PSG-MAE: Robust Multitask Sleep Event Monitoring Using Multichannel PSG Reconstruction and Inter-Channel Contrastive Learning

Yifei Wang, Qi Liu, *Senior Member, IEEE*, Fuli Min, and Honghao Wang

***Abstract*— Polysomnography (PSG) signals are essential for studying sleep processes and diagnosing sleep disorders. With the advancement of deep neural networks (DNNs), automated analysis of PSG data has become increasingly feasible. However, the limited availability of data for certain sleep events often restricts DNNs to single-task learning on a single-source dataset, limiting their ability to generalize to new events and reducing robustness across datasets. To address these challenges, we propose PSG-MAE, a pretraining framework based on the masked autoencoder (MAE). By leveraging self-supervised learning on large volumes of unlabeled PSG data, PSG-MAE trains a robust feature extraction network applicable to diverse sleep event monitoring tasks. Unlike conventional MAEs, PSG-MAE applies complementary masking across PSG channels, integrates a multichannel signal reconstruction mechanism, and incorporates an inter-channel contrastive learning (ICCL) strategy. This design enables the encoder to capture temporal features from each channel while simultaneously modeling latent inter-channel relationships, thereby enhancing the utilization of multichannel information. Experimental results demonstrate that PSG-MAE effectively learns both temporal details and inter-channel dependencies from PSG signals. When the pretrained encoder is fine-tuned with downstream sleep event monitoring networks, it achieves a macro-averaged F1-score of 81.0% for sleep staging and 82.6% for apnea detection on the SHHS dataset. Cross-dataset validation on the MESA dataset further confirms the framework's robustness and broad applicability. The code of PSG-MAE is available at https://github.com/yfw-scut/PSG-MAE.git**

Received 27 March 2025; revised 13 August 2025, 9 October 2025, and 19 November 2025; accepted 13 December 2025. Date of publication 17 December 2025; date of current version 24 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the GJYC program of Guangzhou under Grant 2024D01J0081, in part by the ZJ program of Guangdong under Grant 2023QN10X455, and in part by the Fundamental Research Funds for the Central Universities under Grant 2025ZYGXZR053. *(Corresponding authors: Qi Liu; Honghao Wang.)*

Yifei Wang and Qi Liu are with the School of Future Technology, South China University of Technology, Guangzhou 511400, China (e-mail: ywang634@outlook.com; drliuqi@scut.edu.cn).

Fuli Min and Honghao Wang are with the Department of Neurology, Guangzhou First People's Hospital, and the School of Medicine, South China University of Technology, Guangzhou 510180, China (e-mail: minfuli@163.com; wang_whh@163.com).

Digital Object Identifier 10.1109/TNSRE.2025.3645353

***Index Terms*— Polysomnography signal analysis, multichannel signal reconstruction, pretrained deep learning models, sleep stage classification, sleep apnea detection.**

I. Introduction

SLEEP is a necessity for life maintenance. Consistent and adequate rest is crucial for improving health, productivity, well-being, and quality of life as well as public safety [1]. In recent years, the accelerated pace of global urbanization and rising stress have exacerbated the prevalence of sleep disorders, posing a substantial challenge to public health. Common sleep disorders, such as insomnia, arousal disorders, sleep apnea, rapid eye movement sleep behavior disorder (RBD), and periodic limb movement disorder (PLMD), are associated with a heightened risk of medical complications, including cardiovascular diseases, depression, anxiety, diabetes, and compromised immune function [2]. Consequently, the development of automated screening and intervention methods for sleep disorders is of significant research value.

Currently, combining polysomnography (PSG) with deep neural networks (DNNs) has become a widely explored approach in automated sleep monitoring research [3]. PSG is widely considered the most reliable method in the field of sleep medicine for diagnosing sleep-related disorders, often employed to evaluate both the diagnosis and efficacy of treatment for sleep disturbances [4], [5]. A standard PSG recording gathers data on brain waves (electroencephalography, EEG), eye movements (electrooculography, EOG), chin and leg muscle activity (electromyography, EMG), heart activity (electrocardiography, ECG), chest and abdominal breathing effort, nasal airflow, oxygen saturation, etc [6]. While PSG provides comprehensive documentation of sleep patterns, the analysis and clinical interpretation of these neurophysiological recordings demand rigorous systematic training. Additionally, annotating PSG data is labor-intensive and time-consuming, with 2-3 experts typically spending about 2 hours to annotate an 8-hour sleep recording. Subjective differences among experts can also lead to variability in annotation results [7], [8]. The process of annotating PSG sleep data involves categorizing sleep stages and identifying sleep events. Following the sleep staging guidelines outlined by the American Academy

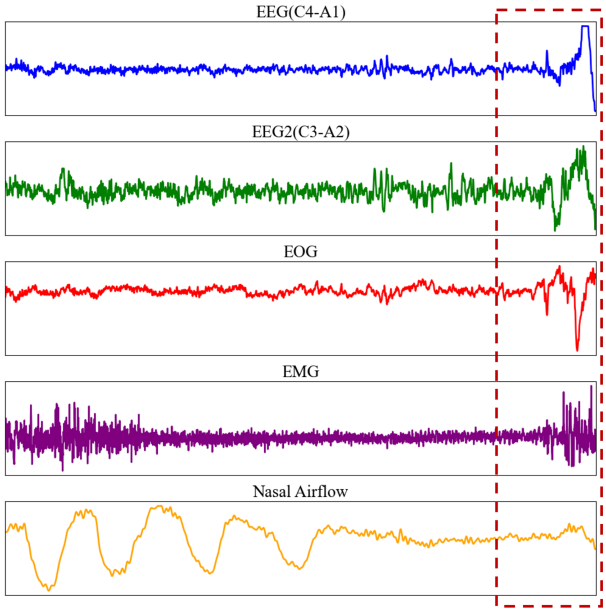

Fig. 1. PSG of one sleep epoch (30s), during which arousal occurs, marked in the dashed box. A sleep event during sleep often induces abrupt fluctuations in multiple channels of the PSG signals. Integrating the variations across different channels can help improve the accuracy of sleep event monitoring.

of Sleep Medicine (AASM), PSG data is divided into 30-second segments along the temporal dimension. Each epoch is then classified into stages, including wake (W), non-rapid eye movement (NREM) stages (N1, N2, N3), and rapid eye movement (REM) sleep [9]. The epoch-based segmentation approach is used in contemporary clinical practice to label sleep events, ensuring consistent analysis across research studies.

The multichannel nature of PSG signals makes them well-suited for integration with machine learning and deep neural networks, enabling the automatic extraction of complex sleep features and effectively modeling nonlinear relationships for accurate sleep event annotation [10], [11]. Current PSG-driven automated sleep events monitoring can be divided into two main areas. One focuses on the automatic sleep staging [12], [13], [14], while the other involves the detection and labeling of sleep behaviors, events, and disorders [15], [16]. However, there are two main challenges in the current sleep event monitoring models. First, the wide variety of sleep events and their different manifestations across populations result in a limited quantity of public datasets for certain sleep events [17], [18]. As a result, many models are trained on small, task-specific datasets, making them sensitive to the feature distribution of the data. This limits their ability to transfer to other sleep event monitoring tasks and hinders a comprehensive, multidimensional evaluation of sleep. Second, most current sleep models rely on only 1-2 PSG signal channels for specific monitoring tasks, neglecting the potential inter-channel interactions. As shown in Fig. 1, sleep events often induce signal changes across multiple channels. By integrating information from multiple channels, we can reduce misjudgments caused by disturbances in individual channels while improving the overall accuracy of event detection.

To address the challenges mentioned above, we propose leveraging unlabeled data through self-supervised learning to improve the model's stability in PSG feature extraction and enhance its performance in multichannel information fusion. To this end, we introduce PSG-MAE, a novel pretraining framework for PSG signals. In contrast to normal masked autoencoders (MAEs), PSG-MAE is based on a complementary-masking strategy for multichannel PSG signal reconstruction, meaning that one epoch (30 seconds) of PSG data is used as input with a pair of complementary masks generated along the channel dimension. Based on this design, we not only present a redesigned channel-level reconstruction loss but also introduce inter-channel contrastive learning (ICCL) to further explore the inter-channel interaction information. PSG-MAE aims not only to capture fine-grained temporal information but also to learn the potential relationships between multiple channels of PSG. After the pretraining phase of PSG-MAE, a robust PSG encoder is built, which can be combined with downstream networks and fine-tuned to adapt to different sleep event monitoring tasks. The contributions of this paper can be summarized as follows:

- We propose PSG-MAE, a novel pretraining framework for PSG signals, which employs a complementary-masking strategy and leverages unlabeled PSG data for self-supervised learning. This approach enhances the feature extraction process, which can be applied to a wide range of sleep event monitoring tasks.
- To better exploit the multichannel nature of PSG signals, we introduce an updated channel-level signal reconstruction loss and a novel ICCL method. These innovations improve PSG-MAE's ability to capture fine-grained temporal information from each channel while also effectively modeling inter-channel interactions.
- Combined with downstream networks, the pretrained PSG encoder demonstrates exceptional discriminative performance and robustness across multiple sleep event monitoring tasks, including sleep staging and sleep apnea detection, showcasing its broader applicability compared to traditional single-task models.

II. Related Work

The research on PSG-driven automated sleep event monitoring currently focuses primarily on addressing the issues of sleep staging and the identification of other sleep events. Research on automated sleep staging generally follows two approaches: using raw signals as input and applying time-frequency transformations (e.g., Fourier transform, continuous wavelet transform) to generate spectra as input [19]. Each channel of PSG signals is typically represented as one-dimensional time-series data. Consequently, feature extraction modules commonly employ models such as 1D convolutional neural networks (1D-CNNs) and recurrent neural networks (RNNs), which are well-suited for processing one-dimensional signals and capturing temporal information. Goshtasbi et al. introduced SleepFCN, a fully convolutional framework that utilizes dual 1D-CNN branches with distinct kernel

sizes to capture information across various EEG frequency bands [20]. Building upon SleepFCN, Zhu et al. developed MS-HNN, which integrates a squeeze-and-excitation block into the dual-kernel 1D-CNN branches to select more informative features. Additionally, MS-HNN employs bidirectional gated recurrent units (Bi-GRU) in the downstream network to learn temporal dependencies [21]. Na et al. proposed using convolutional layers to fuse multichannel PSG data, allowing various physiological signals to contribute to the decision-making process [22]. Physiological signals, such as EEG, exhibit distinct variations in frequency bands and power distributions across different sleep stages. Time-frequency transformations effectively capture these frequency characteristics, especially for applications involving non-stationary signals. Huy et al. introduced SeqSleepNet, which uses short-time Fourier transform to convert PSG signals into time-frequency spectrograms and applies recurrent layers to capture both short-term and long-term dependencies within each epoch [23]. Dai et al. proposed generating multichannel time-frequency spectrograms and employing multiple transformer groups to capture both individual channel features and joint features across channels [24].

Research on automated monitoring of other sleep events, such as sleep disorders, has advanced significantly. Zhao et al. segmented signals from the C3-A2 and C4-A1 EEG channels into five sub-bands, extracted entropy and variance features, and used machine learning to classify obstructive sleep apnea (OSA), central sleep apnea (CSA), and normal breathing events with [25]. Brink-Kjaer et al. used CNN+Bi-LSTM to extract features and temporal information from 5-minute PSG epochs for RBD classification, extending it with latent space transfer to analyze entire night recordings [26]. Qu et al. combined single-channel EEG data with a domain adaptation strategy, using similarity loss between encoders from source and target domains to learn temporal features. The source encoder was then integrated with LSTM networks for insomnia detection [27].

Sleep events are stage-dependent and often causally linked. For example, obstructive apneas and hypopneas occur more frequently during REM or N2/N3 sleep due to reduced muscle tone, while both central and obstructive apneas can induce arousals to restore ventilation. These interdependencies reveal the complexity of sleep physiology and support the need for automatic multitask sleep monitoring. MSleepNet proposes a semi-supervised multiview hybrid neural network that leverages both raw EEG and time–frequency representations, incorporates an attention-enhanced ResNet backbone, and employs a multitask classification loss to enable simultaneous detection of sleep arousal and sleep stage [28]. EEG-CLNet builds on a collaborative learning framework with a dual-branch convolutional neural network that jointly learns sleep stage classification and OSA event detection from a single-channel EEG, using inter-task feature fusion and mutual constraint losses to enhance performance across both tasks [29]. Similarly, MSED presents a multi-modal sleep event detection model that integrates EEG, EOG, EMG, and airflow signals through a hierarchical attention-based network, combining modality-specific feature extraction and cross-event joint learning to detect sleep stages, arousals, and respiratory events simultaneously [30].

Self-supervised learning has been shown to improve the robustness of feature extraction by leveraging unlabeled data. In this context, MAE learns robust feature representations by masking and reconstructing portions of the input signals, and has demonstrated superior performance in various downstream tasks [31], [32], [33], [34]. MAE has been applied to the representation learning of temporal physiological signals. Y.-T. Lan et al. proposed the Corrupted Emotion Autoencoder (CEMOAE) framework to address channel corruption in EEG topographic maps by reconstructing masked signals to learn robust features and fine-tuning a pretrained autoencoder for emotion recognition [35]. H. Ma et al. proposed a novel Region-State Masked Autoencoder (RS-MAE) that reduces redundancy in dynamic functional connectivity matrices, introduces region-state embeddings, and applies data augmentation to enhance classification performance for neuropsychiatric disorders based on resting-state fMRI. The encoder, pretrained in this manner, has been shown to improve downstream task performance by capturing more relevant features [36].

III. Methods

A. Overview

The general framework of PSG-MAE, based on the complementary masking strategy, is shown in Fig. 2. A 30-second multichannel PSG data segment, divided into temporally equal-length sub-segments, serves as the input. After that, the input is processed by applying a pair of randomly generated and complementary masks, forming two masked inputs that are fed into a shared encoder-decoder network to reconstruct the unmasked regions. The model then applies multichannel reconstruction and self-supervised ICCL to capture both temporal features and interactions among channels within the PSG input.

B. Multichannel Signal Reconstruction With Complementary-Masking

The input PSG data $X \in \mathbb{R}^{C\times L}$ has C channels, with each channel containing L time steps. According to the standards of the International Classification of Sleep Disorders (ICSD), X encompasses a 30-second window of PSG data, where $L = 30s \times$ sampling frequency. The input X is partitioned into smaller and manageable sub-segments. These sub-segments are defined by a hyperparameter N_{Patch}, which specifies how many patches the original time series data should be divided into along the time dimension. resulting in a set of sub-segments x_i. This process can be expressed as

$$X = [x_1, x_2, x_3, \ldots, x_N],\ x_i \in \mathbb{R}^{C\times L'},\ L' = \frac{L}{N_{Patch}}. \quad (1)$$

At the beginning of the pretraining phase, the PSG-MAE framework generates a pair of complementary masks, M and $(\mathbf{1} - M) \in \mathbb{R}^{C\times L}$, each having the same size as the original input. The masking process begins by randomly selecting the floor of half the number of channels ($\lfloor C/2 \rfloor$) from each sub-segment x_i, and then combining all selected channels to

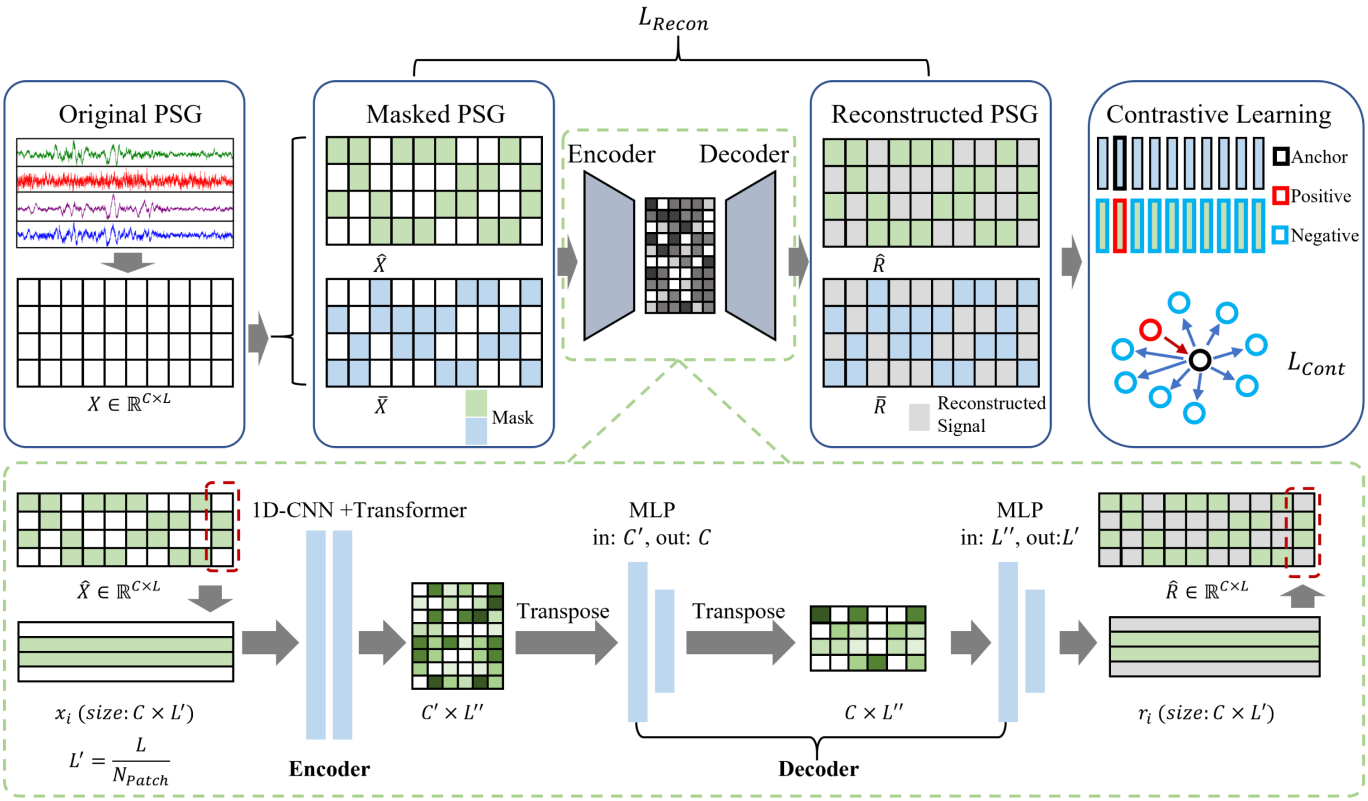

Fig. 2. The framework of **PSG-MAE**: The original PSG signal is divided into sub-segments along the time dimension, followed by the application of complementary masks across the channel dimension. After passing through the encoder-decoder network, the unmasked portions of the signal are reconstructed, with the channel-level reconstruction loss facilitating the learning of temporal features in the original signal. In the pair of reconstructed PSG signals, one sub-segment is treated as an anchor, whose corresponding sub-segment in the other signal is considered as a positive sample, while the remaining sub-segments are negative samples. ICCL is then applied to learn the intrinsic relationships between different channels by maximizing the distance of positive pairs and minimizing that of negative ones.

form the mask M, while the channels that are not selected form the complementary mask $(\mathbf{1} - M)$. In this way, the two masks are complementary across the channel positions. Once the two complementary masks are created, the input matrix X is masked by applying both M and $(\mathbf{1} - M)$ across all sub-segments, and a pair of inputs to the shared encoder is generated as

$$\hat{X} = M * X, \tag{2}$$

$$\bar{X} = (\mathbf{1} - M) * X. \tag{3}$$

The masked data $\hat{X}$ and $\bar{X}$, after undergoing the masking process, are passed through a transformer-based encoder for feature extraction and sequence modeling. This encoder leverages the self-attention mechanism to capture long-range dependencies within the data, enabling it to understand complex temporal patterns across multiple channels. After encoding, the transformed representation is fed into a multilayer perceptron (MLP)-based decoder. The decoder reconstructs the original signals from the encoded feature maps, producing a pair of reconstructed signals$\hat{R}$ and $\bar{R}$. These reconstructed signals are subsequently compared with the original signals to compute the redesigned channel-level reconstruction loss, formulated as

$$L_{Recon} = L_{COS} + L_{MSE}, \tag{4}$$

$$L_{COS} = \frac{1}{C}\sum_{c=1}^{C} L_{channelCOS_c}, \tag{5}$$

$$L_{channelCOS_c} = 1 - \frac{1}{N}\sum_{n=1}^{N} CosineSimilarity\left(r_n^c, x_n^c\right), \tag{6}$$

$$CosineSimilarity\left(r_n^c, x_n^c\right) = \frac{\sum_{t=1}^{N_{patch}} r_n^c(t)\, x_n^c(t)}{\|r_n^c\|\|x_n^c\|}, \tag{7}$$

$$L_{MSE} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}\left(x^c(t) - r^c(t)\right)^2, \tag{8}$$

where reconstruction loss is formulated as a combination of channel-level cosine similarity loss L_{COS} and channel-level mean squared error (MSE) loss L_{MSE}. The L_{COS} is computed by firstly evaluating the cosine similarity of the c-th channel in the reconstructed signal sub-segment r_n^c and the corresponding channel of the masked signal sub-segment x_n^c, n means the n-th segment, as expressed in (7). The resulting cosine similarities are then averaged across both the sub-segments and channels, as outlined in (6) and (5). The L_{MSE} is computed by first calculating the squared difference between the corresponding values of the masked signal $x^c(t)$ and the reconstructed signal $r^c(t)$ for each time step t along the channel c. The squared differences are then averaged over all T time steps and further averaged across all C channels, shown in (8). By integrating these two channel-level loss functions, the cosine similarity loss enforces the preservation of the overall pattern and trend of the reconstructed signals relative to the original data, while the MSE loss refines the relative magnitudes of the numerical

values. This synergy between the two losses ensures that the model captures both the structural integrity and the numerical accuracy of the signals, leading to a faithful reconstruction of the original data.

C. Introduction to ICCL

During sleep monitoring, the physiological data recorded by PSG is multi-source, encompassing signals collected from various sensors (such as EEG, EOG, respiratory airflow sensors, EMG, etc.) distributed across the body. These signal channels are typically synchronized in the time domain so that sleep events are often reflected simultaneously across multiple signal channels. This temporal alignment enables the multi-dimensional data to exhibit complementary information characteristics in multi-task sleep event detection. To fully exploit this synergy, PSG-MAE introduces a novel ICCL strategy, which allows the encoder to uncover the latent commonalities and differences between signals, thereby enhancing the collaborative representation of features across different channels.

Specifically, we apply contrastive learning to the two groups of reconstructed output sub-segments that form $\hat{R}$ and $\bar{R}$. The objective is to ensure that signal blocks from different channels, which correspond to the same time frame (i.e., originate from the same sub-segment), are drawn closer together in the feature space. In contrast, channel blocks from different time frames, which have weak correlations, are pushed further apart. Upon obtaining the reconstructed signals $\hat{R}$ and $\bar{R}$ from the shared decoder, we recursively select sub-segment $\hat{r}_i$ from$\hat{R}$ as anchor sample.Then the corresponding sub-segment $\bar{r}_i$ from $\bar{R}$, which contains complementary channel information relative to $\hat{r}_i$ within the same time interval, is designated as the positive sample. Meanwhile, the remaining sub-segments $\hat{r}_{j\neq i}$ from $\hat{R}$ serve as negative samples. To effectively capture the differences between positive and negative samples for local patch-level modeling, a triplet loss is employed during training to measure the relative distances among the anchor, positive, and negative samples. This strategy enables PSG-MAE to effectively learn and extract shared features across different channels while maintaining the independence and distinctiveness of temporal information. The channel contrastive loss L_{CL} is defined as

$$L_{\mathrm{CL}} = \frac{1}{N_{patch}} \sum_{i=0}^{N_{patch}} F_{MAX}, \tag{9}$$

$$F_{MAX} = \max\left(0, d(\hat{r}_i, \bar{r}_i) - \frac{1}{N_{patch}-1} \sum_{i\neq j} d(\hat{r}_i, \hat{r}_j) + \alpha\right), \tag{10}$$

$$d(x,y) = \sqrt{\sum_{k=1}^{D} (x_k - y_k)^2}, \tag{11}$$

where $d(x, y)$ denotes the Euclidean distance between samples x and y, x and y are two sample vectors with D dimensions. The components x_k and y_k represent the values of the samples in the k-th dimension. in (10), $d(x, y)$ measures the similarity between pairs of signal blocks, and α is a hyperparameter that defines the minimum margin between the positive and negative samples, ensuring that the negative samples are sufficiently far away from the anchor in the feature space.

The loss function for the pretraining framework for PSG signals based on a complementary-masking strategy for multichannel signal reconstruction is defined as

$$L = L_{\mathrm{Recon}} + L_{\mathrm{CL}}. \tag{12}$$

By jointly optimizing these two losses, PSG-MAE effectively preserves the fine-grained temporal details of PSG signals while also capturing the correlated features across different channels. The resulting encoded features provide a robust intermediate representation that can be leveraged for a wide array of downstream sleep-related tasks, including sleep stage classification, apnea detection, and recognition of other sleep events. Consequently, the pretrained encoder can be further fine-tuned for different sleep analysis applications, thereby enhancing model performance across diverse datasets and task-specific scenarios.

D. Downstream Multitask Sleep Events Monitoring

To effectively apply the pretrained PSG-MAE encoder for downstream sleep event monitoring tasks, we design a feature decomposing network as illustrated in Fig. 3. This downstream network comprises several key components aimed at extracting and refining task-relevant information. To maximize the utility of the pretrained features derived from PSG-MAE, we incorporated a multi-branch 1D-CNN architecture. This architecture utilizes filters of varying sizes (1×3, 1×5, and 1×7) to capture multi-scale temporal patterns present in the PSG signals. These extracted features are then concatenated, enabling the network to integrate complementary information from different receptive fields. Subsequently, we apply a dimensionality-reduction step using a 1×1 filter, followed by global pooling, to further distill the feature representation while retaining the most informative components relevant to the task. The final output is then passed through an MLP layer for discrimination, generating task-specific results. A crucial aspect of this approach is the involvement of the pretrained PSG-MAE encoder during the training phase of the downstream network. By integrating the encoder into the backpropagation process, the network can dynamically fine-tune its feature extraction capabilities to learn features that are specifically tailored to the downstream task. Rigorous validation of the effectiveness of the PSG-MAE encoder is conducted through the following experiments on sleep staging and apnea detection tasks.

We employ cross-entropy loss to optimize the models for both downstream tasks. In the sleep staging task, multi-class cross-entropy loss (13) is used, while binary cross-entropy loss (14) is applied in the apnea detection task; they are defined as

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(p_{ij}), \tag{13}$$

$$L_{bcls} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right], \tag{14}$$

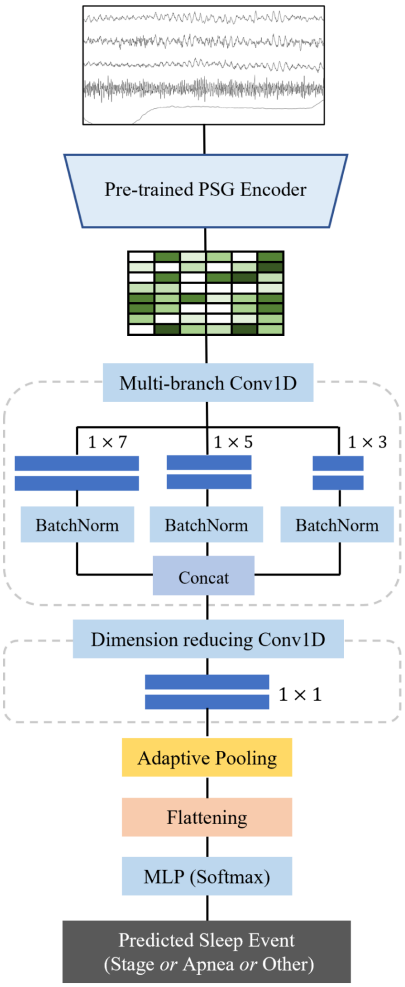

Fig. 3. Basic structure of downstream sleep events monitoring network.

where y is the manually annotated label of sample, p is the sample's predicted probability of each class. Given the class imbalance in both sleep staging and apnea detection tasks (e.g., sleep apnea events typically constitute a small proportion of the total sleep duration), we apply class weights to the cross-entropy loss function, with higher weights assigned to underrepresented classes to mitigate the impact of class imbalance during training.

IV. Experiments

The objective of this study is to achieve robust extraction of both single-channel temporal features and multichannel fusion features from PSG data by proposing the unsupervised learning approach, PSG-MAE. The experiments are designed with two main goals. First, to examine whether the reconstructed signals, facilitated by the complementary-masking and ICCL strategies, can faithfully preserve fine-grained temporal characteristics and capture intrinsic inter-channel dependencies in multichannel PSG data. Second, to evaluate the performance of the pretrained encoder across multiple datasets on two downstream sleep event monitoring tasks—sleep staging and apnea detection—thereby assessing its feasibility, discriminative performance and generalization capability across diverse data distributions.

A. Dataset

To ensure sufficient diversity in the PSG data during the pretraining phase of the PSG-MAE and enhance the encoder's robustness and generalization ability, this study uses four different datasets. These datasets not only help in improving the encoder's performance but also allow for the evaluation of its effectiveness in downstream sleep event monitoring tasks:

The **Sleep Heart Health Study (SHHS)** [37], [38] is a multicenter cohort dataset collected under the National Heart, Lung, and Blood Institute (NHLBI). It includes overnight PSG and other physiological measurements from approximately 6,400 adults. Data were collected in two visits: SHHS1 (1995–1998; 5,793 recordings) and SHHS2 (2001–2003; 2,651 recordings). After data cleaning to meet the channel requirements, SHHS1 retained 5,519 recordings, while SHHS2 retained all 2,651 recordings. Within the cohort, 47.2% of participants were male and 52.8% were female. The mean age was 64.55 ± 11.17 years, with a mean body mass index (BMI) of $28.21 \pm 5.79\,\mathrm{kg/m^2}$ and a mean apnea–hypopnea index (AHI) of 10.67. Based on AHI thresholds, 46.2% of participants exhibited no sleep apnea (AHI$<$ 5), 31.2% had mild sleep apnea (5 $\leq$AHI$<$ 15), 14.4% had moderate sleep apnea (15 $\leq$AHI$<$ 30), and 8.2% had severe sleep apnea (AHI$\geq$ 30) [39].

The **PSG-audio** dataset (Version 3) [40] is sourced from the Sismanoglio – Amalia Fleming General Hospital in Athens, Greece, and was collected and annotated by the hospital's medical team. It comprises 286 synchronized PSG recordings, accompanied by breathing sound recordings from both tracheal and ambient microphones, intended for the analysis of apneic events and the development of home-based apnea detection techniques. Participants ranged in age from 18 to 88 years (mean 52.3), with 61% males and 39% females, and exhibited a BMI of 30.7. The mean AHI was 25.4, with 19% of participants classified as having no sleep apnea, 24% with mild sleep apnea, 21% with moderate sleep apnea, and 36% with severe sleep apnea. The mean nadir oxygen saturation (SpO_2) was $79.2 \pm 7.5\%$.

The **Clinical-PSG** dataset used in this study was collected during clinical sleep tests conducted between 2021 and 2022 at the Department of Neurology, Guangzhou First People's Hospital, Guangdong, China. It comprises PSG data from 371 subjects (187 males and 130 females), recorded during nighttime sleep in a clinical laboratory setting. Professional sleep medicine physicians annotated sleep events in 30-second epochs, including sleep stages, apneas, periodic limb movements during sleep (PLMD), and periodic limb movements while awake (PLMA). The dataset supports sleep disorder diagnosis and research and was approved by the hospital's Ethics Committee (Approval No. K-2025-067-01), with no private information collected. It serves to enrich the diversity of the PSG-MAE encoder's pretraining and provides real clinical data for its downstream task validation. Demographic statistics for the subjects include mean values for AHI (25.16), BMI ($27.40\,\mathrm{kg/m^2}$), Age (57.95), and minimum SpO_2 (83.09%), with detailed distributions presented

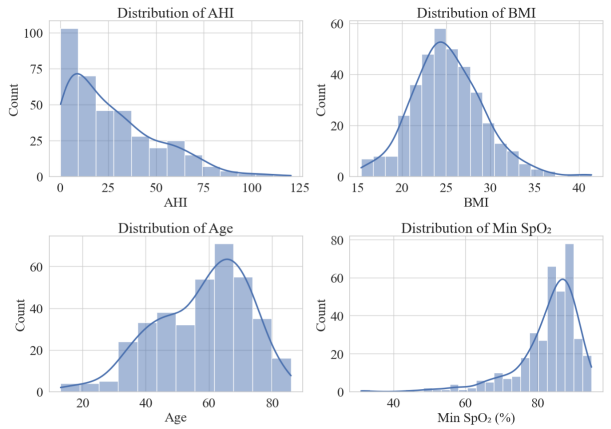

Fig. 4. Statistical distributions of the Clinical-PSG dataset.

in Fig. 4. Based on AHI thresholds, the cohort comprised 17.1% without sleep apnea, 21.4% with mild sleep apnea, 23.8% with moderate sleep apnea, and 37.7% with severe sleep apnea.

The **Multi-Ethnic Study of Atherosclerosis Sleep (MESA)** [37], [41] dataset is a large-scale, multi-ethnic ancillary study of sleep conducted during MESA Exam 5 (2010–2013). It comprises in-home overnight polysomnography (PSG), wrist actigraphy, and standardized sleep questionnaires from 2,237 participants, with raw PSG data available for 2,056 individuals. The cohort had a mean age of 68 years, comprised 48% males and 52% females, and exhibited a mean BMI of 28.0 kg/m^2. The median AHI was 14.0, 22% of participants had no sleep apnea, 32% had mild sleep apnea, 27% had moderate sleep apnea, and 19% had severe sleep apnea [42]. In this study, the MESA dataset was not utilized for pretraining the PSG-MAE encoder but was instead employed for cross-dataset fine-tuning the trained sleep monitoring models on downstream tasks to assess the generalization capability of the proposed method.

B. Experimental Setup

Table I summarizes the selected channels and the data volume used for pretraining and downstream tasks in each dataset. During the pretraining phase, PSG-MAE was trained on a server equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory, using a batch size of 64 and a learning rate of 10^{-4}. The synthetic pretraining dataset comprises 7327 records of PSG data from the SHHS (6670 records are uniformly sampled), PSG-audio, and Clinical-PSG datasets. Combining these datasets ensures that the encoder learns the distribution of physiological signals across different datasets and subjects, enhancing the robustness of feature extraction. To maintain consistency in channel selection across the pretraining data, five common channels are chosen from the datasets: right central brain activity, left central brain activity, left eye movement, chin movement, and pressure-based airflow signal. This choice aligns with our two downstream objectives: EEG/EOG/EMG form the standard triad for sleep staging, while AIRFLOW provides direct respiratory evidence for apnea detection. The PSG data undergoes EEG artifact removal processing to reduce interference from non-brain signals. The sampling frequency of the PSG data is set to 100 Hz, therefore, the shape of the processed sleep epochs, (*channel number*, *data length*), is standardized to $(5, 3000)$. To mitigate individual differences in the PSG data, we apply the Z-score normalization method to each of the five channels in the PSG signals. The formula is

$$x' = \frac{x - \mu}{\sigma}. \tag{15}$$

This approach involves calculating the median (μ) and standard deviation (σ) for each channel, and then using these values to normalize the data accordingly. The training data is randomly shuffled, with 80% allocated for training, 10% used for validation, and the remaining 10% used for testing. Training proceeded for 200 epochs with validation every 10 epochs; upon completion, the best-performing checkpoint was selected for testing and subsequently used as the encoder for the downstream tasks.

In downstream evaluation, we assess PSG-MAE on sleep staging and apnea detection. Experiments are conducted on servers with NVIDIA GeForce RTX 3090/4090 GPUs (24 GB), using a batch size of 256 and a learning rate of 10^{-4}. For SHHS, we use 1,000 records from SHHS1 and 500 from SHHS2 (all excluded from pretraining), and we include all annotated records from the Clinical-PSG dataset; the channel configuration matches pretraining. Both the in-dataset training (on SHHS and Clinical-PSG) and the cross-dataset fine-tuning (on MESA) follow the same subject-wise 5-fold cross-validation protocol: in each fold, 10% of the training subjects are held out for validation. The training schedules differ as follows. For in-dataset training, we train for 100 epochs, validate every 5 epochs, and evaluate on the test fold using the checkpoint with the best validation performance. For adaptation to MESA, we initialize from the best SHHS checkpoint and fine-tune for 10 epochs with validation every

TABLE I
SELECTED PSG CHANNELS AND SAMPLE ALLOCATION FOR PRETRAINING AND DOWNSTREAM TASKS

Dataset	Pretrain	Downstream	PSG Channels
SHHS1	4519	1000	'EEG', 'EEG2', 'EOG (L)', 'EMG' and 'AIRFLOW'
SHHS2	2151	500	
PSG-audio	286	0	'EEG C3-A2', 'EEG C4-A1', 'EOG LOC-A2', 'EMG Chin' and 'Flow Patient (Pressure cannula)'
Clinical-PSG	371	371	'EEG C3-REF', 'EEG C4-REF', 'EOG LOC', 'EMG Chin' and 'Airflow'
MESA	0	2056	'EEG1', 'EEG3', 'EOG-L', 'EMG' and 'Pres'
Overall	7327	3927	

"Pretrain" and "Downstream" refer to the number of samples used for pretraining and downstream task training, respectively.

2 epochs; in each fold, the best validation checkpoint is used for testing, and results are averaged across the five folds. Sleep staging is formulated as a five-class problem over W, N1, N2, N3, and R. Apnea detection is binary, grouping epochs annotated as obstructive apnea, central apnea, or hypopnea into *Apnea*, with all remaining non-W epochs assigned to *Normal*.

We include recently released open-source models for sleep staging and apnea detection as baselines to compare with our proposed approach. For the sleep staging task, SleepTransformer [43], Cross-Modal Transformers [44], MHFNet [45] and MAESleepNet [46] are evaluated under our downstream training pipeline. For the first three models, we use the authors' reported optimal configurations by selecting the required signal channels from the downstream training records (keeping the original sampling rates unchanged to match each model's optimal setting) and then train and fine-tune them under the same cross-validation protocol for comparison with our model. MAESleepNet is first pretrained in a self-supervised manner on the 329 SHHS PSG recordings with regular sleep patterns (Apnea–Hypopnea Index < 5) as described in its original paper, and is subsequently trained on the downstream tasks. For the apnea detection task, we bring the outputs of SE-MSCNN [47] and BAFNet [48] to a 30-second epoch resolution and proportionally adjust their windowing parameters; we then train them on the downstream-task datasets to ensure their outputs are directly comparable to those of our model.

C. Evaluation Metrics

To rigorously evaluate the effectiveness of the pretraining phase and the performance of downstream sleep events monitoring, this study first employs the MSE to quantify the discrepancy between the reconstructed and original signals, thereby assessing the accuracy of the signal reconstruction. The MSE is written as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{pred}} - y_{\text{true}})^2. \tag{16}$$

For the 5-class sleep staging task, performance is evaluated using accuracy (ACC), macro-averaged F1-score (MF1), and the F1-score for each class. In this context, C denotes the total number of sleep stages, N the total number of epochs, and (TP_c, FP_c, TN_c, FN_c) represent the confusion matrix entries for stage $c \in 1, \ldots, C$, then

$$\text{ACC} = \frac{\sum_{c=1}^{C} TP_c}{N}, \tag{17}$$

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{18}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{19}$$

$$F1_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{20}$$

$$\text{MF1} = \frac{1}{C}\sum_{c=1}^{C} F1_c. \tag{21}$$

As in the apnea detection task (binary), apnea events are relatively rare and of short duration, which makes accuracy alone potentially misleading; therefore, performance is evaluated using ACC, MF1, sensitivity, and specificity. Here, (TP, FP, TN, FN) denotes the confusion matrix entries for the normal and apnea classes, then

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{22}$$

$$\text{MF1} = \frac{1}{2}\left(\frac{2TP}{2TP + FP + FN} + \frac{2TN}{2TN + FN + FP}\right), \tag{23}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{24}$$

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{25}$$

V. Experimental Results & Discussion

A. Results of PSG-MAE Pretraining Process

The Fig 5. shows the original PSG signal input and the reconstructed PSG signal output from PSG-MAE during the validation of the pretraining phase. A comparison between (a) and (b) shows that the reconstructed signal accurately replicates the trend of signal variations while preserving the temporal details of the original signal. Furthermore, in the highlighted dashed boxes, fluctuations of multiple signal channels in the original signal are aligned with the time scale of the reconstructed signal, indicating that PSG-MAE is capable of capturing underlying relationships among different channels. Through ICCL, PSG-MAE retains the original signal's trend while demonstrating a certain degree of noise suppression, which indicates the interaction content learned by the encoder can help repair corrupted signal channels. The comparison between (b) and (c) underscores the critical role of ICCL in preserving single-channel information within multi-channel fused features. Without ICCL, PSG-MAE struggles to accurately reconstruct signals due to the loss of channel-specific details. This highlights ICCL's ability to enhance the fusion of multi-channel information during the pretraining phase while minimizing information loss from individual channels. Table II presents the MSE of the selected 5 channels from the reconstructed and original signals, quantitatively reflecting the accuracy of PSG-MAE in signal reconstruction. EEG and EOG signal reconstruction demonstrates good performance, with effective preservation of waveform details and favorable MSE values. Similarly, the PSG-MAE model without ICCL cannot achieve the same level of reconstruction performance. Despite the high noise levels in the original EMG signals, the PSG-MAE model is still able to recover the main trends of the signal. In contrast, airflow signals show greater variability due to the significant time gap between the datasets and variations in data acquisition conditions, which complicate the reconstruction process. The reconstruction performance of some airflow signals is suboptimal, resulting in relatively higher average MSE values.

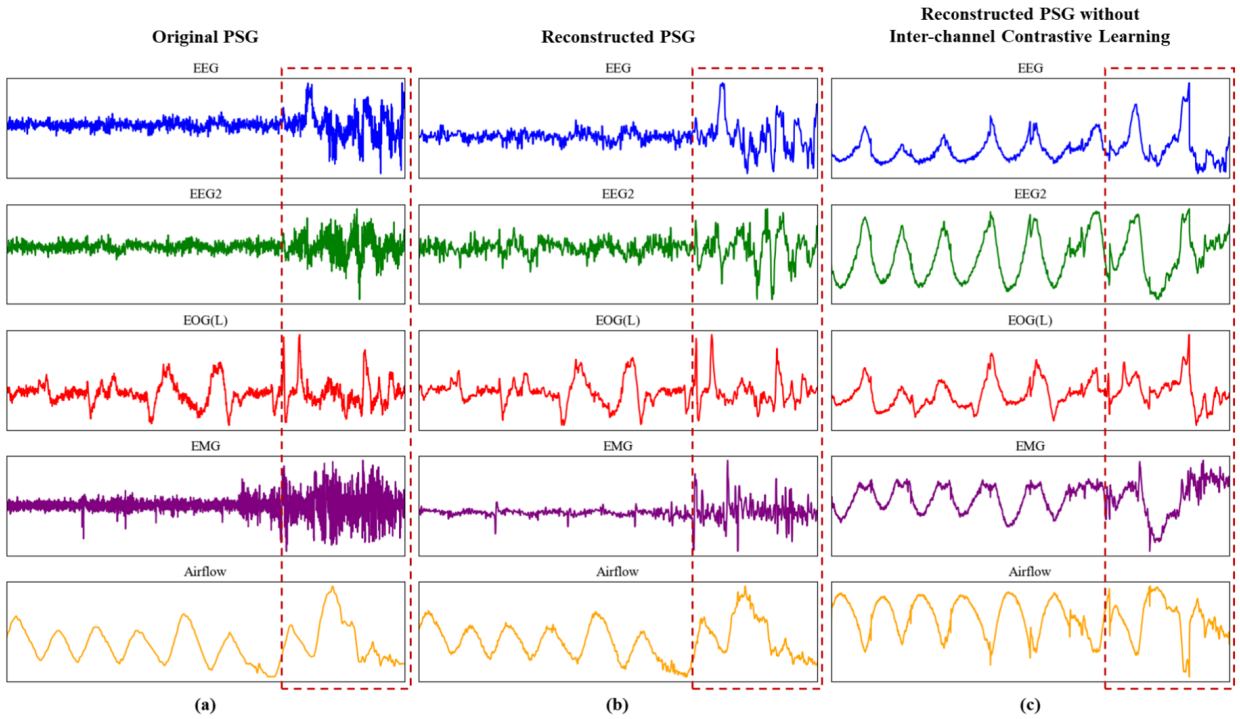

Fig. 5. The signal reconstruction results of PSG-MAE pretraining. Comparison shows that the channel-level signal reconstruction loss and ICCL enable the framework to learn fine-grained temporal information within PSG channels as well as interaction information between channels. In contrast, without ICCL, it becomes difficult to disentangle individual channel information from the fused multichannel features.

TABLE II
THE MSE VALUE OF PSG CHANNELS BETWEEN THE ORIGINAL SIGNAL AND THE RECONSTRUCTED SIGNAL OF PSG-MAE

PSG Channels	EEG	EEG2	EOG (L)	EMG	Airflow
MSE without ICCL	2.34×10^{-2}	1.73×10^{-2}	8.27×10^{-2}	1.39×10^{-2}	1.23
MSE with ICCL	$\mathbf{7.83 \times 10^{-6}}$	$\mathbf{6.55 \times 10^{-6}}$	$\mathbf{3.12 \times 10^{-5}}$	$\mathbf{5.88 \times 10^{-6}}$	$\mathbf{6.26 \times 10^{-2}}$

B. Feature Distribution Changes in Downstream Training

Fig. 6 presents the uniform manifold approximation and projection (UMAP) visualization of features extracted by the pretrained PSG-MAE encoder, where each point corresponds to the reduced feature representation of a single sleep epoch. Subfigures (a)–(b) and (c)–(d) show the changes in feature distribution after the encoder is fine-tuned on the downstream sleep staging and apnea detection tasks, respectively. Before fine-tuning, the extracted features exhibit only subtle clustering. After task-specific training, however, the features form more distinct clusters—corresponding to different sleep stages in the sleep staging task and to normal versus apnea epochs in the apnea detection task—indicating that the encoder has learned task-relevant representations better suited to the feature extraction requirements of both tasks.

C. Sleep Staging Performance

We evaluated sleep staging models composed of a pretrained encoder and a downstream classification network on the SHHS and Clinical-PSG datasets. Table III compares the performance of our approach with that of other existing methods. In the table, *PSG-MAE(P)*, *PSG-MAE(U)*, and *PSG-MAE(F)* refer to a pretrained PSG-MAE encoder, an untrained PSG-MAE encoder, and a pretrained PSG-MAE encoder kept frozen during downstream training, respectively, while *NET* denotes the downstream classification network.

In the 1,500 SHHS downstream training records, compared with recently proposed methods, the combination of *PSG-MAE(P) + NET* outperforms these models, achieving ACC of 89.2% and MF1 of 81.0%, while maintaining superior discriminative ability for N2 (93.6%) and N3 (88.2%). On the Clinical-PSG dataset, which more closely reflects real-world clinical sleep monitoring conditions, *PSG-MAE(P) + NET* consistently outperforms conventional deep classification networks across all metrics. Comparative experiments with *PSG-MAE(U)* and *PSG-MAE(F)* further reveal that, for this relatively small dataset, an untrained encoder tends to overfit and thus generalizes poorly, while a frozen pretrained encoder, despite stronger feature extraction capability, fails to adapt to the specific data distribution without fine-tuning, resulting in suboptimal performance.

To further assess the generalization capability of PSG-MAE to unseen datasets, we perform cross-dataset fine-tuning by initializing from the best-performing *PSG-MAE(P) + NET* weights obtained on SHHS and then fine-tuning on the MESA dataset. The UMAP visualization of features extracted by the fine-tuned encoder (*PSG-MAE(FT)*) in Fig. 6 (e) shows a clear separation among sleep stages, while Table III indicates that

TABLE III
COMPARISON OF PSG-MAE VARIANTS AND EXISTING MODELS FOR THE SLEEP STAGING TASK

Dataset	Model	Overall Metrics		Per-class F1(%)				
		ACC(%)	MF1(%)	W	N1	N2	N3	R
SHHS	SleepTransformer(2022)	84.3	76.6	91.8	43.2	87.8	75.6	84.9
	MAESleepNet(2024)	83.9	74.3	89.6	35.3	88.3	75.0	83.3
	Cross-Modal Transformers(2024)	86.0	78.3	**93.8**	42.9	87.3	80.1	**87.3**
	MHFNet(2025)	86.5	79.4	92.3	**50.5**	88.8	80.9	84.3
	PSG-MAE(P) + NET	**89.2**	**81.0**	91.7	45.7	**93.6**	**88.2**	85.7
Clinical-PSG	CNN-LSTM	78.1	65.2	86.7	53.6	85.1	35.4	65.4
	CNN-Transformer	79.9	68.3	88.3	55.6	86.3	40.6	70.9
	PSG-MAE(U) + NET	69.9	45.4	79.3	14.5	78.9	6.9	47.2
	PSG-MAE(F) + NET	74.1	63.2	85.1	49.5	81.8	32.8	66.9
	PSG-MAE(P) + NET	**83.3**	**73.8**	**92.8**	**61.9**	**86.6**	**51.8**	**75.7**
MESA	SleepTransformer(FT)	75.4	68.0	84.9	31.9	79.8	71.7	71.9
	MAESleepNet(FT)	75.6	66.6	84.6	31.9	81.7	65.4	69.6
	Cross-Modal Transformers(FT)	76.4	69.0	86.3	33.1	79.9	72.1	73.4
	MHFNet(FT)	80.7	71.6	90.7	39.5	85.6	69.3	72.8
	PSG-MAE(FT) + NET(FT)	**84.9**	**77.9**	**92.4**	**49.5**	**88.4**	**80.7**	**78.8**

NOTE: **P**: Pretrained encoder. **U**: Untrained encoder (random initialization). **F**: Fully pretrained encoder but frozen during downstream training. **FT**: Cross-dataset fine-tuned encoder. **NET**: Downstream network. **NET(FT)**: Cross-dataset fine-tuned downstream network. **"+"** denotes a combination of the encoder variant and the downstream network.

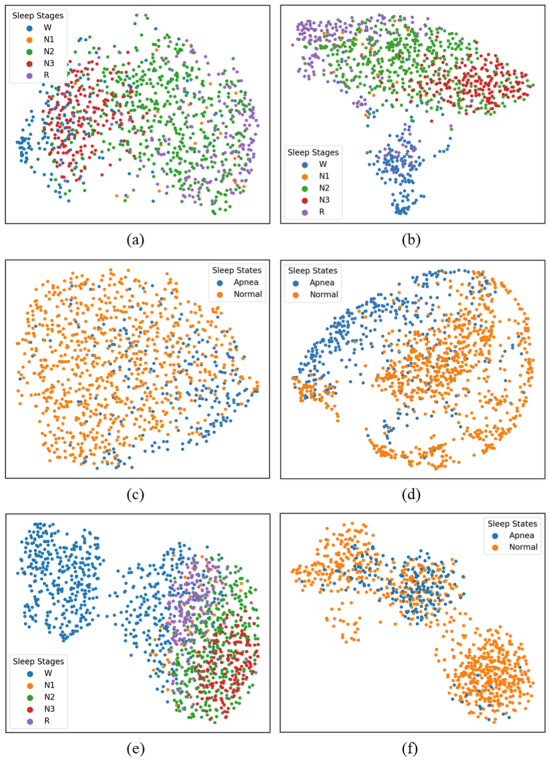

Fig. 6. UMAP visualization of PSG-MAE feature distributions of sleep staging and apnea detection tasks. (a)–(b) and (c)–(d) show distributions of features on the SHHS dataset before and after downstream-task training (within each pair, left: before; right: after). (e) and (f) show the feature distributions after cross-dataset fine-tuning the model on the MESA dataset.

comparison models trained only on a single dataset transfer poorly when fine-tuned to a new dataset, their feature extraction and decision boundaries do not carry over effectively. In contrast, *PSG-MAE(FT) + NET(FT)* achieves strong performance on MESA, with ACC of 84.9% and MF1 of 77.9%, demonstrating that pretraining PSG-MAE on multiple datasets confers robustness and cross-distribution generalization in the sleep staging task.

Finally, the aggregated confusion matrices in Fig. 7 (a)–(c) highlight the strong feature extraction ability of the pretrained PSG-MAE encoder in the sleep staging task. They also reveal that stage N1 is consistently the most challenging to classify across all three datasets. The per-class F1 for N1 is markedly lower than for other stages. This aligns with the UMAP patterns, where the N1 cluster shows the greatest overlap with its neighbors, indicating less separability in the learned feature space. We attribute the performance gap primarily to the inherent ambiguity of N1 and its scarcity in the data. As a transitional state from wakefulness to light sleep, N1 exhibits mixed or weakly expressed spectral signatures (attenuated α, emerging θ), subtle EOG activity, and reduced EMG, which blur boundaries with W and N2. Meanwhile, N1 is relatively rare, about 6–9% of epochs across datasets versus 35–45% for N2, amplifying majority-class bias and yielding under-calibrated decision thresholds.

D. Apnea Detection Performance

We evaluated apnea detection models on the SHHS, Clinical-PSG, and MESA datasets, with the results summarized in Table IV.

On the SHHS dataset, *PSG-MAE(P) + NET* demonstrates superior performance over previously reported methods across most evaluation metrics, achieving notable gains in MF1 (82.6%) and sensitivity (93.8%) while maintaining competitive specificity (88.9%), indicating its strong capability in

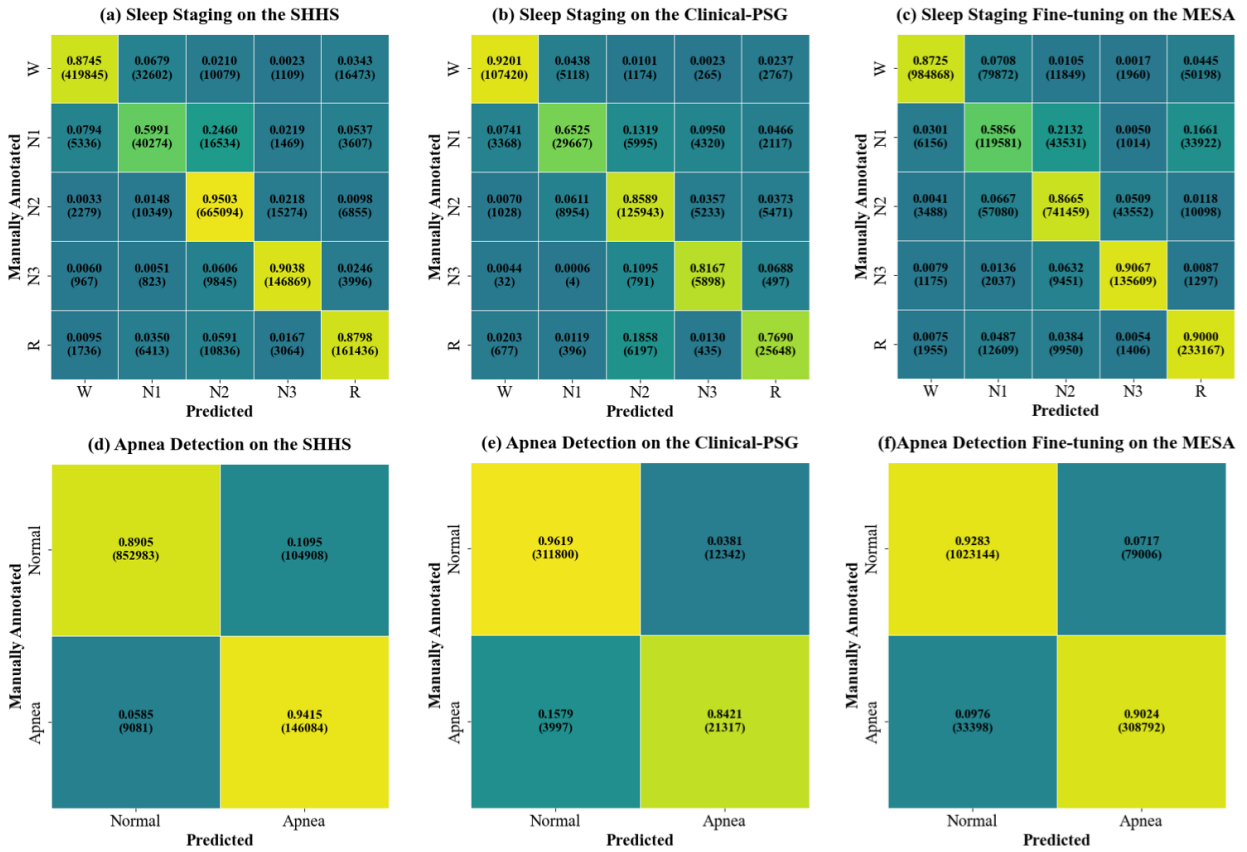

Fig. 7. Aggregated confusion matrices for downstream sleep event monitoring (decimals denote row-normalized percentages; numbers in parentheses are raw counts).

TABLE IV
COMPARISON OF PSG-MAE VARIANTS AND EXISTING MODELS FOR THE APNEA DETECTION TASK

Dataset	Model	ACC(%)	MF1(%)	Sensitivity(%)	Specificity(%)
SHHS	SE-MSCNN (2022)	86.8	77.1	77.7	88.3
	BAFNet (2023)	88.1	79.1	80.3	89.4
	PSG-MAE(P) + NET	**89.6**	**82.6**	**93.8**	**88.9**
Clinical-PSG	CNN-LSTM	91.6	60.1	19.2	97.2
	CNN-Transformer	91.3	67.3	38.9	95.4
	PSG-MAE(U) + NET	91.7	52.4	5.8	98.4
	PSG-MAE(F) + NET	89.4	72.8	78.4	90.2
	PSG-MAE(P) + NET	**94.7**	**83.1**	**82.2**	**95.7**
MESA	SE-MSCNN (FT)	80.7	72.2	53.6	89.2
	BAFNet (FT)	83.5	76.7	62.3	90.1
	PSG-MAE(FT) + NET(FT)	**92.2**	**89.6**	**90.1**	**92.8**

accurately detecting sleep apnea events. On the Clinical-PSG dataset, *PSG-MAE(P) + NET* consistently attains the highest scores across all metrics. Ablation experiments with *PSG-MAE(U)* and *PSG-MAE(F)* further reveal that, for this relatively small dataset, the lack of pretraining results in limited representational capacity and increased susceptibility to overfitting, whereas constraining a pretrained encoder without fine-tuning hampers its ability to adapt to the target domain, thereby limiting overall performance.

To examine the adaptability of PSG-MAE in the context of apnea detection, we also transferred the best-performing *PSG-MAE(P) + NET* weights obtained from SHHS training to the MESA dataset for fine-tuning. As shown in Fig. 6 (f), the UMAP visualization of features extracted by the fine-tuned encoder (*PSG-MAE(FT)*) exhibits a clear clustering of apnea epochs, indicating that the learned representations have effectively adapted to the new domain. In line with this observation, Table IV reports superior performance across all evaluation metrics, with *PSG-MAE(FT) + NET(FT)* achieving high ACC (92.2%), MF1 (89.6%), and balanced sensitivity–specificity (90.1%–92.8%) on the previously unseen dataset. In contrast, models trained only on a single dataset generalize poorly,

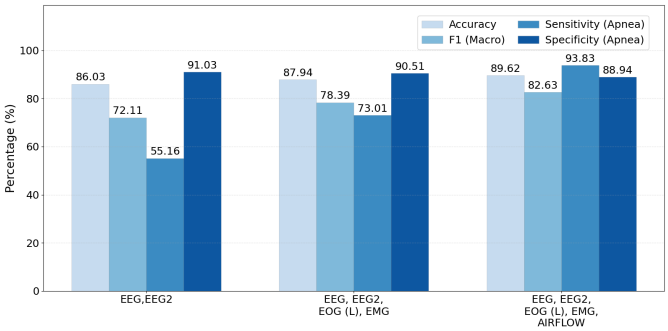

Fig. 8. Apnea detection performance on SHHS in the PSG-MAE framework across different channel combinations.

tending to misclassify apnea epochs and thereby yielding lower sensitivity for apnea. The aggregated confusion matrices for the apnea detection task across the three datasets are shown in Fig. 7(d)–(f), from which it can be observed that PSG-MAE exhibits robust performance in monitoring sleep apnea events.

To substantiate the benefit of PSG-MAE's multi-channel fusion for sleep event detection, we pretrain with different channel combinations and evaluate them on sleep apnea detection. The results are shown in Fig. 8. From the bar chart, it is evident that EEG alone is insufficient to accurately detect apneic epochs: EEG primarily captures sleep background rhythms, yielding high specificity but limited sensitivity to respiratory events. Adding EOG (L) and EMG introduces eye-movement and muscle-tone cues that help capture respiration-related micro-arousals and stage transitions, raising sensitivity to 73.01% and MF1 to 78.39%. With the AIRFLOW channel, which directly reflects airflow restriction/interruption, sensitivity further increases to 93.83% while specificity remains high (88.94%), leading to better overall performance (ACC 89.62%, MF1 82.63%). Overall, channels more closely aligned with respiratory physiology provide larger gains, and multimodal fusion markedly improves sleep apnea detection.

E. Discussion

The experimental findings substantiate the effectiveness of the PSG-MAE pretraining strategy and its adaptability to downstream sleep event monitoring tasks. By applying a pair of complementary masks to multiple channels within PSG segments from the same temporal sub-segment, the proposed channel-level signal reconstruction loss compels the encoder to extract temporal details from multi-channel inputs during reconstruction. Furthermore, complementary channels participate in ICCL with channels from different temporal segments, enabling the encoder to capture inter-channel dependencies at a given time point. Empirically, ICCL proves particularly beneficial for noisy channels, as it refines reconstructed signal trends and ensures that multi-channel reconstructions preserve both channel-specific characteristics and temporal alignment. In addition, channel ablations indicate that our method achieves better sleep event monitoring by effectively fusing information from multiple channels.

When fine-tuned with the downstream sleep event monitoring network, the pretrained encoder achieves state-of-the-art or competitive performance in both sleep staging and apnea detection. On SHHS, *PSG-MAE(P) + NET* matches or exceeds recent methods, exhibiting superior discriminative capability for N2/N3 stages and robust apnea event detection; on the clinically representative Clinical-PSG dataset, it consistently outperforms conventional deep architectures across all metrics, while ablations reveal the necessity of pretraining and fine-tuning for optimal adaptation. Cross-dataset fine-tuning from SHHS to MESA further demonstrates strong transferability: learned representations form well-structured clusters in UMAP for both staging and apnea, whereas single-dataset baselines transfer poorly. These results collectively indicate that PSG-MAE is not only effective for the evaluated tasks but also holds promise as a generalizable foundation for multi-dimensional, comprehensive sleep assessment systems.

VI. Conclusion

We presented *PSG-MAE*, a MAE–based pretraining framework that leverages unlabeled PSG data via self-supervised learning to enhance feature extraction for automated sleep event monitoring. The framework combines channel-level signal reconstruction loss for detailed temporal modeling with ICCL to capture cross-channel dependencies. Pretraining on multi-dataset PSG data with a unified channel configuration improves robustness to diverse data distributions, while fine-tuning with task-specific networks enables effective downstream task learning, including sleep staging and apnea detection.

Comprehensive evaluations show that PSG-MAE learns robust, transferable representations, achieving competitive or superior results compared with existing methods and demonstrating strong generalization to unseen datasets. Future work will focus on extending the framework to support a broader range of sleep event monitoring tasks. Ultimately, PSG-MAE aims to provide a robust and generalizable foundation for multidimensional PSG analysis from a single input PSG segment.

References

[1] K. Ramar et al., "Sleep is essential to health: An American academy of sleep medicine position statement," *J. Clin. Sleep Med.*, vol. 17, no. 10, pp. 2115–2119, Oct. 2021.

[2] M. Lee et al., "Automatic sleep stage classification using nasal pressure decoding based on a multi-kernel convolutional BiLSTM network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 2533–2544, 2024.

[3] M. Yazdi, M. Samaee, and D. Massicotte, "A review on automated sleep study," *Ann. Biomed. Eng.*, vol. 52, no. 6, pp. 1463–1491, Jun. 2024.

[4] J. V. Rundo and R. Downey, "Chapter 25-polysomnography," in *Clinical Neurophysiology: Basis and Technical Aspects* (Handbook of Clinical Neurology), vol. 160, K. H. Levin and P. Chauvel, Eds., Amsterdam, The Netherlands: Elsevier, 2019, ch. 2, pp. 381–392. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444640321000254

[5] H. Zhou, A. Liu, S. Ding, J. Yao, and X. Chen, "An interpretable single-channel EEG sleep staging model based on prototype matching and multitask learning," *IEEE Sensors J.*, vol. 25, no. 2, pp. 3782–3793, Jan. 2025.

[6] H. Zhang, X. Wang, H. Li, S. Mehendale, and Y. Guan, "Auto-annotating sleep stages based on polysomnographic data," *Patterns*, vol. 3, no. 1, Jan. 2022, Art. no. 100371.

[7] I. Perez-Pozuelo et al., "The future of sleep health: A data-driven revolution in sleep science and medicine," *NPJ Digit. Med.*, vol. 3, no. 1, p. 42, Mar. 2020.

[8] E. Khalili and B. Mohammadzadeh Asl, "Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG," *Comput. Methods Programs Biomed.*, vol. 204, Jun. 2021, Art. no. 106063.

[9] R. Berry et al., *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications: Version 2.3*. Darien, IL, USA: American Academy of Sleep Medicine, 2015. [Online]. Available: https://books.google.com/books?id=SySXAQAACAAJ

[10] J. Somanna, D. Joshi, H. Gundu, and G. Srinivasa, "Automated classification of sleep apnea and hypopnea on polysomnography data," in *Proc. 12th Biomed. Eng. Int. Conf. (BMEiCON)*, 2019, pp. 1–5.

[11] A. De and E. Priya, "Sleep apnea sub-type detection from polysomnography signals," in *Proc. IEEE Int. Conf. Interdiscipl. Approaches Technol. Manage. Social Innov. (IATMSI)*, Mar. 2024, pp. 1–6.

[12] S. K. Satapathy, S. Thakkar, A. Patel, and D. Patel, "A machine learning-based models for intelligent automated sleep staging classification system using polysomnography data," in *Proc. IEEE 11th Region 10 Humanitarian Technol. Conf. (R10-HTC)*, Oct. 2023, pp. 267–272.

[13] A. Procházka, J. Kuchyňka, M. Yadollahi, C. P. S. Araujo, and O. Vyšata, "Adaptive segmentation of multimodal polysomnography data for sleep stages detection," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, 2017, pp. 1–4.

[14] D. Zhang, Y. She, J. Sun, X. Yang, X. Zeng, and W. Qin, "SwinSleep: A deep learning framework advancing overnight sleep staging toward clinical practice," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.

[15] X. Li, A. Al-Ani, and S. H. Ling, "Feature selection for the detection of sleep apnea using multi-bio signals from overnight polysomnography," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1444–1447.

[16] A. Bartolo, B. D. Clymer, R. C. Burgess, J. P. Turnbull, J. A. Golish, and M. C. Perry, "An arrhythmia detector and heart rate estimator for overnight polysomnography studies," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 5, pp. 513–521, May 2001.

[17] F. Ehrlich et al., "State-of-the-art sleep arousal detection evaluated on a comprehensive clinical dataset," *Sci. Rep.*, vol. 14, no. 1, p. 16239, Jul. 2024.

[18] H. Lee et al., "A large collection of real-world pediatric sleep studies," *Scientific Data*, vol. 9, no. 1, p. 421, Jul. 2022.

[19] R. N. Sekkal, F. Bereksi-Reguig, D. Ruiz-Fernandez, N. Dib, and S. Sekkal, "Automatic sleep stage classification: From classical machine learning methods to deep learning," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103751.

[20] N. Goshtasbi, R. Boostani, and S. Sanei, "SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2088–2096, 2022.

[21] H. Zhu et al., "MS-HNN: Multi-scale hierarchical neural network with squeeze and excitation block for neonatal sleep staging using a single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2195–2204, 2023.

[22] Y. Na, D. Kim, D. Kim, and J. Lee, "Evaluation of OSA patient sleep stage classification performance using a multi-channel PSG dataset," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2022, pp. 1–4.

[23] P. Huy, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019.

[24] Y. Dai et al., "MultiChannelSleepNet: A transformer-based model for automatic sleep stage classification with PSG," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 9, pp. 4204–4215, Sep. 2023.

[25] X. Zhao et al., "Classification of sleep apnea based on EEG sub-band signal characteristics," *Sci. Rep.*, vol. 11, no. 1, p. 5824, Mar. 2021.

[26] A. Brink-Kjaer, K. M. Gunter, E. Mignot, E. During, P. Jennum, and H. B. D. Sorensen, "End-to-end deep learning of polysomnograms for classification of REM sleep behavior disorder," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 2941–2944.

[27] W. Qu et al., "Single-channel EEG based insomnia detection with domain adaptation," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104989.

[28] H. Liu, H. Zhang, B. Li, X. Yu, Y. Zhang, and T. Penzel, "MSleepNet: A semi-supervision-based multiview hybrid neural network for simultaneous sleep arousal and sleep stage detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–9, 2024.

[29] L. Cheng, S. Luo, X. Yu, H. Ghayvat, H. Zhang, and Y. Zhang, "EEG-CLNet: Collaborative learning for simultaneous measurement of sleep stages and OSA events based on single EEG signal," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.

[30] A. N. Zahid, P. Jennum, E. Mignot, and H. B. D. Sorensen, "MSED: A multi-modal sleep event detection model for clinical sleep analysis," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2508–2518, Sep. 2023.

[31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[32] L. Wang et al., "VideoMAE v2: Scaling video masked autoencoders with dual masking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14549–14560.

[33] Z. Cai et al., "MARLIN: Masked autoencoder for facial video representation LearnINg," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2023, pp. 1493–1504.

[34] Z. Zhang, P. Zhao, E. Park, and J. Yang, "MART: Masked affective representation learning via masked temporal distribution distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 12830–12840.

[35] Y.-T. Lan, W.-B. Jiang, W. Zheng, and B. Lu, "CEMOAE: A dynamic autoencoder with masked channel modeling for robust EEG-based emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2024, pp. 1871–1875.

[36] H. Ma, Y. Xu, and L. Tian, "RS-MAE: Region-state masked autoencoder for neuropsychiatric disorder classifications based on resting-state fMRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 6, pp. 10707–10720, Jun. 2025.

[37] G.-Q. Zhang et al., "The national sleep research resource: Towards a sleep data commons," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018.

[38] S. F. Quan et al., "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997, doi: 10.1093/sleep/20.12.1077.

[39] N.-Y. Kuo, H.-J. Tsai, S.-J. Tsai, and A. C. Yang, "Efficient screening in obstructive sleep apnea using sequential machine learning models, questionnaires, and pulse oximetry signals: Mixed methods study," *J. Med. Internet Res.*, vol. 26, Dec. 2024, Art. no. e51615. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/39699950

[40] G. Korompili et al., "PSG-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies," *Sci. Data*, vol. 8, no. 1, p. 197, Aug. 2021.

[41] X. Chen et al., "Racial/Ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (MESA)," *Sleep*, pp. 877–888, Jun. 2015.

[42] D. A. Dean et al., "A systematic assessment of the association of polysomnographic indices with blood pressure: The multi-ethnic study of atherosclerosis (MESA)," *Sleep*, vol. 38, no. 4, pp. 587–596, Apr. 2015.

[43] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[44] J. Pradeepkumar et al., "Towards interpretable sleep stage classification using cross-modal transformers," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 2893–2904, 2024.

[45] R. Liu et al., "MHFNet: A multimodal hybrid-embedding fusion network for automatic sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 29, no. 5, pp. 3387–3397, May 2025.

[46] L. Ai et al., "Joint fine-grained representation learning and masked relational modeling for EEG-based automatic sleep staging in fabric space," *IEEE J. Biomed. Health Informat.*, early access, Jul. 23, 2025, doi: 10.1109/JBHI.2025.3592249.

[47] X. Chen, Y. Chen, W. Ma, X. Fan, and Y. Li, "SE-MSCNN: A lightweight multi-scaled fusion network for sleep apnea detection using single-lead ECG signals," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1276–1280.

[48] X. Chen, W. Ma, W. Gao, and X. Fan, "BAFNet: Bottleneck attention based fusion network for sleep apnea detection," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 5, pp. 2473–2484, May 2024.