frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 06 June 2023
DOI 10.3389/fnins.2023.1167723

OPEN ACCESS

EDITED BY
Fernando Vaquerizo-Villar,
University of Valladolid, Spain

REVIEWED BY
Matteo Cesari,
Innsbruck Medical University, Austria
Panfeng An,
Nanchang University, China

*CORRESPONDENCE
Zhihong Yang
✉ zhyang@implad.ac.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 16 February 2023
ACCEPTED 08 May 2023
PUBLISHED 06 June 2023

CITATION
You Y, Chang S, Yang Z and Sun Q (2023) PSNSleep: a self-supervised learning method for sleep staging based on Siamese networks with only positive sample pairs. *Front. Neurosci.* 17:1167723. doi: 10.3389/fnins.2023.1167723

PSNSleep: a self-supervised learning method for sleep staging based on Siamese networks with only positive sample pairs

Yuyang You[1†], Shuohua Chang[1†], Zhihong Yang[2*] and Qihang Sun[1]

[1]School of Automation, Beijing Institute of Technology, Beijing, China, [2]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Traditional supervised learning methods require large quantities of labeled data. However, labeling sleep data according to polysomnography by well-trained sleep experts is a very tedious job. In the present day, the development of self-supervised learning methods is making significant progress in many fields. It is also possible to apply some of these methods to sleep staging. This is to remove the dependency on labeled data at the stage of representation extraction. Nevertheless, they often rely too much on negative samples for sample selection and construction. Therefore, we propose PSNSleep, a novel self-supervised learning method for sleep staging based on Siamese networks. The crucial step to the success of our method is to select appropriate data augmentations (the time shift block) to construct the positive sample pair. PSNSleep achieves satisfactory results without relying on any negative samples. We evaluate PSNSleep on Sleep-EDF and ISRUC-Sleep and achieve accuracy of 80.0% and 74.4%. The source code is publicly available at https://github.com/arthurxl/PSNSleep.

KEYWORDS

sleep staging, self-supervised learning, Siamese networks, contrastive learning, positive sample pairs

1. Introduction

Human sleep data are used in a variety of medical diagnoses, health care, and other applications (Wulff et al., 2010), and are commonly collected using polysomnography (PSG). It consists primarily of an electroencephalogram (EEG), an electrooculogram (EOG), an electromyogram (EMG), and an electrocardiogram (ECG). However, they are difficult to recognize and must be annotated by sleep specialists who have been well trained in the field. PSG data are always segmented into epochs of 30 s for analysis. In addition, the sleep stages of each individual are classified by experts according to sleep manuals such as the Rechtschaffen and Kales (R&K; Wolpert, 1969) and the American Academy of Sleep Medicine (AASM; Iber et al., 2007). There are a number of traditional deep learning approaches (Phan et al., 2018; Supratak and Guo, 2020; Zhu et al., 2020; Guillot and Thorey, 2021), which require a large amount of labeled data for training. It is very challenging to apply these methods to sleep classification when labeling these recordings is much more challenging than labeling an image (Mai and Yu, 2021).

The self-supervised learning method has attracted a lot of attention in recent years. It can be used to extract effective representations from unlabeled data and achieve similar performance

to supervised learning with limited annotation information (Jing and Tian, 2019). Among various self-supervised learning methods, contrastive learning is favored by researchers because of its excellent performance (Lê Khắc et al., 2020). Oord et al. (2018) proposed Contrastive Predictive Coding (CPC). The negative samples were selected from the current batch and the entire model was trained using the loss function NCE (Gutmann and Hyvärinen, 2010) known as InfoNCE. As well as proving that self-supervised learning is universal in many different fields, they also proved that it has many advantages. As Wu et al. (2018) demonstrated, self-supervised learning can be achieved by maximizing the distinction between instances. They adopted a memory bank to store representations, which expanded the selection range of negative samples to the entire dataset. He et al. presented Momentum Contrast (MoCo; He et al., 2019), which was used for self-supervised visual representation learning. The updated strategy maximized consistency between negative samples and improved performance (Chen X. et al., 2020). However, the success of such methods depends greatly on the selection of negative samples in the training process (Jaiswal et al., 2020).

An alternative method of self-supervised learning is to learn invariant representations from different views of the original data (Zhou et al., 2021). As a means of achieving self-supervised learning, Siamese networks (Caron et al., 2020; Bardes et al., 2021) are used to maximize the similarity between the outputs (representations) of two branches of the Siamese networks. Chen T. et al. (2020) proposed SimCLR. It simplified contrastive self-supervised learning and did not rely on specific architectures or memory banks. A BYOL method has been proposed by Grill et al. (2020), and a SIMSIAM method has been proposed by Chen and He (2020). A prediction module is included in both of these methods, which introduces asymmetry into the original Siamese network. On the basis of previous research, Zbontar et al. proposed Barlow Twins (Zbontar et al., 2021), which incorporate redundancy reduction strategies. Assran et al. (2022) proposed Masked Siamese networks (MSN). It matched the representation of an image view containing randomly masked patches to the representation of the original unmasked image. A critical aspect of these methods is the composition of the data augmentations, which plays a crucial role in the results (Wang and Qi, 2021).

In recent years, some researchers have tried to apply self-supervised learning to sleep staging, hoping to free sleep experts from the tedious labeling work. SleepDPC was proposed by Xiao et al. (2021) and based on two dedicated learning principles, predictive and discriminative. It could discover underlying semantics from raw EEG signals. Cosleep is a representational learning framework that is based on a multi-view co-training mechanism that was proposed by Ye et al. (2022), along with a memory module that was added to the framework. Chang et al. (2022) proposed DSSNet, which combined the classical framework of DeepSleepNet (Supratak et al., 2017) and the classical self-supervised learning loss function InfoNCE. The TS-TCC was proposed by Eldele et al. (2021), used two different augmentations to get two views and adopted a contextual contrasting module to learn discriminative representations. The above methods have achieved good results. However, some of these studies rely too heavily on the selection of effective negative samples, and some of their network structures are overly complex.

In order to better apply self-supervised learning method to sleep staging, and free it from excessive dependence on negative samples, we propose PSNSleep. It can achieve better performance than other self-supervised methods by using Siamese networks and only a positive sample pair. The positive sample pair is constructed using a simple data augmentation method. Then, two CNNs and a GRU are used as branches of the Siamese network to extract general representations. A network's overall training goal is to maximize the similarity between pairs of positive samples in order to increase its performance. In addition, we introduce asymmetry into the Siamese network and adopt different update strategies for the parameters of the two branches.

In summary, this paper is mainly devoted to developing a novel self-supervised learning method based on Siamese networks for sleep staging. In the representation extraction part of the Siamese network, we adopted two CNNs and a GRU. This network structure is more suitable for sleep data and can extract multi-view representations. At the same time, we introduced an asymmetric structure of a prediction in one branch of the Siamese network to prevent the occurrence of collapse solutions. An augmentation strategy designed to eliminate dependence on negative samples and create positive pairs. We introduced mixup and time shift augmentations. The mixup learns foreground information by mixing different background information. The time shift views adjacent sleep epochs as positive pairs. We evaluated our framework on two public datasets. The results show that our method is effective for sleep staging. Additionally, we conducted ablation experiments to explore the effects of different data augmentation methods.

2. Materials and methods

We propose a new self-supervised learning method to extract general representations from single-channel EEG without expecting to learn differences between the current sample and negative samples. The method is called PSNSleep. Figure 1 illustrates the architecture of PSNSleep.

2.1. Data augmentation blocks

2.1.1. The time shift block

The sleep data are continuous, which means that sleep epochs are generally in the same stage for a period of time. A subject's sleep data throughout a night is visualized in Figure 2; we can see the continuity between two adjacent sleep epochs. And the probability of adjacent two sleep epochs happen to be in the transition stage of sleep, which means they belong to two sleep stages, is very small. As a result, we may ignore this situation and consider two sleep epochs to be in the same sleep stage as a whole. After adopting this augmentation method, we obtained encouraging experimental results, indirectly demonstrating that ignoring this situation does not have much impact on our experiments. We can choose the adjacent sleep epochs at the same stage as a positive sample pair since their waveforms are similar, which indicates that they have a high degree of similarity. For the sleep signal at epoch t in the overnight sleep data, the adjacent epoch $t+1$ can be considered as its positive sample (Mai and Yu, 2021). It should be noted that the continuity of sleep stages is an assumption. For people with diseases, e.g., sleep apnea, narcolepsy, it might not hold due to the sleep fragmentation they suffer from.

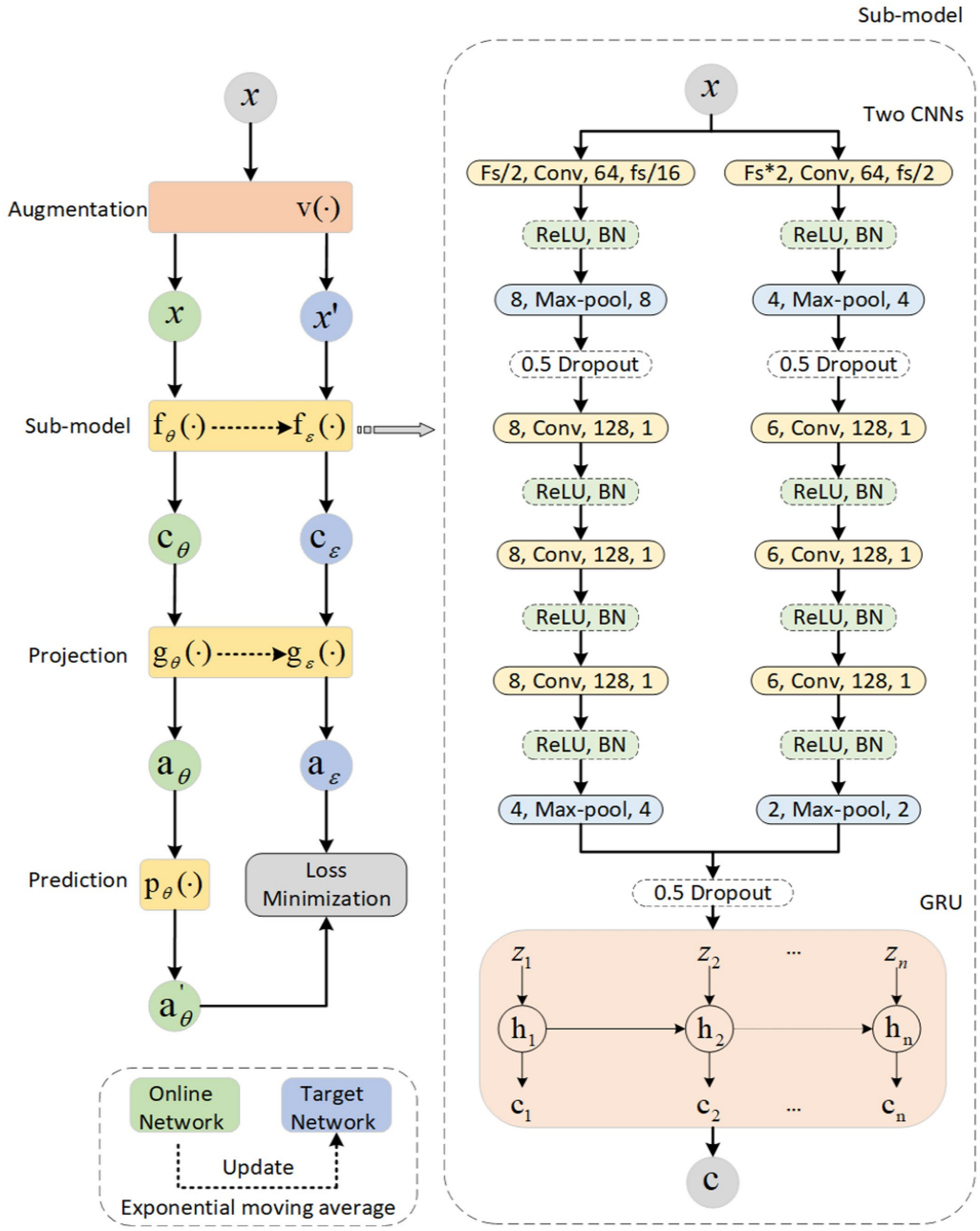

FIGURE 1
The overview architecture of our model. The left part of the figure is a Siamese network. The two branches are the online network and the target network. The parameters of the target network are dynamically updated according to the online network. The right part of the figure shows the structure of the sub-model, consisting of two CNNs and a GRU.

2.1.2. The mixup block and the Gaussian block

A batch of normalized EEG samples is used as input. In the mixup block, each sample is mixed with the other sample in a small ratio to obtain a mixed sample (Niizumi et al., 2021). It is also pertinent to mention that the other sample has been randomly selected from the current batch. The batch size is large enough to ensure the randomness that the sample selection is random. This operation can be viewed as changing the background information of a sample. In detail, the mixed sample adds background that is produced by the mixup block, while the raw sample has no background. The purpose of the above operations is to create a positive sample pair. They share most of the information, which we call foreground when referring to the area of audio recognition, but have different backgrounds. Therefore, in our method, the model can be improved by focusing on the foreground and ignoring different backgrounds in order to learn similar information between positive sample pairs. We actually require a representation of similar information extracted from the positive sample pair.

We adopt a basic mixup calculation method, and the calculation formula is,

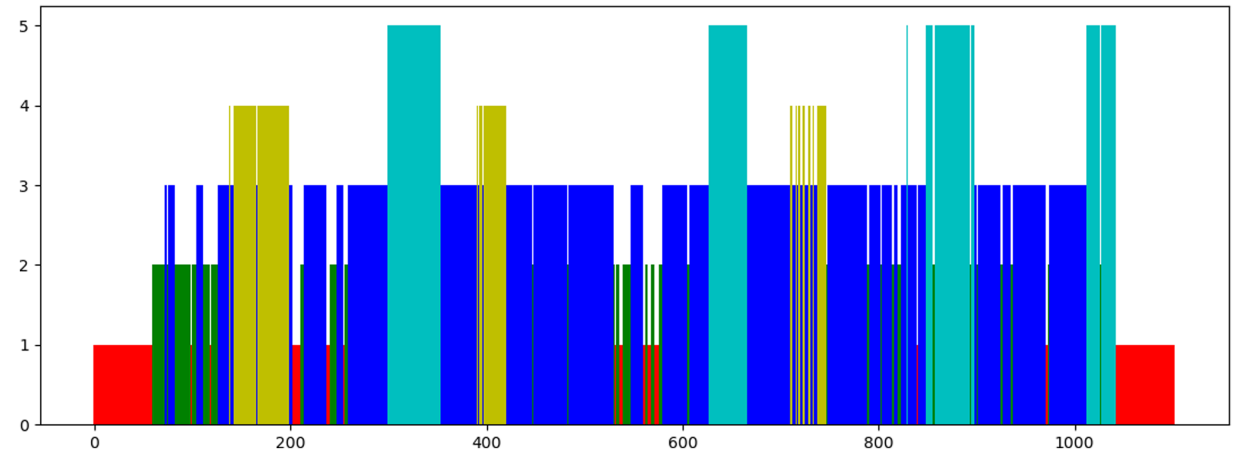

FIGURE 2
Sleep distribution map for a subject throughout the night. Stages of the same type have the same color and height. A *Y*-axis represents the classification of sleep, where 1 represents W, 2 represents N1, 3 represents N2, 4 represents N3, and 5 represents REM. During a period of time, the height and color are the same, indicating that they are in the same phase of sleep.

$$\tilde{x}_i = (1-\gamma)x_i + \gamma x_k \tag{1}$$

where x_i is the current sample, x_k is the other sample that is used to mix, γ is the mixing ratio. Furthermore, a higher mixing ratio implies a greater proportion of background from the other sample. $\tilde{x}_i$ is the mixed sample. x_k is randomly selected from the current batch.

For the Gaussian block, the principle is similar to the mixup block. The difference is that the mixed part is no longer other sleep samples, but rather the Gaussian noise we randomly generate. Considering that the signal about the Gaussian noise is generally small, it is directly superimposed on the original sleep data rather than mixed in a certain ratio. The current sample is superimposed with Gaussian noise generated at random, the calculation formula is,

$$\tilde{x}_i = x_i + g_i \tag{2}$$

where x_i is the current sample and g_i is the gaussian noise.

2.1.3. The random mask block

Assuming that the length of an EEG sample is N, we can say that the sample consists of N patches. In order to maintain global information, we randomly select patches to mask. We set the values of the randomly selected patches to 0 to ensure that the raw data and the augmentation have the same length. For the waveform of sleep data, masking some points randomly does not have a significant impact on the overall trend of the waveform. In other words, the masked waveform contains similar graphic information compared to the original waveform. They have high similarity and can be considered as a positive sample pair. In the training process, our network is used to predict masked patches using the powerful learning capability of neural networks, which facilitates learning off-the-shelf representations more effectively (He et al., 2021).

2.1.4. The scaling block

In addition to mixing with other signals to generate positive samples, we can also only rely on the sample itself to generate. We add random variations to the signal and scale up its magnitude. The specific realization is to scale the sleep data point by point by multiplying it with a random-number. The magnitude scaling can determine the similarity between positive sample pairs to a certain extent. The variation in amplitude can be expressed as follows,

$$\tilde{r}_i = \rho \times r_i \tag{3}$$

where r_i is the magnitude of the current signal, $\tilde{r}_i$ is the scaling magnitude, and ρ is the scaling rate.

2.2. Siamese network

The Siamese network consists of two branches that are called online network and target network. Both of them share a similar architecture. The branch of online network has three parts: a sub-model that is used as a representation extractor, a projection, and a prediction. The prediction is absent from the other branch of the target network, which makes it a bit different. This self-supervised learning method is named PSNSleep.

2.2.1. Representation extractor

The specific structure of the sub-model is shown in Figure 1. The model is composed of convolution layers and a recurrent layer. A convolution layer consists of two branches, one with a big-filter and the other with a small-filter, which allows time-frequency features to be extracted. In actuality, the recurrent layer is a gated recurrent unit (GRU), which is used to learn sequential epoch features. Taking the sub-model of the online network as an example, the process is as follows:

$$f_\theta(\cdot) = f_{(CNN,\theta_1)}(\cdot) + f_{(GRU,\theta_2)}(\cdot) \tag{4}$$

$$z_t = f_{(CNN,\theta_1)}(x_t) = \delta(w_s x_t + b_s) + \delta(w_l x_t + b_l) \tag{5}$$

$$c_t = f_{(GRU,\theta_2)}(z_t) = GRU(z_t, h) \tag{6}$$

Where $f_\theta(\cdot)$ represents the sub-model, $f_{(CNN,\theta_1)}(\cdot)$ and $f_{(GRU,\theta_2)}(\cdot)$ represent the CNNs and the GRU respectively, z_t is the time-frequency feature, and c_t is the final representation we need.

2.2.2. Projection and prediction

It is very helpful to have a projection and a prediction. In previous studies, the projection has been shown to improve performance. In addition, the asymmetry of the two branches of the Siamese network caused by the prediction of the future can help the whole model learn more information and avoid collapsed solutions. Both are composed of two fully-connected layers, and all of the layers have 960 units. The projection also includes two additional operations, batch normalization (BN) and activation using rectified linear units (ReLU). The operations of these two parts are as follows:

$$a_\theta = g_\theta(c_\theta), a_\varepsilon = g_\epsilon(c_\epsilon) \tag{7}$$

$$a'_\theta = p_\theta(a_\theta) \tag{8}$$

where a_θ. and a_ε are the output of representations of c through the projection. a'_θ is the output of the prediction, which is used to predict a_ε

2.2.3. Update strategy of the Siamese network parameters

Although the online network and target network have many similarities, their update strategies are completely different. The parameters θ of the online network are constantly updated during the training of the whole model. In order for the online network to be trained, the regression targets are provided by the target network. The parameters ε in this model are exponential moving averages of the θ. After each training step, we perform the following updates:

$$\theta = optimizer(\theta, \nabla L, \eta) \tag{9}$$

$$\epsilon = \tau\epsilon + (1-\tau)\theta \tag{10}$$

where ∇L is the gradient of loss function L, η is the learning rate of optimizer, and τ is the target decay rate.

2.2.4. Loss function

We duplicate a raw single-channel EEG (which has been normalized) into two copies. One does not require any processing, and the other is processed to get x' through a data augmentation module. We put x and x' into two branches of the Siamese network, online network and target network, to get the outputs a'_θ and a_ϵ respectively. In addition, L2-normalization is applied as well. To measure the similarity of the positive sample pair, we calculate the mean squared error between the normalized prediction and target projection.

$$\overline{a'_\theta} = \frac{a'_\theta}{\| a'_\theta \|_2}, \overline{a_\epsilon} = \frac{a_\epsilon}{\| a_\epsilon \|_2} \tag{11}$$

$$L_{\theta,\epsilon} = \| \overline{a'_\theta} - \overline{a_\epsilon} \|_2^2 = 2 - 2 \cdot \leq \frac{\langle a'_\theta, a_\epsilon \rangle}{\| a'_\theta \|_2 \cdot a_\epsilon \|_2} \tag{12}$$

We symmetricize the loss $L_{\theta,\epsilon}$ by separately feeding x' to the online network and x to the target network to compute $L'_{\theta,\epsilon}$. The loss function can be defined as follows:

$$L = \left(L_{\theta,\epsilon} + L'_{\theta,\epsilon}\right)/2 \tag{13}$$

During each training step, a stochastic optimization step is performed to minimize the loss L.

2.3. Experiments

We evaluate our self-supervised learning method by using single-channel EEG signals from two public datasets: Sleep-EDF (Goldberger et al., 2000; Kemp et al., 2000) and ISRUC-Sleep (Khalighi et al., 2016).

2.3.1. Sleep-EDF

It was an excellent dataset for the study on sleep staging in aging. The data were obtained in a 1987–1991 study of age effects on sleep in healthy Caucasians aged 25–101. The SC cohort of the Sleep-EDF contains 20 healthy subjects. Each PSG recording has two EEG signal channels, Fpz-Cz and Pz-Cz. All of them have the same sampling rate of 100 Hz. According to R&K standards, all recordings are categorized into eight categories (W, N1, N2, N3, N4, REM, MOVEMENT, and UNKNOWN). Data need to be preprocessed in accordance with the AASM standard. N3 and N4 are combined into N3, MOVEMENT, and UNKNOWN are the start and end of the recording, respectively. It is only the Fpz-Cz channel that we use.

2.3.2. ISRUC-sleep

As a test of the fit of our model to a generalized situation, we adopted this dataset to test the performance of the model. It contains 100 subjects. Each PSG recording has six EEG signals with the same 200 Hz sampling rate. All recordings are segmented into 30-s epochs and visually scored by two different sleep experts according to the guidelines of AASM, with the stages: W, N1, N2, N3, and REM. We only use the F3-A2 channel.

In Table 1, the numbers of 30-s EEG epochs for the five stages are presented. We employ zero-mean normalization to improve the speed of convergence of our model. The input data x is normalized to $\tilde{x} = \frac{x-\mu}{\sigma}$, where μ is the average and σ is the standard deviation.

The input size of Sleep-EDF is (batch-size, 1, 3,000) and ISRUC-Sleep is (batch-size, 1, 6,000). Other basic settings include a batch size of 32, a model trained for 200 epochs, and a random seed of 2022. The optimizer we chose is Adam with a learning rate of 0.0001. A decay rate of 0.1 is set for the target network. For the augmentation block, the mixing ratio γ is configured to 0.4. In order to mine the sequence information between extracted representations, we add a simple sequence network before classification, which is composed of a GRU with 64 units. The sequence length is 5. Then a linear classifier is used to evaluate the performance of representations. It consists of two fully connected layers with 960 and 64 units, respectively. We use the sample without any data augmentation as the input of the sub-model from the online network. In the process of evaluation, we adopt its output as a representation. All parameters of the sub-model are frozen when we train the classifier. The optimization is performed using an Adam optimizer with a learning rate of 0.0001. We have also set the number of training epochs to 200. The cross-entropy loss function is minimized by training the classifier.

We adopted 10-fold cross-validation. Datasets are divided into 10 parts. In each fold, nine parts are used in the training process to help update the parameters of the Siamese network and the left part is used to evaluate. A confusion matrix is adopted to clearly show the classification result of each class. We use overall accuracy (acc), macro-averaging F1-score ($MF1$), and Cohen's Kappa coefficient (κ) (Sokolova and Lapalme, 2009) to measure the performance of our network specifically. They can be calculated as follows:

$$acc = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{14}$$

$$MF1 = \frac{\sum_{c=1}^{C} F1_c}{C} \tag{15}$$

where TP_c presents the true positive samples when the class is c, $F1_c$ presents F1-score when the class is c, C presents the number of sleep stages, and N presents the total number of epochs.

TABLE 1 The numbers of 30-s EEG epochs.

Dataset	W	N1	N2	N3	REM	Total
Sleep-EDF	8,285	2,804	17,799	5,703	7,717	42,308
ISRUC-Sleep	20,098	11,062	27,511	17,251	11,265	87,187

TABLE 2 The results of previous methods and PSNSleep.

Method	Sleep-EDF			ISRUC-Sleep		
	acc	MF1	k	acc	MF1	k
DeepsleepNet	0.820	0.769	0.76	-	-	-
SleepDPC	0.701	0.640	-	0.536	0.489	-
Cosleep	0.716	0.558	-	0.579	0.501	-
DSSNet	0.800	0.700	-	0.714	0.663	-
PSNSleep	0.808	0.738	0.737	0.744	0.710	0.668

3. Results

Table 2 shows the results of the previous methods compared with our model, including one supervised learning method and three self-supervised learning methods. We have developed a model that has a similar structure to DeepSleepNet at the representation learning stage. Both SleepDPC and Cosleep contain multi-channel EEG signals, while we only use a single-channel EEG signal. Besides, SleepDPC only used the SC cohort portion of the dataset during the experiment. The previous model DSSNet used a single-channel EEG signal and all data, but it relied too heavily on the selection of negative samples.

For each fold, we use the last epoch's result as the current fold's result. Then, we can calculate the results of all folds to obtain the final performance metrics, as shown in Tables 3, 4. The accuracy of our method is 80.8% on Sleep-EDF and 74.4% on ISRUC-Sleep. It achieves the highest results among self-supervised methods. Our model improves the accuracy of Sleep-EDF and ISRUC-Sleep by 0.8% and 3%, respectively, compared to DSSNet. The accuracy distance between our self-supervised model and the classic supervised learning model DeepSleepNet is further reduced to 1.2%, on the dataset Sleep-EDF. In the validation experiment, as the experimental sleep data includes a certain proportion of aging PSG data, our model is capable of performing both non-aging and aging PSG data sleep staging tasks. Table 4 shows that our self-supervised learning method is applicable to sleep datasets and has achieved state-of-the-art performance. The gap between our method and the traditional supervised learning method is further shortened. PSNSleep representations have high generalization. At the same time, the results also prove that over dependence on negative samples is not necessary in self-supervised learning, and only using positive sample pairs can also achieve positive encouraging performance.

To explore how data augmentation methods affect representation performance, we conduct the ablation experiment

TABLE 3 Performance metrics on Sleep-EDF.

	Predicted					Per-class metrics		
	W	N1	N2	N3	REM	PR	RC	F1
W	6,912	445	136	28	213	80.23	89.37	84.55
N1	585	993	566	15	625	35.98	35.67	35.82
N2	603	500	15,192	598	777	87.15	85.98	86.56
N3	157	5	672	4,856	6	88.27	85.25	86.73
REM	358	817	867	4	5,670	77.77	73.48	75.56

TABLE 4 Performance metrics on ISRUC-Sleep.

	Predicted					Per-class metrics		
	W	N1	N2	N3	REM	PR	RC	F1
W	17,582	1,354	415	28	484	83.11	88.52	85.73
N1	2085	4,422	2,805	50	1,532	47.46	40.59	43.76
N2	869	1943	20,853	1802	1834	73.09	76.38	74.70
N3	74	40	3,194	13,697	190	87.79	79.66	83.53
REM	546	1,558	1,263	25	7,595	65.28	69.13	67.15

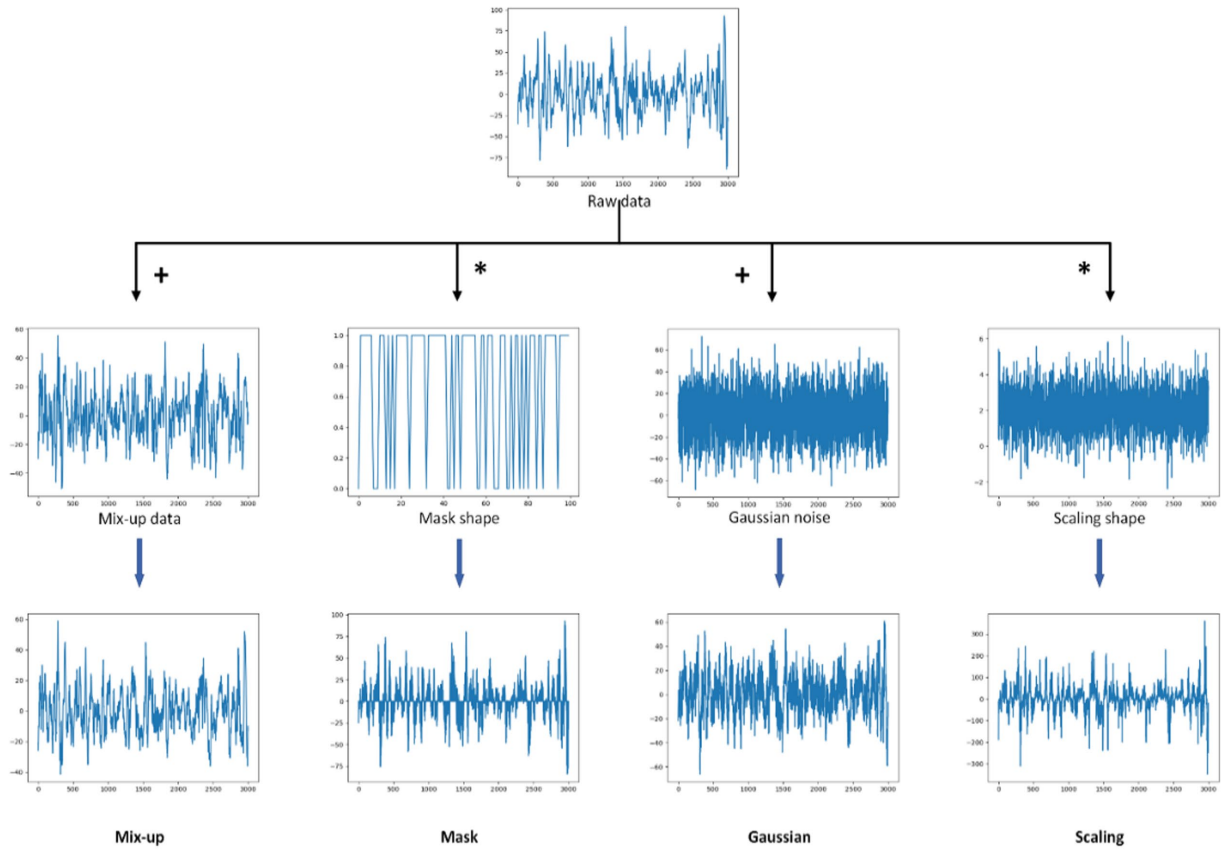

FIGURE 3
Diagram of different data augmentations. From left to right, the methods are mixup, mask, Gaussian, and scaling. The first row is the raw data, the second row is the augmentation data, and the last row is the result graphs. "+" represents the weighted sum of raw data and argumentation data in a certain proportion. "*" refers to the multiplication of elements one by one. A mask shape displays only the first 100 elements in order to facilitate viewing.

TABLE 5 The performance of different data augmentations.

Augmentation	Acc	MF1	κ
None	0.502	0.356	0.234
Gaussian	0.520	0.383	0.279
Scaling	0.535	0.388	0.320
mask1000	0.542	0.419	0.327
mask100	0.550	0.429	0.334
Mixup	0.802	0.710	0.728
Time shift	0.805	0.720	0.731

in this section. In order to maximize the reliability of the comparable results, all experimental settings except the augmentation blocks have been kept the same. As a direct reflection of the performance of different data augmentation methods, we ignore sequence information in order to reflect the performance of the representations. The representations extracted by the self-supervised method are directly fed into the classifier. Our laboratory uses Sleep-EDF to conduct this ablation experiment.

We tested six augmentation blocks: none, Gaussian, scaling, random mask, mixup, and time shift. Their specific operations are shown in Figure 3. The number of masked patches of the random mask is set to 100 (mask100) and 1,000 (mask1000) respectively. The results are shown in Table 5. The accuracy is 50.2%, 52.0%, 53.5%, 54.2%, 55.0%, 80.2%, and 80.5%, respectively. All data augmentation blocks have positive influences on the classification results compared with no augmentation. And the smaller number of mask patches may improve performance to a certain extent. For single augmentation, the time shift block and the mixup block have better results and their accuracy is over 80%.

To explore the sensitivity of our model to the mixing ratio, we varied it to 0, 0.2, 0.4, 0.6, 0.8, and 1.0, respectively. The results are shown in Figure 4. Our model achieves the highest performance when it is set to 0.4. It is also essential that the mixing ratio is chosen appropriately. Too large or too small a mixing ratio will lead to a decline in results.

There are still some limitations. (1) Through ablation experiments, we selected the time shift block as our data augmentation method. However, sleep data continuity in this method is an assumption. For people with diseases such as sleep apnea, narcolepsy, it might not hold due to the sleep fragmentation they suffer from. (2) For comparison with previous studies, we selected F3-A2 and Fpz-Pz channels when the EEG derivations recommended by the AASM are F4-M1, C4-M1, and O2-M1. (3) We selected two datasets to validate the feasibility of our proposed method. However, the effectiveness of joint training and testing on two or more datasets remains to be verified. (4) In addition,

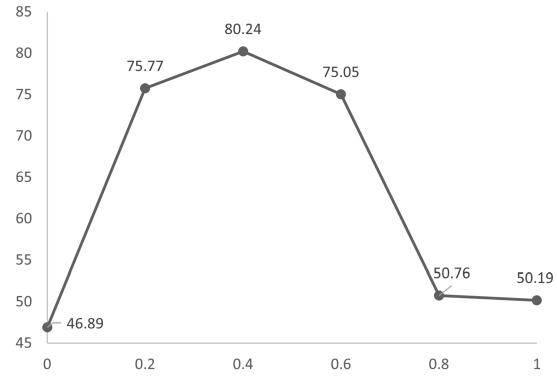

FIGURE 4
Accuracy at different mixing ratios is obtained by 10-fold cross-validation on the Sleep-EDF. The *Y*-axis represents accuracy and the *X*-axis represents the mixing ratio.

information about open-source datasets is limited. We cannot determine whether data leakage occurs during the experimental process.

In the subsequent research: (1) We will explore whether using EEG signals from different channels of the same dataset impacts experimental performance. (2) We will conduct transfer learning on sleep data, which means training on one dataset and testing on another dataset. (3) Cesari et al. (2021) developed an automatic method to model sleep as a continuous and dynamic process and this method predicted aging more accurately. In our future work, we can combine it with our method. Specifically, the self-supervised learning method replaces the manual feature extraction process, aiming to provide features with better generalization performance for the subsequent classification process.

4. Conclusion

In this paper, we propose a novel self-supervised learning method called PSNSleep. It extracts representations from unlabeled EEG signals and achieves the highest performance of self-supervised learning methods. It overcomes the disadvantage that the performance of self-supervised learning depends largely on negative samples in sleep staging. Our architecture consists of a Siamese network with two CNNs and a GRU for representation extraction. A projection component based on previous experience is also included in our model. The use of prediction also facilitates the introduction of asymmetry and improves the performance of our network. The positive pair is constructed from data augmentations, which are essentially different views of the same sample. Data argumentation is one of the keys to ensuring that the method we propose is successful as well. We explored a variety of data augmentation techniques during the course of these experiments. The results show that the time shift block achieves the highest performance. The objective of our model is for the representation of the positive sample pair, which is made up of two branches of the Siamese network, to be highly similar. Our experimental results show that sleep staging based on self-supervised learning can also achieve competitive results when using only positive sample pairs.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YY, SC, and ZY: conceptualization and methodology. YY, SC, and QS: software. YY: validation, data curation, supervision, and project administration. ZY: formal analysis and visualization. QS: investigation. YY and ZY: resources, writing—review and editing, and funding acquisition. SC: writing—original draft preparation. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (Nos. 81973744 and 81473579), CAMS Innovation Fund for Medical Science (CIFMS; Nos. 2022-I2M-1-018 and 2022-I2M-2-001), and the Beijing Natural Science Foundation (No. 7173267).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., et al. (2022). "Masked Siamese networks for label-efficient learning" in *European Conference on Computer Vision (ECCV)*.

Bardes, A., Ponce, J., and LeCun, Y. (2021). VICReg: variance-invariance-covariance regularization for self-supervised learning. ArXiv [Preprint]. doi: 10.48550/arXiv.2105.04906

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. ArXiv [Preprint]. doi: 10.48550/arXiv.2006.09882

Cesari, M., Stefani, A., Mitterling, T., Frauscher, B., Schönwald, S. V., and Högl, B. (2021). Sleep modelled as a continuous and dynamic process predicts healthy ageing better than traditional sleep scoring. *Sleep Med.* 77, 136–146. doi: 10.1016/j.sleep.2020.11.033

Chang, S., Yang, Z., You, Y., and Guo, X. (2022). DSSNet: a deep sequential sleep network for self-supervised representation learning based on Single-Channel EEG. *IEEE Signal Process. Lett.* 29, 2143–2147. doi: 10.1109/LSP.2022.3215086

Chen, X., Fan, H., Girshick, R.B., and He, K. (2020). Improved baselines with momentum contrastive learning. ArXiv [Preprint]. doi: 10.48550/arXiv.2003.04297

Chen, X., and He, K. (2020). "Exploring simple Siamese representation learning 2021" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15745–15753.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.E. (2020). A simple framework for contrastive learning of visual representations. ArXiv [Preprint]. doi: 10.48550/arXiv.2002.05709

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C., Li, X., et al. (2021). Time-series representation learning via temporal and contextual contrasting. ArXiv [Preprint]. doi: 10.48550/arXiv.2106.14112

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, E215–E220. doi: 10.1161/01.CIR.101.23.e215

Grill, J., Strub, F., Altch'e, F., Tallec, C., Richemond, P.H., Buchatskaya, E., et al. (2020). Bootstrap your own latent: a new approach to self-supervised learning. ArXiv [Preprint]. doi: 10.48550/arXiv.2006.07733

Guillot, A., and Thorey, V. (2021). RobustSleepNet: transfer learning for automated sleep staging at scale. *IEEE Trans. Neural Syst. Rehabil. Eng.* 29, 1441–1451. doi: 10.1109/TNSRE.2021.3098968

Gutmann, M.U., and Hyvärinen, A. (2010). "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models" in *International Conference on Artificial Intelligence and Statistics*.

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., and Girshick, R.B. (2021). "Masked autoencoders are scalable vision learners." in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R.B. (2019). "Momentum contrast for unsupervised visual representation learning." in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

Iber, C., Ancoli-Isreal, S., Chesson, A.L., and Quan, S.F. (2007). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Techinical Specifications*. Westchester, IL, USA: American Academy of Sleep Medicine.

Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. ArXiv [Preprint]. doi: 10.48550/arXiv.2011.00362

Jing, L., and Tian, Y. (2019). Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 4037–4058. doi: 10.1109/TPAMI.2020.2992393

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A., and Oberye, J. J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* 47, 1185–1194. doi: 10.1109/10.867928

Khalighi, S., Sousa, T., Santos, J. M., and Nunes, U. J. (2016). ISRUC-sleep: a comprehensive public dataset for sleep researchers. *Comput. Methods Prog. Biomed.* 124, 180–192. doi: 10.1016/j.cmpb.2015.10.013

Lê Khắc, P. H., Healy, G., and Smeaton, A. F. (2020). Contrastive representation learning: a framework and review. *IEEE Access* 8, 193907–193934. doi: 10.1109/ACCESS.2020.3031549

Mai, X., and Yu, T. (2021). "BootstrapNet: an contrastive learning model for sleep stage scoring based on raw Single-Channel electroencephalogram" in *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, 303–308.

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., and Kashino, K. (2021). "BYOL for audio: self-supervised learning for general-purpose audio representation" in *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Oord, A.V., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. ArXiv [Preprint]. doi: 10.48550/arXiv.1807.03748

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and de Vos, M. (2018). SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 400–410. doi: 10.1109/TNSRE.2019.2896659

Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 45, 427–437. doi: 10.1016/j.ipm.2009.03.002

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: a model for automatic sleep stage scoring based on raw Single-Channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1998–2008. doi: 10.1109/TNSRE.2017.2721116

Supratak, A., and Guo, Y. (2020). "TinySleepNet: an efficient deep learning model for sleep stage scoring based on raw single-channel EEG" in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 641–644.

Wang, X., and Qi, G. (2021). "Contrastive learning with stronger augmentations" in *IEEE transactions on pattern analysis and machine intelligence*.

Wolpert, E. A. (1969). A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Electroencephalogr. Clin. Neurophysiol.* 26:644. doi: 10.1016/0013-4694(69)90021-2

Wu, Z., Xiong, Y., Yu, S.X., and Lin, D. (2018). "Unsupervised feature learning via non-parametric instance discrimination" in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Wulff, K., Gatti, S., Wettstein, J. G., and Foster, R. G. (2010). Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nat. Rev. Neurosci.* 11, 589–599. doi: 10.1038/nrn2868

Xiao, Q., Wang, J., Ye, J., Zhang, H., Bu, Y., Zhang, Y., et al. (2021). "Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding." in *ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1290–1294.

Ye, J., Xiao, Q., Wang, J., Zhang, H., Deng, J., and Lin, Y. (2022). CoSleep: a multi-view representation learning framework for self-supervised learning of sleep stage classification. *IEEE Signal Process. Lett.* 29, 189–193. doi: 10.1109/LSP.2021.3130826

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). "Barlow twins: Self-supervised learning via redundancy reduction" in *International Conference on Machine Learning*.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A.L., et al. (2021). iBOT: image BERT pre-training with online tokenizer. ArXiv [Preprint]. doi: 10.48550/arXiv.2111.07832

Zhu, T., Luo, W., and Yu, F. (2020). Convolution- and attention-based neural network for automated sleep stage classification. *Int. J. Environ. Res. Public Health.* 17:4152. doi: 10.3390/ijerph17114152