NeuroNet: A Novel Hybrid Self-Supervised Learning Framework for Sleep Stage Classification Using Single-Channel EEG

Cheol-Hui Lee, Hakseung Kim, Hyun-jee Han, Min-Kyung Jung, Byung C. Yoon and Dong-Joo Kim

Abstract—The classification of sleep stages is a pivotal aspect of diagnosing sleep disorders and evaluating sleep quality. However, the conventional manual scoring process, conducted by clinicians, is time-consuming and prone to human bias. Recent advancements in deep learning have substantially propelled the automation of sleep stage classification. Nevertheless, challenges persist, including the need for large datasets with labels and the inherent biases in human-generated annotations. This paper introduces NeuroNet, a self-supervised learning (SSL) framework designed to effectively harness unlabeled single-channel sleep electroencephalogram (EEG) signals by integrating contrastive learning tasks and masked prediction tasks. NeuroNet demonstrates superior performance over existing SSL methodologies through extensive experimentation conducted across three polysomnography (PSG) datasets. Additionally, this study proposes a Mamba-based temporal context module to capture the relationships among diverse EEG epochs. Combining NeuroNet with the Mamba-based temporal context module has demonstrated the capability to achieve, or even surpass, the performance of the latest supervised learning methodologies, even with a limited amount of labeled data. This study is expected to establish a new benchmark in sleep stage classification, promising to guide future research and applications in the field of sleep analysis. The source code is available at https://github.com/dlcjfgmlnasa/NeuroNet

Index Terms—self-supervised learning, electroencephalogram (EEG), polysomnography, automatic sleep staging

1 Introduction

Sleep constitutes a fundamental determinant of human health and lifespan, serving as a cornerstone in alleviating both mental and physical stress encountered during routine activities, while also contributing to the maintenance of physiological homeostasis [1]. However, many individuals suffer from sleep disorders [2], and polysomnography (PSG) is commonly employed to assess sleep quality as a part of their treatment. PSG entails a measurement of various physiological signals during sleep, including electroencephalogram (EEG), electromyography, and electrocardiogram [3], [4] that require a laborious process of manually analyzing the data and classifying sleep stages by sleep experts.

Given these contexts, research into automatic sleep stage classification is advancing, with studies exclusively utilizing single channel EEG gaining particular attention due to their user convenience. The majority of these studies are based on supervised learning methodologies and have demonstrated acceptable performance enhancements through the utilization of the latest deep learning algorithms. Nonetheless, the implementation of such methodologies in real-world settings can pose several challenges. Firstly, the necessity for an extensive amount of labeled data may render its

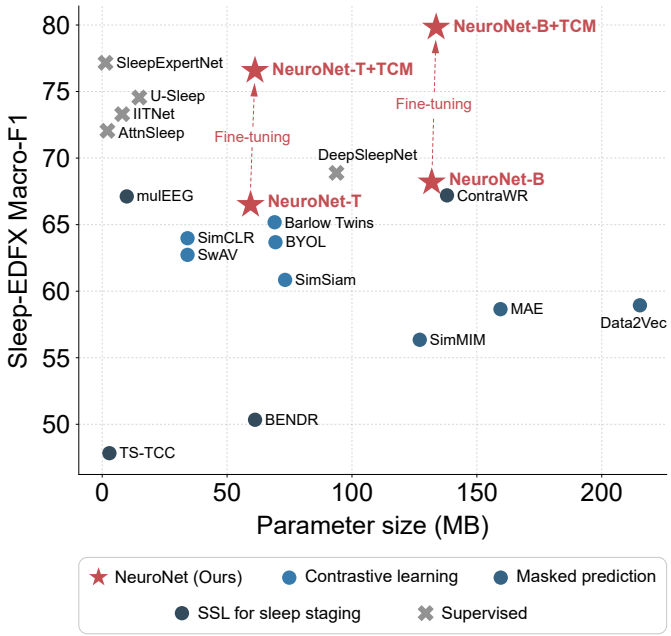

Fig. 1. Performance of Sleep-EDFX across various self-supervised learning and supervised learning.

- *Cheol-Hui Lee and Min-Kyung Jung are with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea; with Interdisciplinary Program in Precision Public Health, Korea University, Seoul, South Korea*
- *Hakseung Kim is with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea*
- *Hyun-jee Han is with the Department of Pharmacology, University of Cambridge, Cambridge, UK*
- *Byung C. Yoon is with the Department of Radiology, Stanford University School of Medicine, VA Palo Alto Health Care System, Palo Alto, CA, USA*
- *Dong-Joo Kim is with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea; with the Department of Neurology, Korea University College of Medicine, Seoul, South Korea; with Interdisciplinary Program in Precision Public Health, Korea University, Seoul, South Korea*
  E-mail: dongjookim@korea.ac.kr

preprint

acquisition impractical. Secondly, the reliability of labels is compromised by low inter-rater agreement rates observed in sleep stage evaluation [5]. Lastly, a model trained on data annotated by a single evaluator is prone to bias towards the opinions and interpretations of that evaluator, thereby posing a substantial risk to its generalizability.

Self-supervised learning (SSL) [6] has emerged as a promising methodology for extracting meaningful representations from unlabeled data. SSL comprises two principal paradigms: masked prediction tasks, which aim to learn the intrinsic feature information of the data, and contrastive learning, which trains models to distinguish between similar and dissimilar pairs of data points. Serving as a pre-training method, SSL trains models using pseudo-labels derived from the inherent features or similarities present in the data. Models pre-trained via SSL can subsequently be fine-tuned even using a small amount of labeled data. This approach not only enables the effective utilization of large volumes of unlabeled data but also mitigates the issues of reduced generalization resulting from inaccurate or biased labels.

The attributes of SSL render it suitable for applications in sleep EEG, where data acquisition is straightforward yet labeling proves challenging. Despite its potential, SSL research in sleep EEG remains relatively limited. A review of existing studies reveals that the majority of studies rely on contrastive learning tasks [7], [8], [9]. These studies have successfully learned representations of EEG, yet their effectiveness significantly depends on the backbone network's performance and on EEG data augmentation techniques [8], [9]. Although masked prediction tasks are not yet prevalent in sleep EEG research, their straightforward architecture, along with the capability to learn rich representations, has garnered popularity in other disciplines, such as computer vision. Nonetheless, they are perceived as less effective at learning discriminative representations, potentially leading to diminished performance in the downstream tasks [10], [11].

This study proposes a model, NeuroNet, a pioneering SSL framework that effectively integrates the capacity to discern inherent information within the dataset via masked prediction tasks with the discriminative representation capabilities afforded by contrastive learning tasks. Additionally, a deep learning methodology known as Mamba [12] is employed in this study, based on the selective state space model as an alternative to Long Short-Term Memory (LSTM) or multi-head attention, which are the conventional components of the temporal context module (TCM) utilized for integrating information across various time-zone EEG epochs.

To the best of our knowledge, the proposed methodology represents an innovative approach yet to be explored within the domain of EEG artificial intelligence research. Our findings underscore the superior performance of NeuroNet over existing SSL methodologies. Furthermore, this study unveils that upon fine-tuning, the model leveraging pretrained NeuroNet in conjunction with Mamba [12] outperforms the latest supervised learning technique trained on extensive labeled datasets even when utilizing only a limited amount of labeled datasets.

2 Related Works

2.1 Sleep Stage Classification

In recent years, research on deep learning-based sleep stage classification has been prolific. Notably, H. Phan et al. have proposed several models for sleep stage classification, as evidenced by several studies [13], [14], [15], which involve the conversion of EEG, electrooculography, and electromyography signals into time-frequency images for utilization as inputs to the models. Specifically, Multi-task CNN [13] employs a convolutional neural network (CNN) architecture incorporating two convolutional layers and two max-pooling layers. SeqSleepNet [14] adopts a Seq2Seq architecture, employing bidirectional long short-term memory (bi-LSTM). Similarly, XSleepNet [15] adheres to the Seq2Seq model structure similar to SeqSleepNet [14]; however, it distinctively integrates raw EEG signals as an additional input, setting it apart from Multi-task CNN [13] and SeqSleepNet [14]. XSleepNet [15] implements a strategic approach that dynamically adjusts the learning rate, increasing it during periods of effective generalization and decreasing it to prevent overfitting when necessary.

For the sake of user convenience, there are studies that opt to utilize solely single-channel EEG. Both DeepSleepNet [16] and IITNet [17] extract representation vectors from EEG signals using CNNs and learn temporal context information through bi-LSTM. While DeepSleepNet [16] focuses on temporal context between inter-epochs, IITNet [17] comprehensively considers temporal context within and between epochs. Recently, multi-head attention has emerged as a primary alternative to Recurrent Neural Network (RNN)-based models for capturing temporal dependencies swiftly and efficiently. Notable amongst those is AttnSleep [18] which is capable of extracting low-frequency and high-frequency features of EEG through multi-resolution CNNs, while adaptive feature recalibration enhances the quality of extracted features by modeling interdependencies between features. Subsequently, multi-head attention is employed to capture temporal dependencies among features. SleepExpertNet [19] extracts representation vectors of EEG via spectral-temporal CNN after signal extraction from disparate frequency bands. This is followed by the implementation of a model combining multi-head attention with bi-LSTM is utilized to learn long- and short-term temporal context information.

2.2 Self-Supervised Learning

SSL, a framework designed to extract highly semantic patterns directly from data without relying on labels, operates through a two-stage pipeline. In the initial stage, it learns generated pseudo-labels via arbitrarily defined tasks. Subsequently, in the second stage, supervised learning is conducted using data with a limited number of labels. For this reason, SSL can be characterized as an intermediate approach between unsupervised and supervised learning methodologies. Currently, SSL research primarily revolves around two principal paradigms: contrastive learning tasks and masked prediction tasks.

2.2.1 Contrastive Learning Task

The objective of the contrastive learning task is to elucidate the interrelationship among multiple samples, employing various methodologies (e.g., negative samples, self-distillation, clustering, and feature decorrelation, etc.). Negative sampling aims to minimize the distance between positive pairs in the latent space while increasing the distance between negative pairs [20], [21]. To achieve this, a large number of contrastive pairs are required, and methods like MOCO [20] and SimCLR [21] utilize memory banks and large batch sizes, respectively. Self-distillation trains online networks to predict the output values of target networks, thereby obviating reliance on negative samples [22], [23]. BYOL [22] and SimSiam [23], both based on self-distillation, present analogous structures. Nonetheless, BYOL [22] involves updating the encoder used in the target network with a momentum encoder during training. In contrast, SimSiam [23] does not use a momentum encoder but employs stop-gradient techniques instead. SwAV [24], a prominent example of clustering, trains representation vectors generated from identical samples to forecast the same prototype class. Feature decorrelation aims to learn decorrelated features [25], [26]. Barlow Twins [25] computes the cross-correlation matrix between outputs of identical networks and trains it to align as closely as feasible with an identity matrix.

2.2.2 Masked Prediction Task

The masked prediction task is a method where parts of an data sample are masked and then restored to learn representations. Examining SSL methodologies based on the masked prediction task, MAE [27] employs a ViT-based asymmetric autoencoder structure. Through the implementation of masking, MAE [27] selectively transforms a segment of the input image into a latent vector via the encoder and then trains the decoder to reconstruct the original image. To minimize spatial redundancy, a substantial proportion of masking (75% or higher) is applied, leading to notable enhancements in generalization performance. BEiT [28] initiates pre-training with a vector quantized-variational autoencoder, subsequently employing it as a tokenizer. It further segments the image into patches and leverages a subset of the masked patches as input for the ViT-based encoder, training it to predict the originally masked content based on the tokenizer. SimMIM [29] adopts a comparable structure to BEiT [28], yet it opts for pixel regression tasks over intricate methodologies like tokenization or clustering. Data2Vec [30] utilizes a teacher network trained on original data and a student network trained on masked data to create representation vectors. The student networks are trained to predict the representation vectors of the target network.

2.2.3 Hybrid Approach: Contrastive Learning Task + Masked Prediction Task

Recent research has initiated discussions on the combination of contrastive learning tasks and masked prediction tasks. CMAE [11] employs a masked autoencoder structure in the online branch and conducts the task of reconstructing the original image. The target branch employs a momentum-updated encoder and receives the entire image to conduct contrastive learning in tandem with the online branch. This model incorporates pixel shifting as a form of data augmentation. CAN [10] also combines contrastive learning and masked autoencoder structures similar to CMAE [11]. Noteworthy is its incorporation of noise prediction, which distinguishes it and enhances its capacity to acquire more refined representations.

2.3 Self-Supervised Learning for Sleep Staging

BENDR [31] integrates a CNN-based module for EEG signal learning alongside a Transformer-based module for learning temporal context between signals. In BENDR [31], output vectors extracted through the CNN and Transformer modules are defined as positive pairs if they are from the same time point, and as negative pairs if they originate from distinct time points. Subsequently, training of these pairs is conducted utilizing InfoNCE loss. ContraWR [32] opts for triplet loss over the conventional InfoNCE loss for contrastive learning. This strategy enables minimization of the distance between positive pairs as well as simultaneous increases in the gap between negative pairs. Negative samples for each sample are substituted with the average value, termed the world representation vector. CoSleep [33] adopts a multi-view SSL approach to concurrently learn signals and spectrogram images. It incorporates a module comprising a queue and a momentum encoder to secure a multitude of negative pairs, with the goal of augmenting representation performance. TS-TCC [8] focuses on temporal representation learning via a temporal contrasting module using weak and strong augmentations applied EEG signals. It maximizes the similarity between contexts originating from the same sample while minimizing the similarity between contexts of different samples. MAEEG [34] engages in representation learning for 6-channel sleep EEG using a masked autoencoder. Fine-tuned MAEEG presents enhanced performance in sleep stage classification even with limited labels provided. mulEEG [9] adopts the same augmentation approach as TS-TCC and has a multi-view SSL structure. It employs EEG signals and spectrograms transformed from EEG signals as input data, implementing diverse loss functions to effectively learn complementary information from multiple views.

3 METHODOLOGY

In this section, we introduce NeuroNet and the Mamba-based TCM. Figure 2 illustrates the detailed framework of NeuroNet. NeuroNet is an SSL framework designed to learn EEG signals, consisting of a total of five training stages. TCM is utilized to effectively capture time-series features or relational information between multiple EEG epochs, akin to a sleep expert. The TCM learns by decoding the intricate temporal patterns inherent within EEG signal, taking the output vector from the pretrained NeuroNet encoder as its input.

3.1 NeuroNet: Contrastive Masked Autoencoder for EEG

3.1.1 Data Preprocessing

Prior to training the model, the EEG signals were bandpass filtered between 1 and 50 Hz and then resampled at a

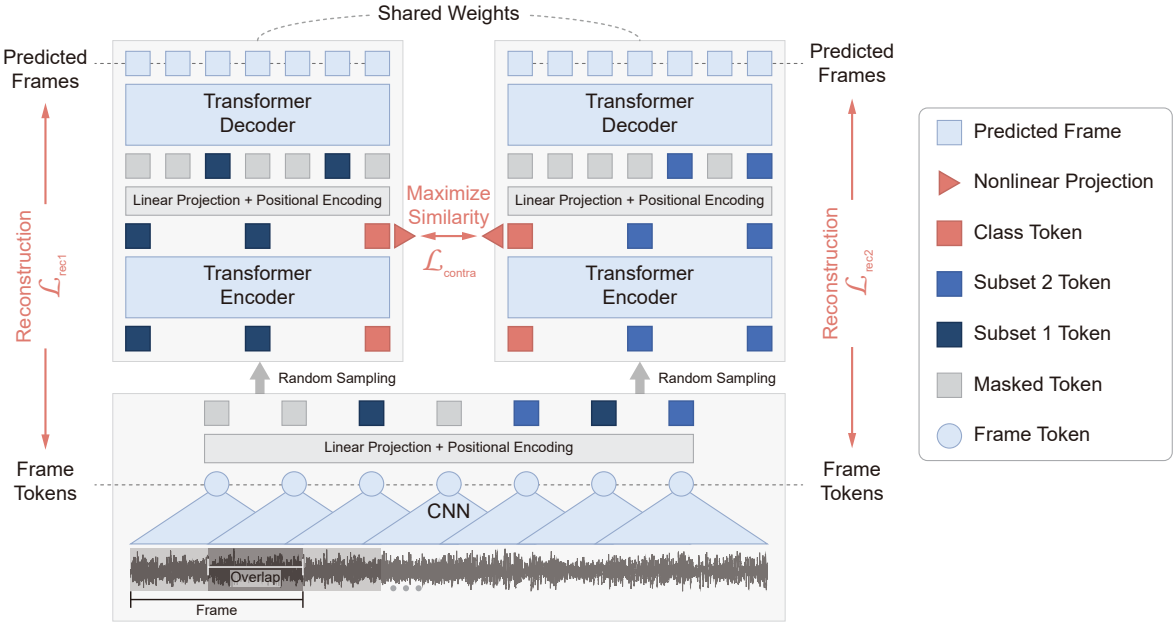

Fig. 2. Overview of the NeuroNet framework architecture.

frequency of 100 Hz. Following this preprocessing, a sliding window methodology was applied to each EEG epoch signal. The sliding window is a technique that involves moving a fixed-size frame at regular intervals across the data, facilitating the analysis which is effectively utilized in identifying specific patterns or events embedded within. The signals extracted through the sliding window served as the input data for the frame network.

3.1.2 Frame Network

The frame network is designed based on a multiscale 1d ResNet [35], specifically tailored for time series classification tasks. This model comprises an initial shared convolution and three parallel feature extractors. The shared convolution includes a 1d convolution layer, batch normalization, and a max pooling layer. Each feature extractor consists of three convolution blocks, with each block including a convolution layer, batch normalization, and an Exponential Linear Unit (ELU) activation function. Following the final block, residual connections and average pooling are applied. Notably, each feature extractor differs by employing 1d convolution layers with varying kernel sizes, which are 3, 5, and 7 respectively. The feature vectors extracted from distinct feature extractors are concatenated and subsequently processed through two fully connected layers to produce the final vector $\{{z_i}^m\}_{m=1}^{M}$, where M represents the total number of frames, and i is the frame index.

3.1.3 Masked Prediction Task

The structure of the MAE has been utilized for the masked prediction task. Figure 2 illustrates that the input data $\{{z_i}^m\}_{m=1}^{M}$ undergoes two separate masked prediction tasks independently. This architecture encompasses an encoder responsible for mapping the input to latent representations and a decoder that reconstructs the original signal from the latent representations. The encoder operates exclusively on the observed portion of the signal, without any mask tokens, while the decoder reconstructs the entire signal from the latent representation and mask tokens.

(Masking) Given the vector $\{{z_i}^m\}_{m=1}^{M}$ extracted through the frame network, a small subset of frames is randomly sampled for training, while the remainder are masked. This subset is denoted as $\{{z_i}^m\}_{m=1}^{\widetilde{M}}$, where $\widetilde{M}$ represents the frame numbers of the sampled subset. Generally, as the masking ratio increases, the amount of information available to the model decreases, which raises the difficulty of the reconstruction task. This enables models to understand the underlying patterns of the data and allows for a more generalized representation of the input data.

(Encoder) The encoder within NeuroNet is charged with the pivotal task of encoding the vector extracted via the frame layer into tokens and subsequently mapping them to the latent space. In NeuroNet, a standard Transformer stack serves as the encoder. The multi-head attention mechanism of the Transformer excels in efficiency capturing temporal information and correlations among frames. The encoder receives as input a vector obtained by concatenating the class token and $\{{z_i}^m\}_{m=1}^{\widetilde{M}}$, and then applying 'linear projection + positional encoding'. Through the Transformer stack, it produces a new vector $\{{h_i}^m\}_{m=1}^{\widetilde{M}+1}$. The class token assumes a crucial role in extracting semantic information from the data, thereby proving instrumental in contrastive learning tasks and predictions.

(Decoder) The decoder receives the vectors $\{{h_i}^m\}_{m=1}^{\widetilde{M}}$ extracted by the encoder, excluding the class token, and the masked vectors as inputs. Similar to the encoder, it applies linear projection and positional encoding to the input vectors. The masked vectors serve to represent vectors that were excluded during the masking phase, encapsulating the information omitted from the input data. In alignment with the asymmetric nature of the MAE, a tiny standard Transformer-based decoder is utilized. However, the decoder is not engaged after the SSL phase.

(Reconstruction Target) To predict the values of the masked vectors, the output vector generated by the decoder is passed through a fully connected layer to be reconstructed to the same size as the vector $\{{z_i}^m\}_{m=1}^M$extracted through the frame layer. Through this process, NeuroNet is able to derive the reconstruction vector $\{{r_i}^m\}_{m=1}^M$, and the loss is calculated using the mean square error, but only for the masked vectors. The formula for the loss function is as follows:

$$L_{rec} = \frac{1}{N\left(M-\widetilde{M}\right)}\sum_{i=1}^{N}\sum_{m=1}^{M-\widetilde{M}}({z_i}^m - {r_i}^m)^2 \quad (1)$$

Here, M and N represent the total number of data points and the batch size, respectively. As shown in Figure 2, since NeuroNet performs two independent masked prediction tasks using a single encoder and decoder, two separate losses (e.g., L_{rec1}, L_{rec2}) are ultimately derived.

3.1.4 Contrastive Learning Task

The objective of the contrastive learning task is to learn valuable representations by maximizing the similarity amongst identical instances while concurrently minimizing the similarity between disparate instances. In NeuroNet, the NT-Xent loss [36] is leveraged, which trains the model to converge positive pairs closer together and simultaneously drive negative pairs further apart within each mini-batch.

In Figure 2, random sampling results in two distinct subsets that are fed into the encoder. From this process, vectors h_i and h_j representing different views are derived through the class token. After mapping h_i and h_j through the projection layer to a new latent space, the normalized c_i, c_j are obtained. If the batch size is N, then c_i and c_j are each generated in sets of N. Consequently, each sample can generate 1 positive pair and $2\,(N-1)$ negative pairs. Thus, by iterating from $k=1$ to $k=2N$, and avoiding references to the same sample for both k and i, the formula is as follows:

$$L_{contra} = \frac{1}{2N}\sum_{k=1}^{N}\left[l\,(2k-1,\ 2k) + l\,(2k, 2k-1)\right] \quad (2)$$

$$l(i,j) = -log\frac{exp(\mathrm{sim}\,(c_i,\ c_j)\,/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} exp(\mathrm{sim}\,(c_i,\ c_k)\,/\tau)} \quad (3)$$

Here, $\tau > 0$ represents the temperature, a hyperparameter, with a default setting of $\tau = 0.5$. The term sim denotes cosine similarity.

3.1.5 The Combined Loss Function

NeuroNet aims to learn more superior representations by combining the masked prediction task, which learns the semantic information of EEG signals, with the contrastive learning task, which learns the relationships between EEG signals. The formula is as follows:

$$L_{total} = \frac{1}{2}\left(L_{rec1} + L_{rec2}\right) + \alpha L_{contra} \quad (4)$$

Here, α is the hyperparameter that balances the two types of loss mentioned above.

3.2 Mamba-based Temporal Context Modules

In the landscape of sequence modelling, previous models such as LSTM and multi-head attention have been situated within a trade-off between effectiveness and efficiency. Recently, Mamba [12], a structured state space sequence model featuring a selective mechanism and scan module, has emerged as a powerful tool in long sequence modeling. The selective mechanism employs input-dependent parameters, unlike space state models that utilize constant transition parameters. This adaptive strategy enhances overall generalization and performance by selectively focusing on relevant information while disregarding noisy or extraneous data. Complementing this, the scan module is deployed across each window of the input sequence, adeptly capturing intricate patterns and dependencies spanning multiple time steps. Additionally, a hardware-aware algorithm facilitates linear expansion of sequence length, thereby optimizing and improving computational efficiency and resource utilization.

According to the American Academy of Sleep Medicine (AASM) [37], when classifying sleep stages, not only the local features occurring in a single-epoch EEG (e.g., K-complex, sleep spindle, etc.) are considered, but also the relationships between adjacent EEG epochs are comprehensively considered to determine the stage of sleep. In this study, a Mamba-based TCM was developed to efficiently capture the temporal characteristics and correlations among multiple EEG epochs. Concretely, sleep stages are classified by a Mamba-based TCM that receives class tokens derived through an encoder from multiple EEG epochs. This process can be formula expressed as follows:

$$Sleep\ Stage = Mamba(\{Cls\ Token_n\}_{n=1}^{N}) \quad (5)$$

$$Cls\ Token = Encoder(EEG\ Epoch) \quad (6)$$

Here, N represents the number of EEG epochs input simultaneously, which is N is 20.

4 Experiments

4.1 Dataset Description

In this study, three PSG datasets were utilized.

4.1.1 Sleep-EDF Expanded Dataset

This study utilized the Sleep-EDF expanded dataset (Sleep-EDFX) [38], which is divided into two subsets: SC and ST. The SC subset contains PSG recordings from 153 healthy individuals aged 25 to 101, designed to explore the effects of age on sleep patterns. On the other hand, the ST subset comprises PSG recordings from 44 individuals aged 18 to 79, specifically curated to examine the impact of temazepam on sleep. The PSG configuration includes two bipolar EEG channels (Fpz-Cz and Pz-Oz), a horizontal EOG channel, and a submental chin EMG channel. Sleep stages were classified every 30 seconds by sleep experts into one of eight categories: {'W', 'N1', 'N2', 'N3', 'N4', 'REM', 'M', and '?'}. In this research, only the SC subset was utilized, selecting the Fpz-Cz EEG signal with a sampling rate of 100 Hz. Additionally, in alignment with the AASM standard, out of the 8 classes, N3 and N4 were merged into N3, and the classes 'M' and '?' were omitted.

TABLE 1
Experimental settings and dataset statistics.

Dataset	*No. of subjects*	*EEG Channel*	*Evaluation Scheme*	*Held-out Validation Set*	*Sampling Rate*	*Class Distribution*					
						Wake	*N1*	*N2*	*N3*	*REM*	*# Total*
Sleep-EDFX	153	Fpz-Cz	5-fold CV	15 subjects	100 Hz	66822 (34.78%)	21522 (11.20%)	69132 (35.99%)	8793 (4.58%)	25835 (13.45%)	192104
SHHS	329	C4-A1	5-fold CV	40 subjects	125 Hz	59129 (17.51%)	10304 (3.05%)	142125 (42.09%)	60153 (17.81%)	65953 (19.53%)	337664
ISRUC-Sleep	100	C4-A1	10-fold CV	10 subjects	200 Hz	22142 (24.55%)	9140 (10.13%)	30499 (33.82%)	16115 (17.87%)	12291 (13.63%)	90187

* CV: Cross Validation

4.1.2 Sleep Heart Health Study

The Sleep Heart Health Study (SHHS) [39], [40] is a comprehensive multi-center cohort study designed to investigate various cardiovascular and other outcomes associated with sleep-disordered breathing. It consists of two subsets: SHHS1, SHHS2. Each subset within the PSG includes two bipolar EEG channels (C4-A1, C3-A2), one EKG channel, two EOG channels, as well as two lower limb EMG channels, snoring detection, pulse oximeters, and a body position sensor. Sleep experts labeled the recordings every 30 seconds into one of eight categories: {'W', 'N1', 'N2', 'N3', 'N4', 'REM', 'Movement', and 'Unknown'}. In this study, 329 participants from the SHHS1 subset, considered to have regular sleep patterns (Apnea Hypopnea or AHI index less than 5) [41], were selected. The C4-A1 EEG signal with a sampling rate of 125 Hz was chosen. Following the AASM standard, the classes N3 and N4 were merged into N3, and 'Movement' and 'Unknown' were omitted.

4.1.3 ISRUC Sleep Dataset

The ISRUC-Sleep dataset [42] consists of 3 subsets and was collected to study both healthy subjects and those taking sleep medication. The first subset comprises data from 100 participants, each with only one PSG recorded. The second subset includes data from 8 participants, each with two PSG sessions. The third subset comprises data from 10 healthy participants, each with only one PSG recorded. This subset proves particularly useful for conducting comparative analyses between healthy participants and individuals afflicted with sleep disorders. In this dataset, two sleep experts labeled each 30-second interval with one of the five classes {'W', 'N1', 'N2', 'N3', 'REM'}. This study utilized the C4-A1 EEG signal from the first subset, sampled at a rate of 200 Hz, and employed the labeling provided by the first sleep expert.

4.2 Other State-of-the-Art Methodologies

Among various studies focusing on sleep stage classification and SSL methods, several with methodologies of significant influence (i.e., a high number of citations) and available published source code were selected and implemented. Such selection underscores the preference for methodologies with high reproducibility, which have been validated by numerous researchers in the field. The selected recent methodologies can be categorized into contrastive learning-based SSL, masked prediction task-based SSL, SSL for sleep staging, and supervised learning for sleep staging.

Firstly, for the contrastive learning-based SSL methodology, SimCLR [21], BYOL [22], SimSiam [23], SwAV [24], and Barlow Twins [25] were selected. The performance of this methodology shows variability depending on the types of data augmentation employed and the structure of the backbone network. Therefore, various experiments were conducted to identify the optimal approach for integrating EEG into contrastive learning-based SSL (detailed explanations are provided in Appendix C). For the masked prediction task-based SSL methodology, MAE [27], SimMIM [29], and Data2Vec [30] were chosen. In contrast to the former methodologies, these methodologies do not involve the data augmentation process, and the backbone network remains fixed as ViT, thus making additional experimental procedures unnecessary. Raw EEG signals were transformed into short-time Fourier transform images and used as input data. For SSL for sleep staging, BENDR [31], ContraWR [32], TS-TCC [8], and mulEEG [9] were selected, while for supervised learning for sleep staging, DeepSleepNet [16], IITNet [17], U-Sleep [43], AttnSleep [18], and SleepExpertNet [19] were chosen.

4.3 Experimental Setting

4.3.1 Evaluation Scheme

The performance of the model was assessed via subject group k-fold cross-validation, as outlined [8], [9], [33]. In detail, the construction of SSL-based methodologies involves dividing the dataset into three distinct groups: train, validation, and test datasets. The train dataset is utilized for SSL training and proceeds without labels. Subsequently, the validation dataset is utilized for linear evaluation and fine-tuning, leveraging a limited amount of labeled data. The test dataset is employed for comprehensive evaluation. In contrast, methodologies based on supervised learning are divided into train and test datasets, which are utilized for training and evaluation purposes.

Additionally, this study employed three evaluation scenarios to compare the proposed model with another approach. For evaluation, it is necessary to attach a classifier network to the backbone network trained via SSL and then proceed with training using a few labeled data. This process is referred to as a downstream task. A detailed description of each evaluation protocol is as follows:

- **(Evaluation Scenario 1, linear evaluation using single-epoch EEG)** The parameters of the backbone network are fixed, and then only the classifier network is trained. This method enables the evaluation

of which SSL methodologies can effectively represent EEG features.

- **(Evaluation Scenario 2, fine-tuning using multi-epoch EEG)** The performance of the final model, which integrates the backbone network with TCM, referred to as NeuroNet+TCM, is evaluated. In this configuration, the backbone network is fixed, except for the last Transformer layer. This method not only facilitates additional training of the data's nonlinear features but also ensures enhanced performance due to the use of multi-epoch EEG. This evaluation scenario is utilized for comparison between the NeuroNet+TCM and supervised learning models.
- **(Evaluation Scenario 3, cross-dataset evaluation)** The performance of the proposed models (NeuroNet and NeuroNet+TCM) is evaluated using datasets that are different from the ones utilized for training. The aim is to determine whether these proposed models achieve outcomes comparable to or surpassing those achieved by supervised learning. For this, z-normalization was applied to the input signals to align the distributions of datasets. Additionally, models trained on each fold were combined using a soft-voting ensemble. This process was similarly applied to supervised learning models.

The impact of trainable parameter size on performance was examined using two models that share identical architecture yet have varying numbers of parameters. These models are designated as NeuroNet-B and NeuroNet-T, with NeuroNet-B possessing a greater number of model parameters than NeuroNet-T. Appendix A lists the hyperparameter values used in each evaluation scenario. These hyperparameter values were derived based on the results of ablation experiments.

4.3.2 Evaluation Metric

For overall performance measurement, overall accuracy (ACC) and macro-F1 score (MF1) were utilized, while per-class F1 score (F1) was used for measuring performance by class. Here, MF1 is a useful metric for evaluating model performance on imbalanced datasets. ACC and MF1 can be calculated if true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class are provided. The formulas are as follows:

$$ACC = \frac{\sum_{i=1}^{K} TP_i}{M} \tag{7}$$

$$MF1 = \frac{1}{K}\sum_{i=1}^{K}\frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{8}$$

Here, for each class, $Precision_i$ and $Recall_i$ are calculated as $Precision_i = TP_i/(TP_i + FP_i)$ and $Recall_i = TP_i/(TP_i + FN_i)$, respectively, where M is the total number of samples, and K is the number of classes. In this context, K represents the 5 stages of sleep (Wake, N1, N2, N3, and REM).

5 Results

5.1 Comparison with State-of-the-Art Methodologies

5.1.1 Evaluation Scenario 1: Linear Evaluation using Single-epoch EEG

NeuroNet-B demonstrated superior performance across three PSG datasets compared to other SSL methodologies (Table 2) in learning representations of EEG signals. Particularly for the SHHS, both class-specific and overall performances were outstanding. For the ISRUC-Sleep, it exhibited exceptional performance across all metrics except for the REM class. Even though the performance of NeuroNet-T exhibited a slight decrease in comparison to NeuroNet-B, it nonetheless showcased a noteworthy degree of efficacy. In the Sleep-EDFX, NeuroNet-T's performance in the Wake class was notably superior, and its overall accuracy was the highest, excluding NeuroNet-B. For the ISRUC-Sleep, NeuroNet-T exhibited higher overall performance compared to other SSL methodologies.

5.1.2 Evaluation Scenario 2: Fine-tuning using Multi-epoch EEG

On closer examination, when compared to the latest supervised learning methodologies, NeuroNet-B+TCM shows the highest performance across all metrics for the Sleep-EDFX (Table 3). In the SHHS, it also displays the highest performance in all classes except for the Wake and REM. The notable encouragement arises from the fact that supervised learning models harness an extensive array of labeled datasets. For the ISRUC-Sleep, it was observed that NeuroNet-B+TCM falls short in overall performance compared to U-Sleep [43]. However, it still exhibits the highest performance in the N1 and N3 classes compared to other methodologies. Similar to evaluation scenario 1, NeuroNet-T shows lower performance than NeuroNet-B. Nonetheless, NeuroNet-T+TCM showcases performance on par with that of supervised learning methodologies. Specifically, in the N1 class, it shows the highest performance excluding NeuroNet-B+TCM. Additionally, in the SHHS, it is noted that NeuroNet-T+TCM, excluding NeuroNet-B+TCM, achieves the highest MF1.

5.1.3 Evaluation Scenario 3: Cross-Dataset Evaluation

Table 4 presents the results of comparing the proposed models with the two most outstanding models among supervised learning methodologies (i.e., U-Sleep [43], SleepExpertNet [19]). Upon closer inspection, it is observed that the proposed models demonstrate superior performance compared to existing supervised learning models. Despite utilizing only single-epoch EEG, NeuroNet exhibits better performance than supervised learning, and NeuroNet+TCM achieves remarkable performance by employing multiple EEG epochs and TCM. Notably, the results for models trained on SHHS (i.e., B$\rightarrow$ A, B $\rightarrow$ C) are outstanding, attributed to the larger dataset size of SHHS compared to other datasets. However, the performance of models trained on ISRUC-Sleep and evaluated on Sleep-EDFX (i.e., C $\rightarrow$ A) is superior in supervised learning methodologies. Nonetheless, excluding C $\rightarrow$ A, the proposed models outperform in all other aspects. It was shown that there are no noticeable performance gaps attributable to differences in model size.

TABLE 2
Comparison with other methodologies about linear evaluation using single-epoch EEG.

	Sleep-EDFX							*SHHS*							*ISRUC-Sleep*						
	Per-Class F1					*Overall*		*Per-Class F1*					*Overall*		*Per-Class F1*					*Overall*	
	W	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*	*W*	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*	*W*	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*
SimCLR	85.95	32.25	79.44	68.65	53.59	71.49	63.98	83.61	23.14	80.64	82.75	71.92	77.75	68.41	82.24	39.41	69.38	78.43	68.31	70.98	67.55
BYOL	87.08	33.40	78.32	64.49	55.12	71.82	63.68	83.20	19.50	81.71	84.52	71.32	78.51	68.05	82.30	40.62	70.42	77.48	67.76	71.22	67.72
SwAV	85.11	32.49	78.92	62.68	54.44	71.98	62.73	81.88	21.80	80.10	82.17	71.93	77.21	67.58	80.22	38.50	68.00	75.60	66.15	68.91	65.69
SimSiam	85.55	29.30	78.92	63.62	46.89	71.34	60.85	84.77	21.06	81.59	83.56	73.09	79.23	68.81	81.55	38.05	68.74	76.13	66.05	69.68	66.10
Barlow Twins	87.85	29.20	**81.84**	69.69	57.33	75.13	65.18	84.36	23.84	81.69	83.82	74.13	79.27	69.57	83.84	40.02	71.24	78.65	68.68	72.23	68.49
MAE	81.23	27.39	76.72	60.99	46.91	68.31	58.65	85.28	18.38	83.95	84.90	75.86	81.08	69.67	81.94	36.41	74.01	83.69	58.42	71.49	66.89
SimMIM	83.68	26.69	75.73	51.32	44.32	69.94	56.35	84.94	22.38	83.14	85.11	75.25	80.45	70.16	82.05	35.75	74.43	84.13	58.79	71.64	67.03
Data2Vec	83.39	28.54	75.41	56.54	50.82	69.78	58.94	78.07	16.65	79.08	81.02	71.24	76.18	65.21	82.22	36.59	73.42	83.09	60.83	71.71	67.23
BENDR	72.15	28.28	67.83	50.94	32.50	57.42	50.34	52.34	08.41	72.72	78.37	54.78	65.08	53.32	52.75	13.83	72.18	78.28	55.69	64.78	54.55
ContraWR	88.40	34.35	81.67	68.81	62.78	75.79	67.20	85.78	25.51	84.20	85.79	77.53	81.65	71.76	84.09	40.23	73.26	82.55	**71.18**	74.07	70.26
TS-TCC	73.28	21.15	66.01	41.39	37.33	61.45	47.83	70.05	17.68	75.33	73.23	62.00	70.43	59.66	80.91	32.06	70.13	80.27	64.17	70.17	65.51
mulEEG	89.09	**36.52**	80.69	69.62	59.66	74.92	67.12	83.67	21.41	83.06	85.82	74.16	79.94	69.62	80.87	36.69	71.25	82.56	65.77	71.58	67.43
NeuroNet-T	**89.90**	30.21	81.51	71.39	59.51	76.26	66.50	83.97	13.95	83.30	85.45	73.75	80.45	68.09	84.51	39.40	76.15	84.68	67.63	75.12	70.47
NeuroNet-B	89.17	36.24	81.74	**69.97**	**63.82**	**76.74**	**68.19**	**88.27**	**30.76**	**86.20**	**87.56**	**79.41**	**84.13**	**74.44**	**85.08**	**42.11**	**76.84**	**85.74**	71.04	**76.47**	**72.16**

TABLE 3
Comparision between supervised learning-based methodologies and NeuroNet+TCM.

	EEG Epoch	*Sleep-EDFX*							*SHHS*							*ISRUC-Sleep*						
		Per-Class F1					*Overall*		*Per-Class F1*					*Overall*		*Per-Class F1*					*Overall*	
		W	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*	*W*	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*	*W*	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*
DeepSleepNet	25	90.84	35.56	81.42	68.59	68.02	77.49	68.89	83.84	18.87	83.55	84.67	76.80	81.02	69.55	81.55	38.25	68.90	81.17	62.17	69.84	66.41
IITNet	10	92.59	45.77	83.61	63.65	80.90	81.48	73.30	89.32	47.38	85.57	80.96	87.58	84.74	78.16	84.60	40.51	78.39	85.27	79.15	77.89	73.59
U-Sleep	35	92.71	47.72	84.65	65.23	82.44	82.42	74.55	89.56	48.08	86.53	81.76	**88.82**	85.59	78.95	**86.34**	44.16	**79.07**	85.38	**81.48**	**78.89**	**75.29**
AttnSleep	1	91.49	40.44	83.84	72.28	72.18	79.68	72.05	87.28	28.13	83.83	85.39	77.72	81.98	72.47	84.54	42.41	75.81	83.45	69.92	75.72	71.23
SleepExpertNet	20	92.80	52.75	85.92	73.40	80.93	83.13	77.16	**90.84**	40.71	86.60	84.04	87.94	85.94	78.03	72.59	14.25	66.61	72.69	58.05	64.89	56.84
NeuroNet-T+TCM	20	92.27	53.01	85.19	75.23	77.13	82.67	76.57	86.44	50.17	85.97	85.97	83.99	84.62	78.51	83.75	44.73	77.50	86.61	72.04	76.79	72.93
NeuroNet-B+TCM	20	**93.15**	**58.80**	**87.21**	**76.97**	**83.00**	**85.24**	**79.82**	89.05	**55.29**	**88.09**	**86.48**	87.25	**86.88**	**81.23**	84.50	**46.09**	77.25	**86.86**	72.57	77.05	73.45

TABLE 4
Cross-dataset evaluation experiment applied to different PSG datasets.

	EEG Epoch	*A → B*		*A → C*		*B → A*		*B → C*		*C → A*		*C → B*		*Average*	
		ACC	*MF1*	*ACC*	*MF1*	*ACC*	*MF1*	*ACC*	*MF1*	*ACC*	*MF1*	*ACC*	*MF1*	*ACC*	*MF1*
U-Sleep	35	56.50	46.22	56.71	49.73	71.39	59.09	70.08	64.72	**59.38**	45.62	60.08	54.36	62.36	45.00
SleepExpertNet	20	58.63	48.65	58.46	51.93	70.75	59.21	71.19	65.75	58.15	**47.22**	67.41	60.77	64.10	55.59
NeuroNet-T	1	66.94	52.83	61.17	52.71	79.44	69.57	73.50	63.28	50.18	42.21	74.85	64.81	67.68	57.56
NeuroNet-B	1	66.68	52.83	61.06	53.19	79.80	70.40	74.21	63.98	50.20	42.14	75.83	65.57	67.93	58.01
NeuroNet-T+TCM	20	73.82	63.20	**66.93**	**57.92**	84.93	78.32	82.09	75.29	51.74	44.18	**83.01**	**75.66**	**73.75**	65.76
NeuroNet-B+TCM	20	**73.86**	**64.12**	65.29	56.32	**85.12**	**79.98**	**83.95**	**76.59**	50.93	42.03	82.86	75.59	73.67	**65.77**

*A: Sleep-EDFX, B: SHHS, C: ISRUC-Sleep

In conclusion, it has been confirmed that the proposed models showcase superior generalization performance when compared with supervised learning methodologies, effectively operating on datasets extending beyond the scope of their training dataset.

5.2 Ablation Experiments

An ablation experiment was conducted to derive the optimal settings information for the proposed model. Across all experiments, NeuroNet-B served as the base backbone, with Sleep-EDFX serving as the reference dataset. In light of the results obtained from these experiments, hyperparameters were established.

5.2.1 Evaluation Scenario 1: Linear Evaluation using Single-epoch EEG

(Frame Design) The performance is evaluated based on various frame designs. Looking at Table 5, it indicates a trend where decreasing the frame size and overlap step generally leads to performance enhancements, but they significantly increase the training speed. Specifically, the configuration with a frame size of 3 and an overlap step of 0.375 achieves the highest performance, exhibiting an ACC of 76.74% and a MF1 of 68.19%. Therefore, in this study, the frame size and overlap step were fixed at 3 and 0.375, respectively. Considering the impact on training speed, the process of reducing these two values was omitted.

(Masking Ratio) At high masking ratio, NeuroNet demonstrates excellent performance. Figure 3 illustrates the configurations for analyzing the effects on two tasks. NeuroNet w/o masked prediction shows highest accuracy at masking ratios of 70% to 75%, while NeuroNet w/o contrastive learning exhibits exceptional performance at ratios of 75% to 90%. NeuroNet, applying both tasks, outperforms single-task across all masking ratios, indicating their mutual complementarity. NeuroNet achieves its highest performance at a masking ratio of 75%.

TABLE 5
Linear evaluation under different frame size and overlap size.

Frame Setting (sec)		Performance		Training Time / Epoch (min)
Frame Size	*Overlap Step*	*ACC*	*MF1*	
3	**0.375**	**76.74**	**68.19**	**20:33**
	0.75	76.52	67.98	10:35
	1.5	76.49	67.62	05:46
4	0.5	76.36	67.26	15:36
	1	76.32	67.33	07:47
	2	76.42	67.14	04:19
5	0.625	76.28	67.07	11:34
	1.25	76.65	67.38	06:13
	2.5	76.07	67.00	03:35
6	0.75	76.12	67.01	09:43
	1.5	76.11	66.59	05:08
	3	75.95	66.46	02:59

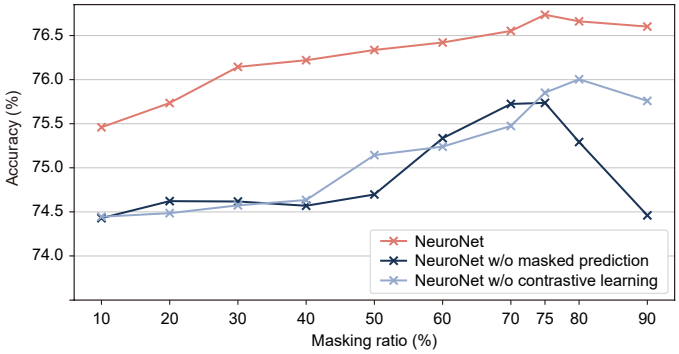

Fig. 3. Impact of different masking ratios on NeuroNet performance.

(Decoder Depth and Width) Upon reviewing the linear evaluation under different decoder dimensions and decoder depths (Table 6), NeuroNet tends to achieve superior performance when using smaller decoders. This is attributed to the characteristic of a smaller decoder, which requires a greater amount of semantic information to successfully accomplish the reconstruction task. Consequently, this necessitates the encoder to generate representations imbued with richer information, thereby enhancing the overall performance of the model. However, it was observed that if the decoder is too small, performing the reconstruction task becomes excessively challenging, leading to degraded performance. Therefore, NeuroNet achieves its highest performance when the dimension and depth of the decoder are set to 256 and 3, respectively.

5.2.2 Evaluation Scenario 2: Fine-tuning using Multi-epoch EEG

(Temporal Context Module) The optimal structure for effectively analyzing temporal variations or correlations among multiple EEG epochs was explored (Table 7). It was observed that the Mamba-based structure outperforms the widely used LSTM or multi-head attention-based structures in previous studies. Furthermore, examining the performance of Mamba based on the context length revealed that the best performance is achieved when the context length is 20.

TABLE 6
Linear evaluation under different decoder dimensions and decoder depths.

Decoder		*Performance*		*Model Size (MB)*
Dim	*Depth*	*ACC*	*MF1*	
192	1	76.05	66.83	123.96
	2	76.21	67.34	125.74
	3	76.23	67.24	127.52
	4	76.32	67.35	129.30
256	1	76.05	66.83	125.62
	2	76.51	67.40	128.78
	3	**76.74**	**68.19**	**131.94**
	4	75.94	66.56	135.10
512	1	76.03	66.65	136.20
	2	76.02	67.17	148.81
	3	76.02	66.48	161.42
	4	75.84	66.52	174.03

TABLE 7
Comparison of modules and context lengths comprising temporal context module.

Model	*Context Length*	*Performance*	
		ACC	*MF1*
LSTM	20	80.65	74.67
Multi-Head Attention	20	81.56	75.27
LSTM + Multi-Head Attention	20	80.90	74.62
Mamba-based TCM (ours)	10	83.74	77.75
Mamba-based TCM (ours)	**20**	**85.24**	**79.82**
Mamba-based TCM (ours)	30	85.13	79.80

5.3 Hypnograms

Figure 4 depicts the predicted results (i.e., hypnograms) for one subject from each of the three PSG datasets. The top row represents the labels annotated by sleep experts, the middle row corresponds to NeuroNet+TCM, and the bottom row to NeuroNet alone. The difference between the 2 figures is that the former shows results for NeuroNet-B and the latter for NeuroNet-T. In detail, it can be observed that across all three PSG datasets, the application of TCM yields results more closely aligned with those annotated by sleep experts. This underscores the significance of effectively incorporating temporal context information, thereby contributing to performance improvement. Moreover, it is clear that NeuroNet-B, despite its increased number of parameters, generally tends to show higher accuracy compared to NeuroNet-T.

6 Discussions

NeuroNet is a novel SSL framework that effectively combines the contrastive learning task with the masked prediction task. This study demonstrates that greater accuracy is achieved in NeuroNet performance when the contrastive learning task and masked prediction task are combined. It suggests that these two tasks mutually complement each other and lead to improved stability and the acquisition of higher-level representations. Consequently, NeuroNet showcases superior performance in comparison to the recent SSL methodologies (Table 2). Furthermore, it has been observed that when subjected to fine-tuning with a sparse number of labeled data, the NeuroNet+TCM configuration not only contends with but surpasses the performance of the latest supervised learning methodologies trained on substantially larger labeled data (Table 3, 4).

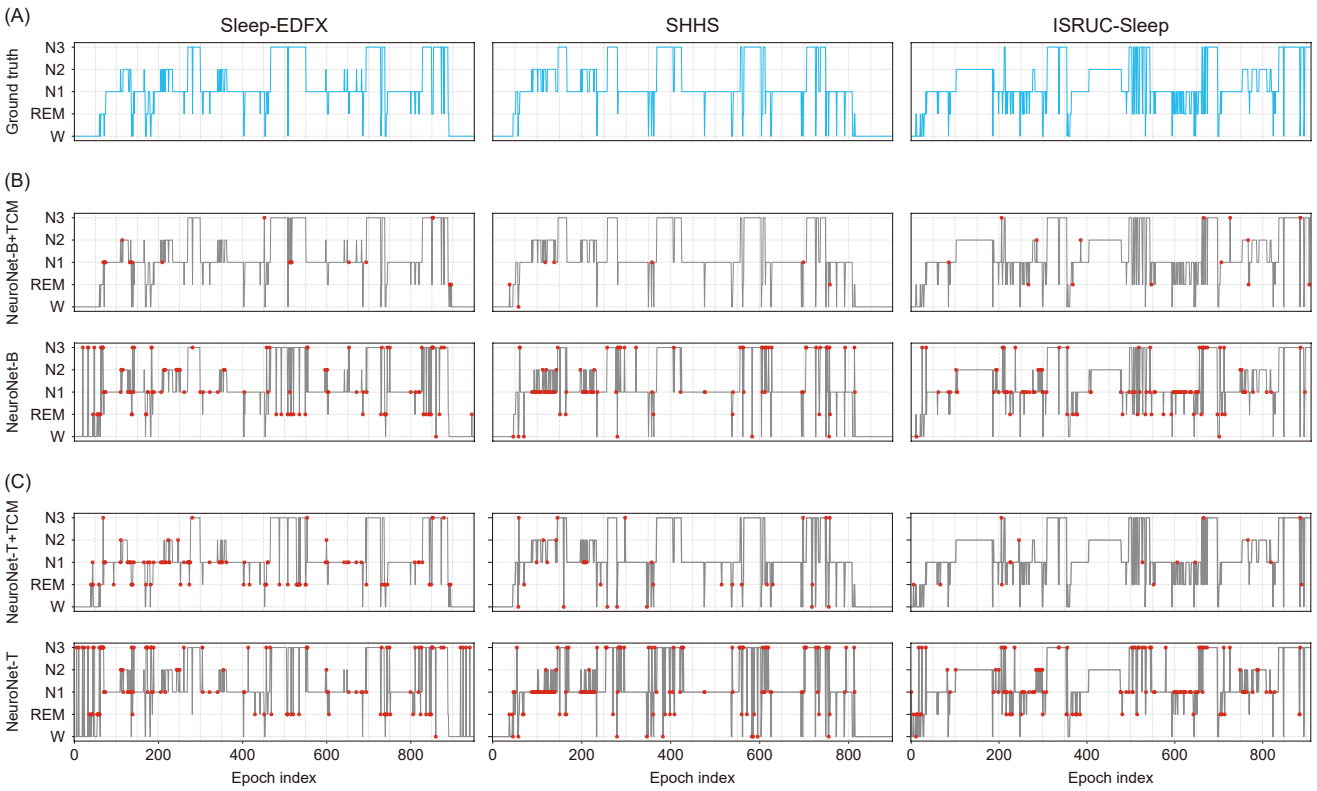

Fig. 4. The output hypnograms across five sleep stages. The first, second, and third columns correspond to #sc4031e0, #shhs1-204928, and #subject-53 within Sleep-EDFX, SHHS, and ISRUC, respectively. (A) is manually scored by a sleep expert. (B) and (C) respectively represent NeuroNet-B and NeuroNet-T. The first row for both (B) and (C) displays the results for NeuroNet+TCM, while the second row shows the results for NeuroNet. The errors are marked by the red dots.

NeuroNet demonstrates exceptional performance and offers the advantage of omitting contrived EEG data augmentation for contrastive learning tasks. Generally, SSL methodologies based on contrastive learning necessitate the process of contrived data augmentation, posing a range of inherent challenges. Firstly, unlike images, the application of data augmentation to EEG signals risks diluting their intrinsic meaning. Secondly, selecting appropriate EEG data augmentation could be challenging, with suboptimal choices of data augmentation leading to ineffective SSL outcomes [8], [21], [22]. Lastly, even when data augmentation is optimized for a specific model or dataset, there is a possibility it may not perform effectively with different models or datasets [8], [9], [32]. NeuroNet inputs two subsets, randomly sampled differently, into the encoder, resulting in output vectors corresponding to different views, and then conducts a contrastive learning task on these vectors. This approach can be viewed as a challenge of determining whether the partially obscured portions of the entire EEG signal are the same or different. Compared to conventional methods, this approach is much simpler and effectively resolves the issues associated with contrived EEG data augmentation.

NeuroNet-B demonstrates higher performance compared to NeuroNet-T in most cases, primarily due to the relatively large scale of PSG datasets (Table 3, 4). The higher number of parameters enables the capture of more intricate and diverse patterns, thereby conferring an advantage in augmenting performance. Despite this, NeuroNet-T demonstrates superior performance over other SSL methodologies on the Sleep-EDFX and ISRUC-Sleep datasets and is competitive on the SHHS dataset. This suggests that although NeuroNet-T has fewer parameters, it still effectively captures the characteristics of EEG signals. Given the inherent challenges associated with EEG data acquisition, which often result in limited dataset capacity, leveraging NeuroNet-T with its fewer parameters may be a pragmatic strategy for optimal performance under such constraints.

The architecture of TCM, intricately designed to discern the relationships between different EEG epochs and based on the Mamba, emerged as a key driver for performance improvement. The overall performance exhibited a notable increase of approximately 4-5% upon Mamba-based TCM, compared to the methodologies predominantly employed in prior studies, such as LSTM or multi-head attention-based methodologies (Table 7). Consequently, the model combining NeuroNet and the Mamba-based TCM (= NeuroNet+TCM), despite being trained on a limited amount of labeled data, demonstrated superior or comparable performance to supervised learning-based sleep staging trained on a vast amount of labeled data. This illustrates that the combination of NeuroNet, which effectively represents EEG features, and Mamba, specialized in intricate sequence modeling, yields highly efficacious outcomes. In particular, Mamba has addressed inefficiencies in long sequences and also improved performance by allowing the parameters of the SSM to be a function of the input.

Despite these advantages, there are still issues that need to be addressed. Firstly, SSL methodologies are designed to

leverage unlabeled data by learning representations without explicit supervision. However, they do incorporate a small amount of labeled data at some point in the model development process. This means that the quality and accuracy of the initial labels may affect the performance of SSL-based methodologies. This can be particularly problematic during fine-tuning, where the quality of labels directly influences the model's ability to generalize from the learned representations to specific downstream tasks. This issue is especially pertinent in this study, which utilized public PSG data, where the unreliability of labels can and likely will be present. Therefore, future research will focus on conducting studies on "noisy label classification" optimized for sleep EEG signals to solve the issue of label reliability. Simultaneously, we plan to employ a large number of highly skilled sleep experts to select several support sets representing each sleep stage and then supplement them through few-shot or zero-shot learning to achieve accurate results without further training. Secondly, NeuroNet has been found to improve performance as the frame size and overlap size decrease, but this comes with a significant increase in computing cost (Table 5). This is because the core component of NeuroNet, the Transformer, struggles with efficiently processing samples with long sequences. Therefore, future research is expected to explore replacing Transformer with Mamba to achieve superior EEG representations along with more efficient computation, which would improve inference speed.

APPENDIX A
TRAINING SETTINGS AND HYPERPARAMETERS

The training and evaluation of the model were conducted on a computer equipped with an Intel I9-9980XE CPU at 3.00GHz, 128GB RAM, and an NVIDIA GPU 3090. Furthermore, all data processing and algorithm development was carried out using Python version 3.9, with the Pytorch version 1.10 library being utilized. The detailed hyperparameters are described in Table A1. Evaluation scenario 3 shares the same hyperparameters as scenario 2, as it does not involve an additional training process.

TABLE A1
Hyperparameters for evaluation scenario for NeuroNet.

	Scenario 1	*Scenario 2*
	Self-Supervised Learning	
epoch	50	
batch size	1024	
frame size	3	
overlap step	0.75	
encoder dim	T: 512 / B: 768	
encoder depth	T: 4 / B: 4	
encoder head	8	
decoder dim	T: 192 / B: 256	
decoder depth	T: 1 / B: 3	
decoder head	8	
projection hidden	(1024, 512)	
temperature scale	0.5	
mask ratio	~	
optimizer	AdamW	
optimizer momentum	(0.9, 0.999)	
learning rate	2e-05	
	Downstream Task	
epoch	300	100
batch size	512	128
optimizer	AdamW	AdamW
optimizer momentum	(0.9, 0.999)	(0.9, 0.999)
learning rate	1e-05	5e-03
temporal context length	-	20
mamba_d_state	-	16
mamba_d_conv	-	4
mamba_expand	-	2

APPENDIX B
CONTRASTIVE LEARNING BASED SSL WITH SINGLE-CHANNEL EEG

SSL methodologies based on contrastive learning, such as SimCLR [21], BYOL [22], SwAV [24], SimSiam [23], Barlow Twins [25], etc., have demonstrated remarkable performance in the field of computer vision. However, compared to SSL methodologies based on masked prediction tasks (e.g., MAE [27], Data2Vec [30], etc.), their representation performance can vary significantly across different scenarios, such as data augmentation techniques and types of backbone networks. Thus, in this research, the following steps were taken to effectively apply contrastive learning-based SSL methodologies to single-channel EEG.

B.1 Data Augmentation

For effective training using contrastive learning-based SSL methodologies, selecting appropriate data augmentations is crucial [21], [44]. Inspired by [7], [8], [44], this research implemented five data augmentations optimized for single-channel EEG. During training, two data augmentations were randomly selected and applied to each data sample. Detailed descriptions of each data augmentation follow.

- **(Random Gaussian Noise)** Adds Gaussian noise to the original signal.
- **(Random Crop)** Randomly crops a portion of the original signal and then interpolates it back to the original signal size.
- **(Random Bandpass Filtering)** Selects a frequency band at random and applies band-pass filtering to the original signal.
- **(Random Temporal Cutout)** Randomly selects a segment of the original signal and replaces it with the mean value of the original signal.
- **(Random Permutation)** Segments the signal randomly, shuffles these segments in a random order, and then merges them.

B.2 Backbone Network

To compare and analyze the performance of SSL methodologies based on different backbone networks, three backbone networks were selected and implemented. Each selected backbone network is a deep learning algorithm designed to analyze sleep stages using single-epoch EEG as input. Detailed descriptions are as follows:

- **(DeepSleepNet [16])** Utilizes two CNNs with different kernel sizes to extract low- and high-frequency features. Each CNN comprises four convolution layers and two max pooling layers. For this study, the bi-LSTM component of DeepSleepNet, typically used

for processing single-epoch EEG, was omitted to solely focus on input from a single epoch.

- **(IITNet, intra-epoch version [17])** Designed to process a single imbued EEG by dividing it into sub-epochs at fixed intervals, which are then fed into a ResNet to extract a representation vector. This vector is subsequently used as the input for a bi-LSTM, enabling the capture of temporal context.
- **(STFT Encoder [9], [32])** This model is designed to convert raw EEG signals into short-time Fourier transform spectrograms, which are then fed into four ResNet to facilitate the learning of sleep EEG signal features. This model has been widely adopted as a backbone model for SSL

B.3 Results

Table B2 presents the results of conducting 5-fold cross-validation on the Sleep-EDFX and SHHS, and 10-fold cross-validation on the ISRUC-Sleep dataset. Detailed examination reveals that the 'STFT Encoder' demonstrated the highest performance across all datasets, with the exception of Sleep-EDFX. Compared to other SSL methodologies, the Barlow Twins [25] method showed superior performance overall.

APPENDIX C
NORMALIZED CONFUSION MATRICES

Appendix Figures 1 and 2 display the normalized confusion matrices. The distinction lies in that Figure 2 presents results with the TCM, unlike Figure 1. Examination of both figures reveals that the application of TCM contributes to an overall enhancement in performance, with NeuroNet-B demonstrating greater accuracy compared to NeuroNet-T.

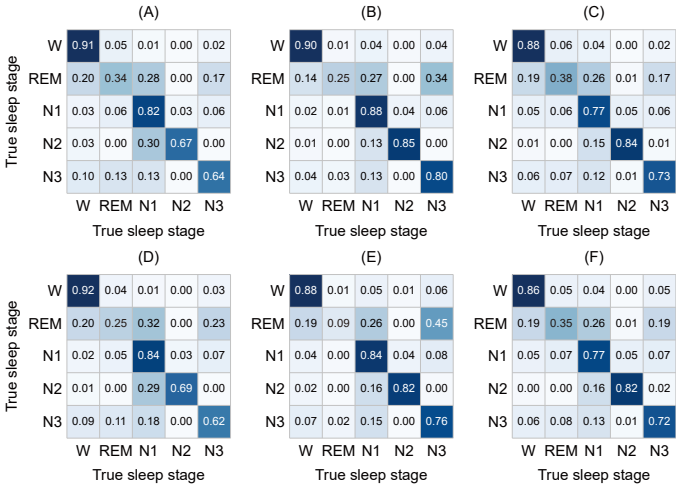

Fig. 1. The confusion matrices for sleep stage classification from evaluation scenario 1. The columns correspond to Sleep-EDFX, SHHS, and ISRUC-Sleep, respectively. Moreover, the first row signifies NeuroNet-B, and the second row depicts NeuroNet-T.

ACKNOWLEDGMENTS

We extend our sincere gratitude to Hyun-Ku Kang for his diverse contributions during the initial phases of our work; This work was supported by a National Research Foundation of Korea (NRF) Grant funded by the Korean government (Ministry of Science and ICT, MSIT) (No. 2022R1A2C1013205);

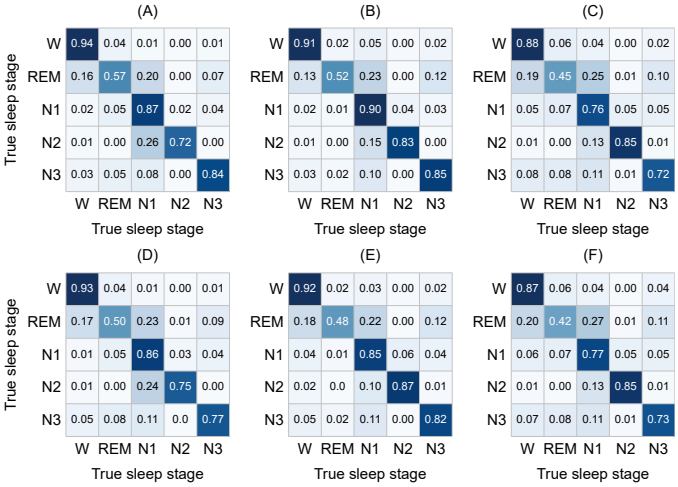

Fig. 2. The confusion matrices for sleep stage classification from evaluation scenario 2. The columns correspond to Sleep-EDFX, SHHS, and ISRUC-Sleep, respectively. Moreover, the first row signifies NeuroNet-B+TCM, and the second row depicts NeuroNet-T+TCM.

REFERENCES

[1] J. M. Siegel, "Clues to the functions of mammalian sleep," *Nature*, vol. 437, no. 7063, pp. 1264–1271, 2005.
[2] M. W. Mahowald and C. H. Schenck, "Insights from studying human sleep disorders," *Nature*, vol. 437, no. 7063, pp. 1279–1285, 2005.
[3] S. D. Davis, E. Eber, A. C. Koumbourlis *et al.*, *Diagnostic tests in pediatric pulmonology*. Springer, 2015.
[4] E. Alickovic and A. Subasi, "Ensemble svm method for automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, 2018.
[5] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring," *Journal of clinical sleep medicine*, vol. 9, no. 1, pp. 81–87, 2013.
[6] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.
[7] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.
[8] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," *arXiv preprint arXiv:2106.14112*, 2021.
[9] V. Kumar, L. Reddy, S. Kumar Sharma, K. Dadi, C. Yarra, R. S. Bapi, and S. Rajendran, "muleeg: a multi-view representation learning on eeg signals," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 398–407.
[10] S. Mishra, J. Robinson, H. Chang, D. Jacobs, A. Sarna, A. Maschinot, and D. Krishnan, "A simple, efficient and scalable contrastive masked autoencoder for learning visual representations," *arXiv preprint arXiv:2210.16870*, 2022.
[11] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
[13] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.

TABLE B2
Linear evaluation based contrastive learning using single-epoch EEG.

		Sleep-EDFX							***SHHS***							***ISRUC-Sleep***						
		Per-Class F1					*Overall*		*Per-Class F1*					*Overall*		*Per-Class F1*					*Overall*	
		W	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*	*W*	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*	*W*	*N1*	*N2*	*N3*	*REM*	*ACC*	*MF1*
SimCLR	*DeepSleepNet*	84.70	26.41	**81.61**	**69.01**	**54.54**	**73.32**	63.25	72.69	00.00	78.43	82.65	64.76	74.13	59.71	78.88	33.41	**70.24**	**79.88**	56.33	68.74	63.75
	IITNet	**85.95**	32.25	79.44	68.65	53.59	71.49	**63.98**	81.63	14.07	79.09	79.85	67.27	75.53	64.38	74.17	26.93	63.95	76.59	53.33	62.87	58.99
	STFT Encoder	85.14	**34.33**	78.72	65.31	55.72	71.97	63.84	**83.61**	**23.14**	**80.64**	**82.75**	**71.92**	**77.75**	**68.41**	**82.24**	**39.41**	69.38	78.43	**68.31**	**70.98**	**67.55**
BYOL	*DeepSleepNet*	86.41	31.14	**81.16**	**68.85**	45.78	**72.71**	62.67	78.54	00.57	79.81	83.83	68.34	76.54	62.22	78.93	33.34	**70.71**	78.03	60.38	69.09	64.28
	IITNet	**87.08**	**33.40**	78.32	64.49	**55.12**	71.82	**63.68**	83.20	**19.50**	81.71	**84.52**	71.32	78.51	**68.05**	79.44	31.35	68.31	**78.23**	62.45	68.13	63.96
	STFT Encoder	85.84	30.10	79.13	65.26	47.52	71.69	61.57	**85.06**	07.16	**82.00**	83.93	**73.36**	**79.91**	66.30	**82.30**	**40.62**	70.42	77.48	**67.76**	**71.22**	**67.72**
SwAV	*DeepSleepNet*	83.62	24.07	**81.54**	**68.38**	53.90	**72.71**	62.30	63.85	00.00	75.27	80.71	60.53	70.05	56.07	79.40	35.88	**69.46**	**77.00**	57.75	68.08	63.90
	IITNet	**85.45**	32.25	79.09	65.91	50.40	70.59	62.62	79.83	13.07	76.43	76.49	66.10	72.95	62.38	71.05	24.88	63.37	74.17	50.88	61.38	56.87
	STFT Encoder	85.11	**32.49**	78.92	62.68	**54.44**	71.98	**62.73**	**81.88**	**21.80**	**80.10**	**82.17**	**71.93**	**77.21**	**67.58**	**80.22**	**38.50**	68.00	75.60	**66.15**	**68.91**	**65.69**
SimSiam	*DeepSleepNet*	81.45	23.97	76.46	**66.93**	28.00	67.79	55.36	71.44	00.06	73.71	73.84	57.87	69.20	55.38	78.36	34.81	**71.20**	**79.17**	61.16	69.47	64.94
	IITNet	83.98	**29.55**	77.14	61.75	46.81	69.05	59.85	72.48	01.75	74.10	79.47	59.11	70.31	57.38	75.23	24.97	63.35	73.06	56.51	63.20	58.62
	STFT Encoder	**85.55**	29.30	**78.92**	63.62	**46.89**	**71.34**	**60.85**	**84.77**	**21.06**	**81.59**	**83.56**	**73.09**	**79.23**	**68.81**	**81.55**	**38.05**	68.74	76.13	**66.05**	**69.68**	**66.10**
Barlow Twins	*DeepSleepNet*	**87.85**	29.20	**81.84**	**69.69**	**57.33**	**75.13**	**65.18**	77.70	00.00	80.21	**83.87**	67.68	76.58	61.89	80.97	36.49	**71.27**	**79.64**	64.29	70.80	66.53
	IITNet	86.46	28.99	77.99	61.75	52.97	70.98	61.63	82.94	19.59	80.95	83.34	72.01	78.04	67.77	78.92	27.91	67.16	77.65	57.96	66.74	61.92
	STFT Encoder	86.03	**31.59**	80.40	66.12	55.48	73.18	63.92	**84.36**	**23.84**	**81.69**	83.82	**74.13**	**79.27**	**69.57**	**83.84**	**40.02**	71.24	78.65	**68.68**	**72.23**	**68.49**

[14] ——, "Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.

[15] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5903–5915, 2021.

[16] A. Supratak, H. Dong, C. Wu, and Y. 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, 2017.

[17] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, "Intra- and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg," *Biomedical signal processing and control*, vol. 61, p. 102037, 2020.

[18] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[19] C.-H. Lee, H.-J. Kim, Y.-T. Kim, H. Kim, J.-B. Kim, and D.-J. Kim, "Sleepexpertnet: high-performance and class-balanced deep learning approach inspired from the expert neurologists for sleep stage classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2022.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. 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.

[21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[22] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, and M. Gheshlaghi Azar, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[23] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[24] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.

[26] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[28] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[29] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[30] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[31] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. 15, p. 653659, 2021.

[32] C. Yang, D. Xiao, M. B. Westover, and J. Sun, "Self-supervised eeg representation learning for automatic sleep staging," *arXiv preprint arXiv:2110.15278*, 2021.

[33] J. Ye, Q. Xiao, J. Wang, H. Zhang, J. Deng, and Y. Lin, "Cosleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification," *IEEE Signal Processing Letters*, vol. 29, pp. 189–193, 2021.

[34] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng, "Maeeg: Masked auto-encoder for eeg representation learning," *arXiv preprint arXiv:2211.02625*, 2022.

[35] F. Wang, J. Han, S. Zhang, X. He, and D. Huang, "Csi-net: Unified human body characterization and pose recognition," *arXiv preprint arXiv:1810.03064*, 2018.

[36] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[37] M. M. Grigg-Damberger, "The aasm scoring manual: a critical appraisal," *Current opinion in pulmonary medicine*, vol. 15, no. 6, pp. 540–549, 2009.

[38] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *J. circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[39] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.

[40] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, and J. M. Samet, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[41] P. Fonseca, N. Den Teuling, X. Long, and R. M. Aarts, "Cardiorespiratory sleep stage detection using conditional random fields,"

IEEE journal of biomedical and health informatics, vol. 21, no. 4, pp. 956–966, 2016.
[42] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "Isruc-sleep: A comprehensive public dataset for sleep researchers," *Computer methods and programs in biomedicine*, vol. 124, pp. 180–192, 2016.
[43] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-sleep: resilient high-frequency sleep staging," *NPJ digital medicine*, vol. 4, no. 1, p. 72, 2021.
[44] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," *arXiv preprint arXiv:2007.04871*, 2020.