right]. \quad (6)$$

Combining the two learning objectives, we get the final objective function:

$$\mathcal{L} = \mathcal{L}_{\text{PC}} + \lambda \mathcal{L}_{\text{DC}}, \quad (7)$$

where λ is the parameter adjusting the weight of $\mathcal{L}_{\text{DC}}$.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

We evaluate our method on two publicly available datasets, Sleep-EDF [18] and ISRUC [19], which are described in Table 1.

Table 1. Statistics of Sleep-EDF and ISRUC.

Sleep-EDF				**ISRUC**			
	# epochs				# epochs		
W	4828	# subjects	20	W	5325	# subjects	10
N1	1416			N1	2208		
N2	7821	# channels	2	N2	4726	# channels	6
N3	3004			N3	3310		
REM	3091	sampling rate	100Hz	REM	1949	sampling rate	200Hz
Total	20160			Total	17522		

Sleep-EDF contains the PSG recordings of 153 subjects offered by PhysioNet [20]. Due to the prohibitively expensive computation, we focused our experiment on selected 20 healthy subjects. Since the wake stage dominates sleep stages, we truncate the leading and the tail wake sleep epochs and preserve only 30 minutes for them respectively to alleviate the class-imbalance problem. ISRUC is a sleep research aimed corpus containing three subgroups. We focus our evaluation on subgroup-III since it contains only healthy subjects.

4.2. Training Strategy and Experimental Setting

The training procedure contains two steps: unsupervised pre-training and supervised fine-tuning. Given an **unlabeled dataset**, the encoder $f_\theta(\cdot)$, the aggregator $g_\psi(\cdot)$ and the predictor $\phi(\cdot)$ are pre-trained with the objective described in Section 3.2. Latter, in the fine-tuning step, the last layer of SleepDPC is substituted by an one-layer perceptron $h_\eta(\cdot)$ for classification. During this phase, we use a small portion (10%) of labeled samples and representations from the aggregator to feed the linear classification layer. Only $h_\eta(\cdot)$ is optimized during the fine-tuning stage while parameters of $f_\theta(\cdot)$ and $g_\psi(\cdot)$ are freezed. Since we focus on the effectiveness of representation from self-supervised learning, only a simple one-layer fully-connected $h_\eta(\cdot)$ network was carried out on the features learned from pre-training to produce the classification results.

To get reliable experimental results, we perform cross validation on our proposed and baseline models. Namely, each subject is recursively selected for evaluation, and the rest of subjects are selected for training. Table 2 illustrates the details of hyperparameter settings. No dedicated training tricks such as dynamic learning rate or early stopping are used.

Table 2. Hyperparameter settings.

Hyperparameter	SleepEDF	ISRUC
Segment Length	10	10
Prediction Steps	5	5
Feature Dimension	128	128
Batch Size	64	32
Learning Rate	0.001	0.001
Pre-training Epochs	50	50
Fine-tuning Epochs	10	10
Optimizer	Adam	Adam

4.3. Performance Versus Supervised Methods

We investigate the performance of SleepDPC versus supervised methods. SleepDPC is compared with three competing methods, including SVM [6], DeepSleepNet [4] and HHT-CNN [5]. Those methods are state-of-the-art and representative in different perspectives, i.e., SVM in machine learning models with handcraft features, DeepSleepNet in deep convolutional models and HHT-CNN in time-frequency utilized models. All the supervised baselines are trained with the **fully labeled** dataset while SleepDPC is fine-tuned with **limited (only 10%)** labels.

Table 3. The Accuracy and F1-macro performance of SleepDPC (**10% labels**) versus supervised baselines (100% labels).

Dataset	SleepEDF		ISRUC	
	Accuracy	F1-macro	Accuracy	F1-macro
SVM	0.767 ± 0.008	0.618 ± 0.004	0.744 ± 0.005	0.705 ± 0.003
HHT-CNN	0.709 ± 0.003	0.621 ± 0.003	0.676 ± 0.004	0.623 ± 0.003
DeepSleepNet	0.816 ± 0.002	0.772 ± 0.003	0.632 ± 0.027	0.603 ± 0.023
SleepDPC (ours)	0.701 ± 0.008	0.640 ± 0.015	0.536 ± 0.015	0.489 ± 0.018

The accuracy and F1-macro performance of SleepDPC and three competing methods are shown in Table 3. SleepDPC performs comparable results on the two datasets in the respective accuracy and F1-macro. In terms of the SleepEDF dataset, SleepDPC performs comparably well concerning supervised baselines, achieving 0.701 and 0.640 of accuracy and F1-macro respectively; in terms of ISRUC dataset, the performance gaps between SleepDPC and competing methods are relatively small, achieving 0.536 and 0.489 of accuracy and F1-macro respectively, despite ISRUC is more challenging than SleepEDF w.r.t. the patient characteristics, dataset quality and etc. The experimental results show that SleepDPC efficiently leverages the information universally contained in the data itself, resulting in high-quality semantic representations; while the competing methods rely on manual labeling or feature engineering, resulting in weak capability of extracting generalized representations and thus slight improvement w.r.t. SleepDPC. The experimental results prove that SSL is promising to arrive competitive performance with supervised methods in the biomedical area.

4.4. Effectiveness of Representations

In this section we focus on the effectiveness of representations learned by SleepDPC. We compare the quality of our learned representations with three baselines including Autoencoder, PCA and Random Weights. Autoencoder consists of an encoder that uses the same architecture with $f_\theta(\cdot)$ of SleepDPC and a decoder that inverts the operations of $f_\theta(\cdot)$. Random Weights is a simplified version of SleepDPC initialized with random weights. A two layer perceptron is used in the fine-tuning phase for Autoencoder and PCA.

Figure 2 shows the accuracy and F1-macro of different methods w.r.t. different percentages of labeled data. It is remarkable that SleepDPC achieves best performance compared with other representation methods on two datasets. In terms of SleepEDF, SleepDPC arrives 0.70 accuracy and 0.64 F1-macro (with 10% labels), outperforming all baselines with accuracy improvements of 79.9%, 55.6% and 75.4% respectively (with order of PCA, Autoencoder and Random Weights); similar results can also be observed on ISRUC, which achieves improvements of 139%, 43.7% and 75.2% in accuracy. Due to the two proposed learning principles, SleepDPC learns more efficient representations than baselines which fail to extract high-level semantics. Figure 2 also demonstrates that the performance grows extremely slow with a larger percentage of labels. The reason is that the bottleneck of the classification layer is reached where the representations are fixed.

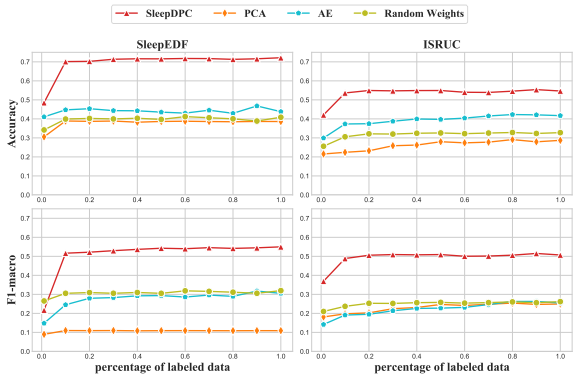

Fig. 2. The performance of SleepDPC and representation learning methods w.r.t. different percentage of labeled data.

5. CONCLUSION

In this paper, we have creatively presented the framework of SleepDPC for sleep stage classification by learning high-level semantics from raw EEG signals. Our approach arrives competitive performance over supervised methods with limited access to labeled data on two public datasets. Future work could focus on incorporating frequency domain information to achieve better interpreted semantic representations.

6. REFERENCES

[1] Edward A Wolpert, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects.," *Archives of General Psychiatry*, vol. 20, no. 2, pp. 246–247, 1969.

[2] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al., "The aasm manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, vol. 176, pp. 2012, 2012.

[3] Emina Alickovic and Abdulhamit Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, 2018.

[4] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo, "Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, Nov 2017.

[5] Liangjie Wei, Youfang Lin, Jing Wang, and Yan Ma, "Time-frequency convolutional neural network for automatic sleep stage classification based on single-channel eeg," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2017, pp. 88–95.

[6] B Koley and Debangshu Dey, "An ensemble system for automatic sleep stage classification using single channel eeg signal," *Computers in biology and medicine*, vol. 42, no. 12, pp. 1186–1195, 2012.

[7] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[11] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[12] Tengda Han, Weidi Xie, and Andrew Zisserman, "Video representation learning by dense predictive coding," in *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. 2019, pp. 1483–1492, IEEE.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Pritam Sarkar and Ali Etemad, "Self-supervised learning for ecg-based emotion recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. 2020, pp. 3217–3221, IEEE.

[15] Hubert J. Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," *CoRR*, vol. abs/2007.16104, 2020.

[16] Laurenz Wiskott and Terrence J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, 2002.

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. 2006, pp. 1735–1742, IEEE Computer Society.

[18] Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Josefien J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, 2000.

[19] Sirvan Khalighi, Teresa Sousa, José Moutinho dos Santos, and Urbano Nunes, "Isruc-sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, 2016.

[20] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.