SleepAC: Less Dependency on Manual Annotations, More Reliable Sampling for Automatic Sleep Staging

Saisai Lv, Donghai Guan, Weiwei Yuan, and Çetin Kaya Koç, *Life Fellow, IEEE*

***Abstract*—Accurate sleep staging is essential for assessing sleep quality and diagnosing sleep disorders, yet it heavily depends on large-scale, expertly labeled datasets, which are costly and time-consuming to produce. While existing methods aim to reduce this reliance, they often utilize data from a limited number of subjects, thereby restricting data diversity and hindering model generalization. To address these challenges, we propose SleepAC, a novel model designed to reduce the dependence on extensive manual annotations. It employs an adaptive sample selection strategy that prioritizes informative and diverse samples, starting with simpler ones and gradually adding more complex ones, while incorporating sleep-specific factors., enabling accurate classification with fewer labeled samples. Furthermore, SleepAC integrates a contrastive learning framework that generates hard negative samples across different sleep stages, effectively enhancing the classification of transitional stages, which are particularly difficult due to limited annotations. Experiments on four public datasets demonstrate that SleepAC achieves competitive accuracy and F1-scores, attaining approximately 95% of the fully supervised performance using only 20% of labeled data. These results underscore its effectiveness in low-resource settings, showcasing promising generalization across complex sleep dynamics while significantly reducing annotation costs.**

***Index Terms*—Sleep stage classification, sample selection, contrastive learning, deep learning.**

I. Introduction

SLEEP plays a fundamental role in maintaining human health, influencing cognitive function, metabolism, and overall well-being. Accurate sleep staging is crucial for assessing sleep quality and diagnosing sleep disorders. According to the American Academy of Sleep Medicine (AASM) guidelines [1], sleep is classified into five stages: wake (W), rapid eye movement (REM), and three non-REM stages (N1, N2, N3), each reflecting different depths of sleep and unique physiological patterns. Traditionally, sleep stage classification relies on expert annotation of polysomnography (PSG) recordings, where trained specialists manually examine segmented 30-second epochs to determine sleep stages. However, this process is time-consuming and labor-intensive, making large-scale data annotation infeasible.

While deep learning has enhanced sleep stage classification by automating feature extraction [2], it remains reliant on large labeled datasets. To mitigate this, recent research has explored alternative methods to reduce dependence on annotated PSG data. One such direction is unsupervised learning, which seeks to uncover inherent patterns from unannotated sleep recordings [3]. By leveraging large amounts of unlabeled data, these methods attempt to build representations of sleep stages without the need for manual labels. However, these methods often struggle to meet the high accuracy demands of sleep staging [4]. Without task-specific supervision, these methods have difficulty identifying the temporal and frequency patterns essential for distinguishing sleep stages. This limitation is particularly evident in detecting transitional stages like N1, which serves as the primary transition from wakefulness to sleep [5]. These stages involve subtle physiological changes that are hard to capture without precise guidance. Consequently, unsupervised models typically lack the granularity and detailed representations necessary for accurate classification, making it challenging to perform at the level required for high-stakes medical applications [6].

Beyond unsupervised learning, efforts have been made to explore semi-supervised learning [7], [8] and self-supervised learning with fine-tuning [9], [10], aiming to bridge the gap between fully supervised and unsupervised approaches. These methods combine a small amount of labeled data with larger sets of unlabeled data, employing techniques such as pre-training for feature extraction or generating pseudo-labels to expand the labeled dataset [11], [12]. However, they often rely on labeled data from a narrow set of subjects, limiting diversity and hindering generalization across variations in age, health conditions, and individual sleep patterns [13]. As a result, these approaches tend to struggle with distribution mismatches between training data and the broader target population, especially when classifying complex or transitional sleep stages where physiological signals can vary significantly. Although these methods reduce

Received 25 October 2024; revised 10 April 2025 and 26 May 2025; accepted 5 July 2025. Date of publication 10 July 2025; date of current version 8 January 2026. This work was supported by the National Natural Science Foundation of China under Grant 62472220. *(Corresponding author: Donghai Guan.)*

The authors are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China (e-mail: lvsaisai@nuaa.edu.cn; dhguan@nuaa.edu.cn; yuanweiwei@nuaa.edu.cn; cetinkoc@ucsb.edu).

Our code is publicly available at: https://github.com/LyuuCode/SleepAC.

Digital Object Identifier 10.1109/JBHI.2025.3587351

the need for extensive labeled data, they still face challenges in generalizing as effectively as fully supervised models.

In this work, we focus on achieving strong in-domain performance with limited labeled data within the target cohort. To this end, we propose SleepAC, an innovative sleep stage classification model that reduces reliance on manual annotation while maintaining high accuracy. SleepAC employs an adaptive sample selection strategy to prioritize informative and diverse PSG samples, optimizing performance with fewer labels. It also incorporates a contrastive learning framework that generates hard negative samples, enhancing the model's ability to distinguish transitional sleep stages. By combining temporal-frequency feature extraction with a reconstruction strategy, SleepAC effectively captures key features, enabling accurate classification of complex sleep stages even with limited annotations. Our main contributions are as follows.

1) We design an adaptive sample selection mechanism that prioritizes simpler, feature-rich samples during early training and gradually shifts to more complex, uncertain samples as training progresses. By incorporating sleep-specific factors, this strategy not only enhances model performance with fewer labeled samples but also improves its ability to generalize across challenging sleep stages.
2) We develop a contrastive learning framework that generates hard negative samples, which closely resemble anchor samples from different sleep stages. This approach enhances the model's ability to capture subtle transitions between sleep stages while preserving physiological patterns.
3) Extensive experiments on four public sleep datasets demonstrate that SleepAC achieves approximately 95% of fully supervised performance with only 20% labeled data, highlighting its robustness and efficiency across diverse sleep conditions while significantly reducing the reliance on manual annotations.

II. Related Work

A. Sleep Sltage Classification

Early sleep stage classification methods mainly used traditional machine learning techniques like SVM and Random Forests [14], [15], which relied on handcrafted features that limited scalability and adaptability. Deep learning revolutionized the field by automating feature extraction, with various architectures developed to model sleep dynamics more effectively. Recurrent Neural Networks (RNNs) [16], [17] capture temporal dependencies in sleep stage transitions for more context-aware classification. L-SeqSleepNet [18] extends this capability by considering entire sleep cycles rather than isolated segments. Convolutional Neural Networks (CNNs) [19], [20] excel at extracting local morphological features by learning spatial representations, enabling effective modeling of time-frequency relationships. U-Sleep [21] uses a fully convolutional network to process EEG and EOG signals, achieving high accuracy. Meanwhile, Graph Neural Networks (GNNs) [22], [23] formulate physiological signals as graph-structured data, enabling systematic modeling of cross-channel interactions. Recently, transformer-based models [24] have gained prominence by using self-attention to capture long-range dependencies. SleepTransformer [25] enhances this by extracting features at both epoch and sequence levels for comprehensive sleep staging. In parallel, hypnodensity graph analysis models sleep stages as continuous probability distributions rather than discrete labels, improving robustness to scorer variability [26], [27]. Despite these advancements, current methods still depend heavily on large-scale labeled datasets, limiting scalability in real-world clinical applications with label inconsistencies and data variability.

B. Active Learning

Effective sample selection is vital for enhancing model performance with limited labeled data. Active learning, a key strategy in deep learning, aims to minimize annotation costs by selecting the most informative samples for labeling. This approach has been extensively applied in fields such as computer vision and medical imaging [28]. Various sample selection strategies have been developed to improve active learning efficiency. For instance, BatchBALD [29] maximizes mutual information to select diverse samples, while Yoo et al. [30] prioritize uncertain samples by predicting classification errors. ALFA-Mix [31] improves diversity by analyzing inconsistencies in interpolated feature representations. Despite their success in other fields, the application of active learning in sleep staging remains limited. Hossain et al. [32] employed an importance-weighted approach to reduce manual labeling, and Macas et al. [33] introduced a confidence-based approach for EEG annotation. However, these methods mainly rely on uncertainty-based sampling, which can introduce noise by prioritizing complex samples early in training. They also overlook unique characteristics of sleep signals, such as waveform patterns and noise sensitivity, limiting their effectiveness in sleep stage classification.

C. Contrastive Learning

Contrastive learning, an advanced representation learning technique, enhances model generalization by learning discriminative representations through positive-negative sample contrast. It is widely utilized across various research domains [34]. In the field of sleep stage classification, contrastive learning has shown promising results. For example, methods like TsC-EA [10] and CoSleep [35] have employed innovative strategies to enhance model performance using unlabeled data. TsC-EA leverages contrastive learning by integrating EEG signal features through temporal segmentation and electrode autoencoders, while CoSleep utilizes a multi-view co-training mechanism and memory modules to improve feature representation. Similarly, Shen et al. [36] addresses the issue of false negatives in contrastive learning by incorporating temporal augmentation and false-negative suppression. Despite recent advancements, contrastive learning methods for sleep stage classification still face challenges. Many current approaches generate 'simple negatives' that are too distant from positive samples in feature space, limiting their effectiveness in helping models learn deeper feature representations, thus reducing learning efficiency and overall performance.

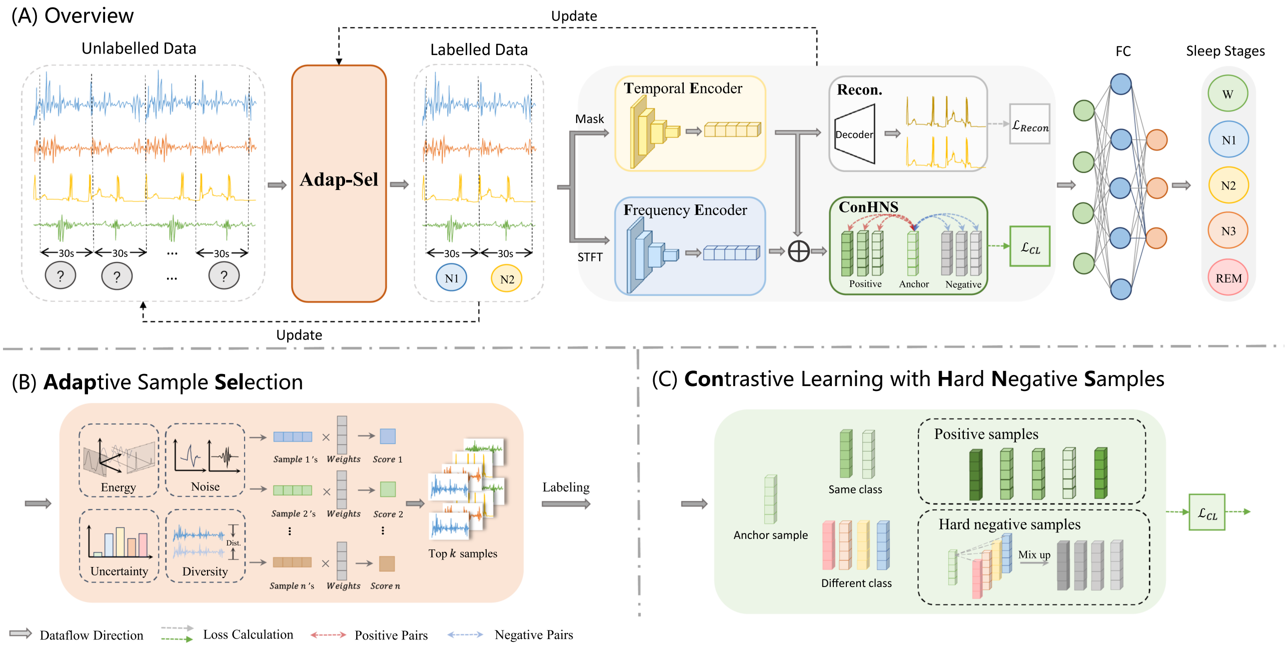

Fig. 1. Overview of SleepAC approach, shown in (A), consisting of Adap-Sel, TF-FE, ConHNS, and Recon four components. It begins with Adap-Sel to identify the most informative samples from the unlabeled dataset for annotation. Temporal and frequency features are then extracted from these annotated samples using temporal encoder and frequency encoder in TF-FE. The temporal features are reconstructed in the Recon module to preserve essential time-based information. Feature discrimination is further refined through contrastive learning in the ConHNS module with hard negative samples. Subfigures (B) and (C) denote the specific process of Adap-Sel and ConHNS.

III. Method

A. Preliminaries

In the context of sleep stage classification, we adopt an *epoch-to-epoch* processing paradigm, where full-night PSG recordings are segmented into consecutive 30-second intervals. This approach aligns with clinical annotation practices. Each interval is independently labeled with one of five distinct stages: Wake, N1, N2, N3, and REM. Formally, a segmented PSG sequence is defined as $X_i = \{x_i^1, x_i^2, \ldots, x_i^M\} \in \mathbb{R}^{M\times T}$, where M denotes the number of PSG signal channels, and T denotes the number of time points in each channel signal x_i^k for $k \in \{1, 2, \ldots, M\}$. We further denote the unlabeled dataset as $D_U = (X_i)_{i=1}^{N_U}$ and the labeled dataset as $D_L = \{(X_i, y_i)\}_{i=1}^{N_L}$, where N_U and N_L represent the number of unlabeled and labeled samples, respectively, and $y_i \in \{1, 2, 3, 4, 5\}$ indicates the sleep stage label. Our objective is to strategically select the most informative samples from D_U, label them, and expand D_L. The ultimate aim is to train a robust classification model $F(\cdot)$ capable of accurately predicting the sleep stage label $\hat{y}_i$ for any given input X_i with minimal manual annotation effort.

B. Overview

The overall framework of our proposed SleepAC is illustrated in Fig. 1. It consists of four main components: Adaptive Sample Selection (Adap-Sel), Temporal-Frequency Feature Extraction (TF-FE), Contrastive Learning with Hard Negative Samples (ConHNS), and Reconstruction (Recon). It begins with the Adap-Sel module, which systematically selects the most informative PSG signal segments for labeling and training. Then, TF-FE module processes these signals to extract both temporal and spectral features. ConHNS module introduces hard negative samples to improve the model's ability to differentiate between closely related sleep stages. Finally, Recon module focuses on reconstructing masked inputs to preserve temporal information and further enhance feature representations. Each module plays a specific role in the SleepAC framework, contributing to the overall design for efficient sleep stage classification.

C. Adaptive Sample Selection

We propose an adaptive sample selection strategy designed to improve labeling and training efficiency in sleep stage classification tasks. This approach systematically prioritizes the most valuable PSG signal segments, which reduces the need for extensive labeling while maintaining model performance. By evaluating samples based on signal clarity and informative content, simpler samples are introduced first, allowing the model to build a solid foundation. As the training progresses, more complex samples are gradually introduced, enhancing the model's ability to generalize and improve classification accuracy.

1) Signal Clarity: Signal clarity is a crucial factor in determining the learnability of each sample, assessed through its energy and noise levels. These intrinsic characteristics help identify samples that are either easier or more challenging for the model to learn from, guiding the progressive introduction of data during training.

The energy of a signal is calculated based on common waveforms in sleep signals, such as alpha waves, delta waves, and sleep spindles, which exhibit distinct patterns across different sleep stages. High-energy samples typically present clear and distinguishable features, making them ideal for early-stage learning. The energy of a sample X_i is quantified as:

$$S_e(X_i) = \sum_{k \in F} |\mathrm{FFT}(X_i)[k]|^2 \tag{1}$$

where F is the set of indices within the desired frequency range, such as those characteristic of specific sleep-related waveforms; k represents the FFT frequency index.

Noise impacts the clarity of a sample's features, with lower noise levels enhancing the informativeness of the data, particularly in the initial phases of training. The noise level of a sample X_i is computed as:

$$S_d(X_i) = \exp\left(-\frac{1}{T}\Sigma_{t=1}^{T}|X_{i,t} - \tilde{X}_{i,t}|^2\right) \tag{2}$$

where $X_{i,t}$ represents the original signal, and $\tilde{X}_{i,t}$ denotes the denoised signal obtained via wavelet transform at time point t. This expression calculates the mean squared error between the original and denoised signals, quantifying the amount of noise present. Higher values indicate cleaner signals with less noise, which enhances feature clarity, especially beneficial during the early stages of model training.

2) Informative Value: The informative value of each sample is assessed based on its uncertainty and diversity, which enable the model to capture complex and varied features essential for robust generalization.

Uncertainty indicates the model's confidence in its predictions. Samples with high uncertainty (low confidence) are often more informative because they highlight areas where the model is uncertain or prone to errors, making these samples crucial targets for labeling. The uncertainty of a sample X_i is quantified as:

$$V_u(X_i) = -\max\left(\frac{\exp(z_k/\tau)}{\sum_{j=1}^{K}\exp(z_j/\tau)}\right) \tag{3}$$

where z_k represents the logit value corresponding to the predicted class, and τ is a temperature parameter that controls the distribution sharpness.

Diversity measures the uniqueness of a sample by calculating its distance from other samples in the dataset. High diversity indicates that a sample covers distinct features, enhancing the model's ability to generalize across varied data. The diversity of a sample X_i is defined as:

$$V_d(X_i) = -\frac{1}{|F_L|}\sum_{f_i \in F_U}\log\sum_{f_j \in F_L}\exp(\mathrm{sim}(f_i, f_j)) \tag{4}$$

where F_U and F_L are the sets of features from the unlabeled and labeled samples, respectively. The term $\mathrm{sim}(\cdot,\cdot)$ represents the similarity between features. Specifically, cosine similarity is employed as the similarity metric:

$$\mathrm{sim}(u, v) = \frac{u^\top v}{\|u\|_2\,\|v\|_2}. \tag{5}$$

By prioritizing diverse samples, the method ensures that the model learns a broader range of features, improving overall adaptability and performance.

3) Sample Selection: To integrate the above metrics, we define a scoring function $\mathrm{Score}(X_i)$ that balances signal clarity and informative value. This score guides the prioritization of samples for labeling and training:

$$\mathrm{Score}(X_i) = \alpha_1 S_e(X_i) + \alpha_2 S_d(X_i) + \beta_1 V_u(X_i) + \beta_2 V_d(X_i) \tag{6}$$

where α_1, α_2, β_1, β_2 are weights adjusting the contribution of each metric. Initially, β_1 and β_2 are set to zero,emphasizing simpler samples. As training progresses, these weights increase, integrating more complex and informative samples. Samples are selected by:

$$X^* = \arg\max_{X_i \in D_U} \mathrm{Score}(X_i). \tag{7}$$

This adaptive strategy effectively guides the model, prioritizing valuable samples and enhancing performance while minimizing labeling efforts. Notably, our strategy aligns naturally with the epoch-level processing, as each segment's signal clarity and informativeness are assessed independently. This enables selective labeling of key segments, boosting annotation efficiency and reducing manual effort.

D. Temporal-Frequency Feature Extraction

To accurately classify sleep stages, our model employs both time-domain and frequency-domain feature extraction methods to capture the key characteristics of PSG signals comprehensively.

The time-domain feature extractor captures effective temporal features from raw time-series signals. It consists of four residual blocks, each containing two standard convolutional layers and a projection shortcut connection. This architecture helps mitigate the vanishing gradient problem and improves feature propagation through the network. The varying number of convolutional filters captures features at different scales, enhancing the model's adaptability to diverse temporal patterns. Each convolutional layer employs ReLU activations and batch normalization to improve training stability:

$$z_i^{temp} = \mathrm{TemporalEncoder}(X_i) \tag{8}$$

where z_i^{temp} represents the extracted time-domain features.

The frequency-domain feature extractor processes the signal in the frequency domain. It applies a Short-Time Fourier Transform (STFT) to convert the signal into a spectral representation, followed by multiple dilated convolution blocks to capture spectral features over larger receptive fields. This method is crucial for identifying long-term dependencies and periodic patterns in sleep signals:

$$z_i^{freq} = \mathrm{FrequencyEncoder}(\mathrm{STFT}(X_i)) \tag{9}$$

where z_i^{freq} denotes the extracted frequency-domain features.

These time-domain and frequency-domain features are concatenated to form a comprehensive representation:

$$z_i = \mathrm{CONCAT}(z_i^{temp}, z_i^{freq}). \tag{10}$$

The comprehensive feature representation z_i is then passed through the classification head to produce sleep stage logits. The classification head consists of a fully connected layer followed by a softmax activation, formulated as:

$$\hat{y}_i = \text{Softmax}(W z_i + b) \tag{11}$$

where W and b are trainable parameters. The output $\hat{y}_i$ represents the probability distribution over the five sleep stage.

E. Contrastive Learning With Hard Negative Samples

Accurately identifying transitional sleep stages is a significant challenge in classification, as traditional methods often fail to capture their subtle features, compromising overall performance. To address this challenge, we leverage the principles of contrastive learning, specifically focusing on constructing hard negative samples. By introducing hard negatives that are closely related to positive samples yet belong to different classes, we refine the model's ability to differentiate between nuanced features.

In our framework, we construct hard negative samples using linear interpolation. For each anchor sample z_i in a batch $\mathcal{B} = \{z_1, z_2, \ldots, z_{|\mathcal{B}|}\}$, we generate hard negative samples z_i^{neg} as follows:6

$$z_i^{neg} = \lambda z_i + (1 - \lambda) z_j \tag{12}$$

where j denotes a sample from the batch $\mathcal{B}$ that belongs to a class with a transitional relationship to i, and $\lambda \epsilon (0, 0.5]$ ensures the anchor sample's influence is less than that of the negative sample. The selection of the transitional relationship is based on feature similarity measures or domain knowledge to better capture subtle transitions. Positive samples z_i^{pos} are selected from the same class to ensure high similarity with the anchor sample.

We employ a contrastive loss function to effectively leverage these hard negative samples:

$$\mathcal{L}_{CL} = \sum_{i \in \mathcal{B}} \frac{-1}{|P(i)|} \log \frac{exp(z_i \cdot z_i^{pos/} \tau)}{exp(z_i \cdot z_i^{pos}/\tau) + exp(z_i \cdot z_i^{neg}/\tau)} \tag{13}$$

where $P(i) = \mathcal{B} \setminus \{z_i\}$, and τ is a temperature parameter that controls the sensitivity of the loss. This approach enhances the model's capacity to discern sleep stage features, particularly during transitional phases.

F. Reconstruction

Inspired by the success of masked autoencoders such as BERT [11] and MAE [37], we implement a reconstruction module to preserve crucial temporal information and enhance reconstruction capabilities. Our framework utilizes a masked autoencoding strategy where random masking is applied to the original signal to create partially observed signals. These signals are then projected into a latent space through a temporal encoder, and the complete signal is reconstructed from this representation.

The reconstruction loss measures the difference between the original and reconstructed signals:

$$\mathcal{L}_{Recon} = \frac{-1}{|\mathcal{B}|} \sum_{i \in |\mathcal{B}|} \left\| M_i \odot \left(\hat{X}_i - X_i \right) \right\|_2^2 \tag{14}$$

where $\odot$ denotes element-wise multiplication, M_i is the mask matrix, and $\hat{X}_i$ and X_i represent the reconstructed and original signals, respectively. This loss focuses on masked positions, enabling the model to learn underlying temporal patterns effectively.

To balance the objectives of classification, contrastive learning, and reconstruction, we combine cross-entropy loss, contrastive loss, and reconstruction loss into an overall loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{CL} + \lambda_3 \mathcal{L}_{Recon} \tag{15}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss, $\mathcal{L}_{CL}$ denotes the contrastive loss, and $\mathcal{L}_{Recon}$ indicates the reconstruction loss. The parameters λ_1, λ_2, and λ_3 are weighting factors that balance the contributions of each loss component to the total loss. This comprehensive loss function ensures classification accuracy while enhancing the model's generalization across transitional sleep stages.

IV. Experiments and Discussion

A. Datasets

We employed four publicly available datasets in this study: ISRUC-S1 [38], ISRUC-S3 [38], Sleep-EDF-20 [39], and Sleep Heart Health Study (SHHS) [40]. In line with AASM guidelines, PSG signals were segmented into 30-second epochs, with each epoch labeled as Wake, N1, N2, N3, or REM. A summary of the datasets is shown in Table I.

1) ISRUC-S1: This dataset includes overnight PSG recordings from 100 subjects (55 males, 45 females) with sleep disorders, aged 20 to 85 years. The recordings include 12 channels, we selected chin EMG, EEG, EOG, and ECG channels in the experiments. To ensure consistency, the original 200 Hz signals were down-sampled to 100 Hz. This dataset provides insights into sleep pathologies and abnormal sleep patterns in clinical populations.

2) ISRUC-S3: This dataset contains overnight PSG recordings from 10 healthy subjects (9 males, 1 female) aged 30 to 58 years. It contains the same channels as ISRUC-S1 and was down-sampled to 100 Hz. Unlike ISRUC-S1, this dataset offers a view of normative sleep patterns in healthy individuals.

3) Sleep-EDF-20: This dataset includes 39 full-night PSG recordings from 20 healthy subjects (10 males, 10 females), aged 25 to 34 years. Each subject has two recordings, except for subject No. 13. Recordings were trimmed starting from 30 minutes before ‘lights off’ to 30 minutes after ‘lights on’. We adopted the Fpz-Cz and Pz-Oz EEG channels in the experiments.

4) Shhs: This dataset is a large-scale, multi-center study comprising PSG data from 5,793 subjects aged between 39

TABLE I
DATASETS DESCRIPTION

Dataset	Subjects	Recordings	Conditions	Experimental Setup	Epochs	W(%)	N1(%)	N2(%)	N3(%)	REM(%)
ISRUC-S1	100	100	Various	10-Fold CV	87187	23.1	12.7	31.6	19.8	12.9
ISRUC-S3	10	10	All Healthy	10-Fold CV	8549	19.3	14.2	30.5	23.6	12.4
Sleep-EDF-20	20	39	All Healthy	20-Fold CV	42348	19.7	6.6	42.0	13.5	18.2
SHHS	5793	5793	Various	train/test: 0.7/0.3	5863207	28.8	3.7	40.9	12.6	13.9

and 90 years, from diverse demographic backgrounds. Unlike the previous datasets, SHHS includes a wide range of sleep-disordered breathing conditions, making it essential for evaluating model generalizability. For our experiments, we adopted the C4-A1 EEG channel, as it is commonly used in sleep staging studies.

B. Experiment Settings

1) Implementation Details: The proposed model is implemented in Python 3.9 with PyTorch 2.1.1. All experiments are conducted on a server with Nvidia 3090 GPUs. The model is trained for 80 epochs using the Adam optimizer with an initial learning rate of 0.001. A cosine annealing learning rate scheduler is applied to gradually reduce the learning rate. The batch size is set to 64, and a dropout rate of 0.1 is used to improve robustness. In the contrastive learning framework, a temperature parameter of $\tau = 0.07$ is used, while all other hyperparameters follow PyTorch defaults.

For evaluation, all experiments are conducted under an in-domain setting. ISRUC-S1 and ISRUC-S3 adopt subject-independent 10-fold cross-validation (CV), ensuring that data from the same subject does not appear in both training and test sets. Sleep-EDF-20 follows a subject-independent 20-fold CV strategy. In contrast, SHHS is divided into a fixed training (70%) and test (30%) split. Performance metrics are averaged across all CV folds or reported based on the test set for SHHS, providing a comprehensive assessment of the model's generalizability within each domain.

2) Evaluation Metric: In this study, we employed three key evaluation metrics: Accuracy (Acc), Macro-averaged F1-score (MF1), and Cohen's Kappa coefficient (κ), to comprehensively assess the performance of the proposed SleepAC model in sleep stage classification. These metrics are essential for evaluating the model's overall accuracy and its ability to handle class imbalance in sleep stage data.

C. Comparison With Supervised Methods

We evaluate the proposed model on the ISRUC-S3, ISRUC-S1, Sleep-EDF-20 and SHHS datasets, comparing its performance with various state-of-the-art (SOTA) methods, including RNN models [41], [42], [52], CNN models [45], [47], [48], [49], [50], [51], GCN models [22], [23], [44], transformer-based models [24], [25], [55] and contrastive learning-based models [46]. We also compare with U-Sleep [21], a large-scale supervised model trained on fully annotated multi-center data and evaluated on ISRUC datasets, which serves as a benchmark for large-scale supervised learning performance.

As shown in Table II, our model consistently outperforms baselines in both accuracy and F1-score across the ISRUC-S3, ISRUC-S1, and Sleep-EDF-20, demonstrating strong classification performance. On the large-scale SHHS dataset, it achieves performance comparable to leading baselines such as L-SeqSleepNet and Cross-modal Transformer, while maintaining a highly competitive F1-score. This underscores the model's scalability and robustness under high inter-subject variability and diverse sleep conditions.

To assess the model's efficiency under limited-label scenarios, we conduct experiments using only 20% of the labeled data. As shown in Table III, the model retains over 97% of the SOTA performance on ISRUC-S3, ISRUC-S1, and Sleep-EDF-20, demonstrating strong capability in low-resource settings. This makes it well-suited for small- and medium-scale applications such as research laboratories and clinical studies, where manual annotation is costly and time-consuming. In particular, as shown in Table II, while U-Sleep relies on large-scale fully labeled datasets for cross-domain generalization, our model achieves better in-domain performance on ISRUC-S1 and ISRUC-S3 with significantly less annotated data, demonstrating strong adaptability under limited supervision.

On the SHHS dataset, a more pronounced performance drop is observed under the same label proportion. This can be attributed to the increased inter-subject variability, more complex sleep patterns, and a higher degree of class imbalance inherent in SHHS. Nevertheless, the model retains around 91% of its fully supervised performance, showing strong adaptability even in large-scale, diverse conditions.

Averaging across all four datasets, our model achieves around 95% of the full-supervision performance, demonstrating robustness and adaptability under limited-label conditions. Despite the challenges of the SHHS dataset, the approach significantly reduces annotation costs while maintaining strong performance. Additionally, Fig. 3 shows the predicted sleep stage sequence for Subject 1 in the ISRUC-S3 dataset, with red 'x' markers indicating misclassified epochs. Despite a few such errors, the predicted sequence closely aligns with the ground truth, highlighting the model's reliability.

D. Comparison With Self-Supervised Methods

We compare the performance of our model with several SSL baselines, including CoSleep [35], TS-TCC [56], CA-TCC [57], and SA-TSC [58], on ISRUC-S3, Sleep-EDF-20 and SHHS. For fair comparison, all SSL baselines are fine-tuned using 20% of the data from a subset of subjects, following common SSL practices [58]. In contrast, SleepAC uses 20% of the data

TABLE II
PERFORMANCE COMPARISON WITH THE SOTA SUPERVISED METHODS ON DATASETS WITH FULL LABELS

Dataset	Model	Overall performance			F1-score for each class				
		ACC	MF1	Kappa	W	N1	N2	N3	REM
ISRUC-S3	DeepSleepNet [41]	78.8	77.9	0.730	88.7	60.2	74.6	85.8	80.2
	SeqSleepNet [42]	78.9	76.3	0.725	83.6	43.9	79.3	87.9	86.7
	GraphSleepNet [22]	79.9	78.7	0.741	87.8	57.4	77.6	86.4	84.1
	U-Sleep [21]	-	77.0	-	90.0	55.0	78.0	74.0	85.0
	MSTGCN [23]	82.1	80.8	0.769	89.4	59.6	80.6	89.0	85.6
	XSleepNet1 [43]	82.5	80.8	0.774	90.1	58.6	82.5	88.7	84.3
	XSleepNet2 [43]	82.6	81.0	0.774	89.9	59.0	82.6	88.4	84.9
	SleepTransformer [25]	81.8	80.3	0.779	89.2	57.4	81.3	88.5	85.2
	JK-STGCN [44]	83.1	81.4	0.782	90.0	59.8	82.6	90.1	84.5
	3DSleepNet [45]	83.2	81.4	0.783	89.6	59.6	83.2	90.9	83.8
	SleePyCo [46]	82.8	80.8	0.774	90.1	58.4	82.9	88.2	84.4
	MixSleepNet [47]	83.0	82.1	0.782	89.9	62.5	81.9	89.9	86.0
	Ours	**84.1**	**82.4**	**0.788**	**89.7**	**64.7**	**82.3**	**89.8**	**85.3**
ISRUC-S1	DeepSleepNet [41]	73.9	72.2	0.654	84.9	50.5	72.8	85.4	65.3
	SeqSleepNet [42]	77.0	68.3	0.648	84.4	12.4	76.9	85.3	79.4
	GraphSleepNet [22]	76.3	72.7	0.715	85.1	41.8	76.1	82.3	77.3
	U-Sleep [21]	-	77.0	-	89.0	52.0	79.0	77.0	88.0
	MSTGCN [23]	79.2	77.2	0.752	86.8	53.2	78.3	85.3	82.3
	XSleepNet2 [43]	80.5	78.4	0.751	89.4	50.6	80.2	87.5	84.4
	XSleepNet1 [43]	80.6	78.2	0.748	89.4	50.4	80.0	87.1	83.9
	SleepTransformer [25]	80.2	77.8	0.758	89.1	49.6	79.6	85.9	84.6
	JK-STGCN [44]	82.0	78.7	0.752	88.5	53.9	79.9	87.6	83.8
	3DSleepNet [45]	82.0	79.7	0.768	90.8	53.4	80.8	88.0	85.5
	SleePyCo [46]	80.8	78.1	0.749	89.6	50.0	80.1	87.0	83.9
	MixSleepNet [47]	81.3	78.7	0.757	90.8	51.2	79.9	87.1	84.4
	Ours	**82.3**	**80.1**	**0.771**	**90.7**	**55.6**	**81.0**	**87.1**	**85.6**
Sleep-EDF-20	DeepSleepNet [41]	81.9	76.6	0.763	86.7	45.5	85.1	83.3	82.6
	SeqSleepNet [42]	85.2	78.4	0.801	90.5	45.4	88.1	86.4	81.8
	TinySleepNet [48]	85.4	80.5	0.799	90.1	51.4	88.5	88.3	84.3
	AttnSleep [49]	84.4	78.1	0.790	89.7	42.6	88.8	90.2	79.0
	XSleepNet1 [43]	86.0	80.0	0.810	91.3	49.5	88.0	86.9	84.2
	XSleepNet2 [43]	86.3	80.6	0.813	92.2	51.8	88.0	86.8	83.9
	SleepFCN [50]	84.8	78.8	0.791	89.6	44.6	89.1	90.6	80.3
	SleepTransformer [25]	85.7	79.6	0.794	90.0	47.0	88.2	88.6	84.2
	Fang et al. [51]	85.8	78.5	0.795	90.3	40.1	89.5	90.6	81.9
	L-SeqSleepNet [18]	86.3	79.3	0.813	91.6	45.3	88.5	86.2	85.2
	Zhao et al. [52]	85.6	81.1	0.800	90.4	53.7	88.3	88.3	85.1
	SleePyCo [46]	86.2	81.2	0.820	91.5	50.0	89.4	89.0	86.3
	Ours	**86.3**	**81.2**	**0.801**	**89.7**	**54.0**	**88.3**	**89.0**	**84.9**
SHHS	CNN [53]	86.8	78.5	0.810	-	-	-	-	-
	SeqSleepNet [42]	87.2	80.2	0.820	91.8	49.1	88.2	83.5	88.2
	IITNet [54]	86.7	79.8	0.810	-	-	-	-	-
	AttnSleep [49]	84.2	75.3	0.780	86.7	33.2	87.1	87.1	82.1
	FCNN+RNN [43]	86.7	79.5	0.813	91.1	48.7	88.0	82.6	87.1
	XSleepNet1 [43]	87.5	81.0	0.826	91.6	51.4	88.5	85.0	88.4
	XSleepNet2 [43]	87.6	80.7	0.826	92.0	49.9	88.3	85.0	88.2
	SleepTransformer [25]	87.7	80.1	0.828	92.2	46.1	88.3	85.2	88.6
	SleepViTransformer [24]	88.1	79.8	0.830	93.4	44.4	88.5	84.5	88.3
	L-SeqSleepNet [18]	88.4	81.4	0.838	93.1	51.1	89.0	84.9	89.8
	Cross-modal transformer [55]	87.7	81.4	0.829	92.8	52.5	88.3	83.7	89.8
	SleePyCo [46]	87.9	80.7	0.830	92.6	49.2	88.5	84.5	88.6
	Ours	**87.8**	**81.3**	**0.832**	**92.4**	**51.3**	**88.7**	**84.5**	**89.6**

TABLE III
PERFORMANCE COMPARISON BETWEEN OUR METHOD (WITH 20% LABELS) AND SOTA (WITH FULL LABELS) PERFORMANCE ON DIFFERENT DATASETS

Dataset	Model	Overall performance			F1-score for each class				
		ACC	MF1	Kappa	W	N1	N2	N3	REM
ISRUC-S3	SOTA*	83.2	81.4	0.783	89.6	59.6	83.2	90.9	83.8
	Ours	81.6	79.0	0.754	86.3	57.5	80.8	88.9	82.3
	Performance Ratio	98.08%	97.05%	96.30%	96.32%	96.48%	97.12%	97.80%	98.21%
ISRUC-S1	SOTA*	82.0	79.7	0.768	90.8	53.4	80.8	88.0	85.5
	Ours	79.8	77.9	0.739	87.6	51.7	80.0	86.8	83.3
	Performance Ratio	97.32%	97.74%	96.22%	96.48%	96.82%	99.01%	98.64%	97.43%
Sleep-EFD-20	SOTA*	86.3	81.2	0.813	91.6	45.3	88.5	86.2	85.2
	Ours	84.2	79.3	0.786	88.3	50.4	88.2	87.5	81.8
	Performance Ratio	97.68%	97.66%	95.85%	96.50%	100.80%	98.66%	98.31%	94.79%
SHHS	SOTA*	88.4	81.4	0.838	93.1	51.1	89.0	84.9	89.8
	Ours	81.8	72.9	0.798	84.9	41.1	83.5	77.7	77.1
	Performance Ratio	92.5%	89.5%	95.2%	91.2%	80.4%	93.8%	91.5%	85.9%

* The SOTA baselines are 3DSleepNet for ISRUC-S1 and ISRUC-S3, and L-SeqSleepNet for Sleep-EDF-20 and SHHS.

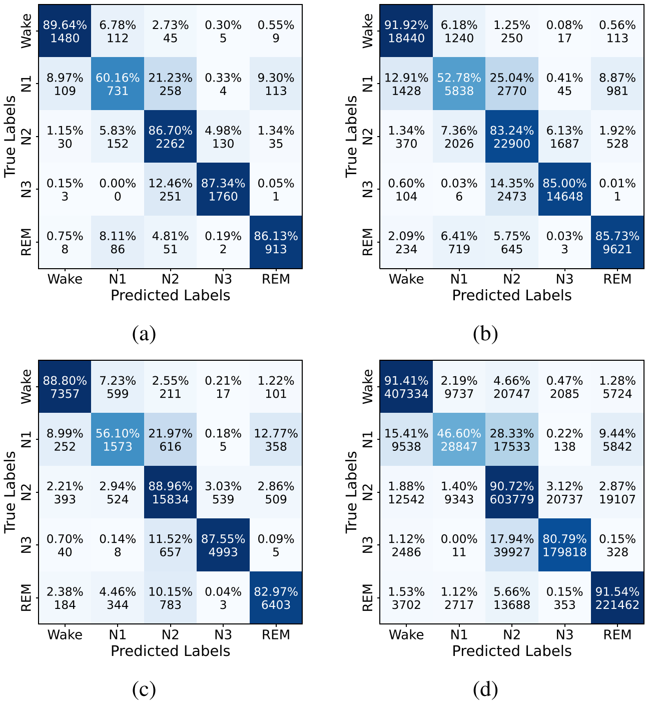

Fig. 2. The confusion matrix of SleepAC on different datasets. (a) ISRUC-S3. (b) ISRUC-S1. (c) Sleep-EDF-20. (d) SHHS.

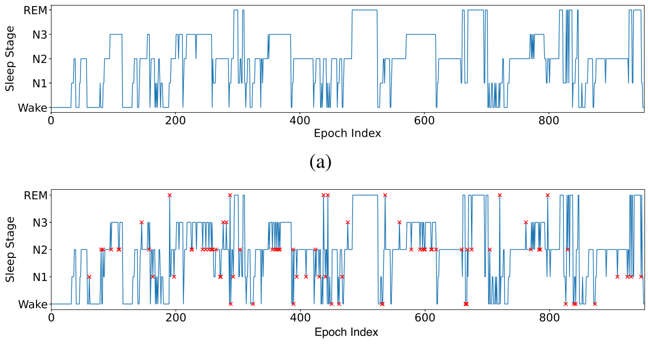

Fig. 3. The comparison between ground truth and predicted sleep stage of subject 1 by SleepAC in ISRUC-S3 dataset. (a) Ground truth. (b) prediction.

sampled across all subjects without fine-tuning, providing a more comprehensive representation of the dataset.

As shown in Table IV, SleepAC significantly outperforms the SSL methods on ISRUC-S3 and Sleep-EDF-20, achieving accuracies of 81.6% and 84.2%, respectively. On the larger SHHS dataset, SleepAC's performance remains competitive, achieving an accuracy of 81.8% and an MF1 of 72.9%. Although SSL methods show comparable performance on SHHS, this can be partially attributed to their pretraining on the full unlabeled dataset, which allows them to learn generalized representations from a wide range of subjects. This advantage is particularly beneficial in large-scale datasets like SHHS. Nevertheless, SleepAC still demonstrates strong robustness and generalization, even in the presence of such complexity.

TABLE IV
PERFORMANCE COMPARISON WITH THE SOTA SELF-SUPERVISED METHODS ON ISRUC-S3, SLEEP-EDF-20 AND SHHS DATASETS

Model	ISRUC-S3		SleepEDF		SHHS	
	Acc	MF1	Acc	MF1	Acc	MF1
CoSleep	57.7	53.8	71.4	61.5	78.3	62.8
TS-TCC	68.3	65.5	78.2	70.8	73.6	63.4
CA-TCC	75.3	71.9	76.9	67.5	80.2	70.2
SA-TSC	74.6	70.8	78.0	71.7	79.2	68.6
Ours	**81.6**	**79.0**	**84.2**	**79.3**	**81.8**	**72.9**

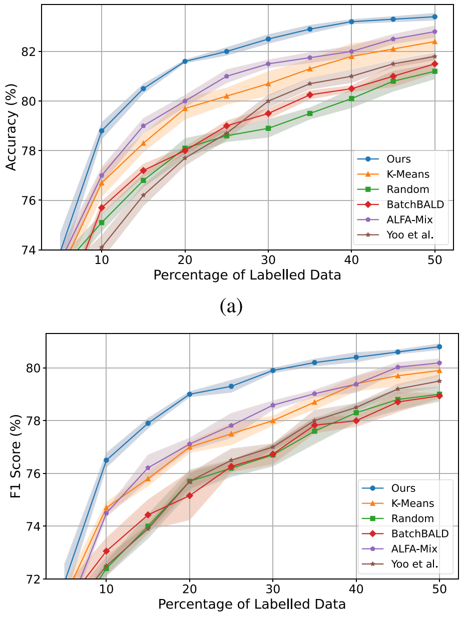

Fig. 4. Comparison of Accuracy and F1-score between Adaptive and Other Sample Selection Methods on the ISRUC-S3 Dataset. (a) Acc. (b) F1-score.

Despite using only 20% of the labeled data, SleepAC maintains strong performance across all datasets, demonstrating its effectiveness in low-label scenarios. This strong performance can be attributed to our SleepAC's ability to capture a broader range of physiological variations, such as age, health conditions, and individual sleep patterns, which contributes to better generalization across different populations. By reducing distribution mismatch and learning from a more representative subset of the dataset, SleepAC preserves strong classification performance while minimizing the need for extensive labeled data. These results underscore SleepAC's robustness in real-world sleep data applications, where labeled data is often limited.

E. Evaluation of Adaptive Sample Selection Method

We compared our adaptive sample selection method with several active learning strategies, including BatchBALD [29], Yoo et al. [30], and ALFA-Mix [31], as well as traditional methods like Random and K-Means sampling, on the ISRUC-S3 dataset. As shown in Fig. 4, our method consistently outperforms these baselines. For instance, with only 20% labeled data, our approach improves F1-score by 1.9% compared to the best baseline, demonstrating its ability to select informative samples for more efficient model training.

The key strength of our method lies in its ability to account for the unique characteristics of sleep data, combined with a strategy that prioritizes simpler samples initially and gradually incorporates more complex ones. While Random sampling and K-Means clustering are widely used, they fail to consider the specific properties of sleep signals, such as distinct waveform patterns and the influence of noise. Similarly, most existing active learning strategies focus on general uncertainty-based criteria, but neglect sleep-specific characteristics. In addition, these strategies tend to select highly uncertain and complex samples early in training, which may introduce noise and hinder model convergence. In contrast, our method adopts a progressive sampling strategy, starting with easier samples and incrementally introducing more challenging ones as the model matures. By explicitly incorporating sleep-specific characteristics, our approach enables more efficient learning and better generalization across diverse subjects, even under limited labeled data conditions.

Moreover, our strategy is not limited to sleep analysis and has broader applicability. It can be applied to other physiological signal analysis tasks, such as ECG classification for heart disease, EEG anomaly detection for epilepsy, and time-series analysis. By reducing labeling requirements and improving model performance, our approach has the potential to make a significant impact across these domains.

It is important to note that, like other active learning strategies, our method requires waiting for the selected samples to be labeled before retraining the model. While this may introduce some delay, it substantially reduces the total annotation effort by focusing on the most informative samples, which is particularly valuable in expert-labeled domains like sleep staging.

F. Ablation Study

To better understand the impact of each component in our model, we conducted an ablation study using four variants of the model on the ISRUC-S3 dataset:

1) *Variant A (Temporal):* Uses only time-domain features with a ResNet architecture as a baseline.
2) *Variant B (Variant A + Frequency):* Adds frequency-domain features to explore multimodal benefits.
3) *Variant C (Variant B + Reconstruction):* Integrates a reconstruction module to enhance feature capture.
4) *Variant D (Variant C + Contrastive Learning):* Adds a contrastive learning module to improve handling of transitional sleep stages.

Asx shown in Table V, Variant A shows limited capability in modeling sleep stage complexity, particularly during transitions, resulting in relatively low accuracy and MF1. Adding frequency-domain features in Variant B yields clear performance gains, indicating the advantage of multimodal representations. Variant C further enhances accuracy and MF1 with the addition of a reconstruction module, effectively handling signal complexity and transitions. Finally, Variant D, which integrates the contrastive learning module, achieves the best performance, with a 1.2% improvement in accuracy and a 2.2% increase in MF1 compared to Variant C. This demonstrates the module's ability to enhance feature representation and improve classification of transitional sleep stages, leading to the highest overall accuracy and MF1 across all variants. Fig. 5 further shows improvements in per-class F1-scores, emphasizing the contributions of time-frequency features, the reconstruction module, and contrastive learning.

TABLE V
ABLATION EXPERIMENT RESULTS OF DIFFERENT VARIANTS MODELS ON ISRUC-S3

Variant	Acc	MF1	κ
Variant A	81.1	77.9	74.7
Variant B	82.3	79.9	76.5
Variant C	82.9	80.2	77.2
Variant D	84.1	82.4	78.8

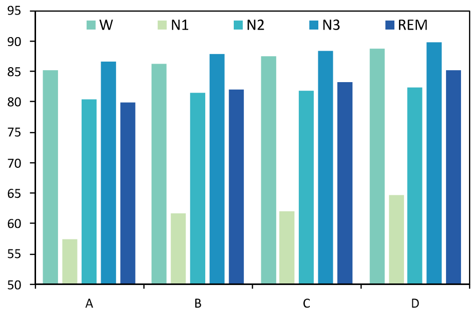

Fig. 5. Specific F1-scores for each class of ablation experiments on ISRUC-S3 dataset with different variant models.

V. Conclusion

In this study, we introduce SleepAC, a novel model for sleep stage classification that balances high accuracy with reduced reliance on extensive manual annotations. Central to SleepAC is an adaptive sample selection mechanism that prioritizes informative and diverse samples while accounting for sleep-specific characteristics. This strategy begins with simpler samples and gradually incorporates more complex ones, enhancing performance with fewer labeled samples and lowering annotation costs. To improve classification of transitional stages, SleepAC integrates a contrastive learning framework that generates hard negative samples targeting these challenging transitions. Combined with time- and frequency-domain feature extraction and a reconstruction module, SleepAC effectively identifies key features for accurate classification, even with limited annotations. Experiments on ISRUC-S3, ISRUC-S1, Sleep-EDF-20, and SHHS demonstrate SleepAC's superior performance, achieving high accuracy and F1-scores with reduced labeled data. Ablation studies further validate the contributions of each component to overall effectiveness.

While our method improves sleep stage classification within the target domain using limited labeled data, it currently focuses on individual epochs and may overlook long-range temporal dependencies important for sleep dynamics. Moreover, it does not address cross-domain generalization, which may limit its applicability to data from different populations or centers. Future work will explore incorporating sequence-level modeling to capture broader temporal context and extend the approach to improve generalizability across diverse domains.

References

[1] R. B. Berry et al., "Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events: Deliberations of the sleep apnea definitions task force of the american academy of sleep medicine," *J. Clin. Sleep Med.*, vol. 8, no. 5, pp. 597–619, 2012.

[2] X. Zhang, X. Zhang, Q. Huang, Y. Lv, and F. Chen, "A review of automated sleep stage based on EEG signals," *Biocybernetics Biomed. Eng.*, vol. 44, no. 3, pp. 651–673, 2024.

[3] Y. Zhu, S. Tu, L. Zhang, and L. Xu, "Multi-source unsupervised domain-adaptation for automatic sleep staging," in *Proc. 2023 IEEE Int. Conf. Bioinf. Biomed.*, 2023, pp. 2437–2440.

[4] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Self-supervised learning for label-efficient sleep stage classification: A comprehensive evaluation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1333–1342, 2023.

[5] P. An, Z. Yuan, and J. Zhao, "Unsupervised multi-subepoch feature learning and hierarchical classification for EEG-based sleep staging," *Expert Syst. Appl.*, vol. 186, 2021, Art. no. 115759.

[6] C. Yoo, H. W. Lee, and J.-W. Kang, "Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network," *IEEE J. Biomed. health Inform.*, vol. 26, no. 3, pp. 1273–1284, Mar. 2022.

[7] Y. Zhang et al., "SHNN: A single-channel EEG sleep staging model based on semi-supervised learning," *Expert Syst. Appl.*, vol. 213, 2023, Art. no. 119288.

[8] H. Liu, H. Zhang, B. Li, X. Yu, Y. Zhang, and T. Penzel, "MsleepNet: A semi-supervision based multi-view hybrid neural network for simultaneous sleep arousal and sleep stage detection," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 4002909.

[9] C.-H. Lee, H. Kim, H.-j. Han, M.-K. Jung, B. C. Yoon, and D.-J. Kim, "NeuroNet: A novel hybrid self-supervised learning framework for sleep stage classification using single-channel eeg," 2024, *arXiv:2404.17585*.

[10] J. Li, Q. Chen, J. Pan, and H. Huang, "A novel self-supervised learning method for sleep staging and its pilot study on patients with disorder of consciousness," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 46, 2024.

[11] Y. Hu, K. Ye, H. Kim, and N. Lu, "BERT-Pin: A BERT-based framework for recovering missing data segments in time-series load profiles," *IEEE Trans. Ind. Informat.*, vol. 20, no. 10, pp. 12241–12251, Oct. 2024.

[12] B. Haoran and L. Guanze, "Semi-supervised end-to-end automatic sleep stage classification based on pseudo-label," in *Proc. 2021 IEEE Int. Conf. power Electron., Comput. Appl*, 2021, pp. 83–87.

[13] Q. Liu et al., "ActiveSleepLearner: Less annotation budget for better large-scale sleep staging," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 9, no. 2, pp. 1756–1765, Apr. 2025.

[14] I. A. Zapata, Y. Li, and P. Wen, "Rules-based and SVM-Q methods with multitapers and convolution for sleep EEG stages classification," *IEEE Access*, vol. 10, pp. 71299–71310, 2022.

[15] P. Memar and F. Faradji, "A novel multi-class EEG-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, Jan. 2018.

[16] B. Yang, W. Wu, Y. Liu, and H. Liu, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2505313.

[17] J. Phyo, W. Ko, E. Jeon, and H.-I. Suk, "TransSleep: Transitioning-aware attention-based deep neural network for sleep staging," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4500–4510, Jul. 2023.

[18] H. Phan et al., "L-SeqSleepNet: Whole-cycle long sequence modelling for automatic sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 10, pp. 4748–4757, Oct. 2023.

[19] Y. Lin, M. Wang, F. Hu, X. Cheng, and J. Xu, "Multimodal polysomnography based automatic sleep stage classification via multiview fusion network," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2504112.

[20] J. Bao et al., "A feature fusion model based on temporal convolutional network for automatic sleep staging using single-channel EEG," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 11, pp. 6641–6652, Nov. 2024.

[21] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-sleep: Resilient high-frequency sleep staging," *NPJ Digit. Med.*, vol. 4, no. 1, p. 72, 2021.

[22] Z. Jia et al., "GraphSleepNet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, vol. 2021, pp. 1324–1330.

[23] Z. Jia et al., "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1977–1986, 2021.

[24] L. Peng, Y. Ren, Z. Luan, X. Chen, X. Yang, and W. Tu, "SleepViTransformer: Patch-based sleep spectrogram transformer for automatic sleep staging," *Biomed. Signal Process. Control*, vol. 86, 2023, Art. no. 105203.

[25] 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.

[26] J. B. Stephansen et al., "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 5229.

[27] L. Fiorillo, D. Pedroncelli, V. Agostini, P. Favaro, and F. D. Faraci, "Multi-scored sleep databases: How to exploit the multiple-labels in automated sleep scoring," *Sleep*, vol. 46, no. 5, 2023, Art. no. zsad028.

[28] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura, "A survey on deep active learning: Recent advances and new frontiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 5879–5899, Apr. 2025.

[29] A. Kirsch, J. Van Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 7026–7037.

[30] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 93–102.

[31] A. Parvaneh, E. Abbasnejad, D. Teney, G. R. Haffari, A. Van Den Hengel, and J. Q. Shi, "Active learning by feature mixing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12237–12246.

[32] H. S. Hossain, S. R. Ramamurthy, M. A. A. H. Khan, and N. Roy, "An active sleep monitoring framework using wearables," *ACM Trans. Interactive Intell. Syst.*, vol. 8, no. 3, pp. 1–30, 2018.

[33] M. Macas, N. Grimova, V. Gerla, L. Lhotska, and E. Saifutdinova, "Active learning for semiautomatic sleep staging and transitional EEG segments," in *2018 IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 2621–2627.

[34] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, "A comprehensive survey on contrastive learning," *Neurocomputing*, 2024, Art. no. 128645.

[35] 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 Process. Lett.*, vol. 29, pp. 189–193, 2021.

[36] F. Shen, Z. Zhang, Y. Peng, H. Guo, L. Chen, and H. Gao, "Self-supervised learning for sleep stage classification with temporal augmentation and false negative suppression," in *Proc. 2024 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1761–1765.

[37] C. J. Reed et al., "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4088–4099.

[38] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, 2016.

[39] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen, and J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

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

[41] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[42] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, 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.

[43] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5903–5915, Sep. 2022.

[44] X. Ji, Y. Li, and P. Wen, "Jumping knowledge based spatial-temporal graph convolutional networks for automatic sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1464–1472, 2022.

[45] X. Ji, Y. Li, and P. Wen, "3DSleepNet: A multi-channel bio-signal based sleep stages classification method using deep learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3513–3523, 2023.

[46] S. Lee, Y. Yu, S. Back, H. Seo, and K. Lee, "SleePyCo: Automatic sleep scoring with feature pyramid and contrastive learning," *Expert Syst. Appl.*, vol. 240, 2024, Art. no. 122551.

[47] X. Ji, Y. Li, P. Wen, P. Barua, and U. R. Acharya, "MixSleepNet: A multi-type convolution combined sleep stage classification model," *Comput. Methods Programs Biomed.*, vol. 244, 2024, Art. no. 107992.

[48] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. 2020 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 641–644.

[49] E. Eldele et al., "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[50] 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.

[51] Y. Fang, Y. Xia, P. Chen, J. Zhang, and Y. Zhang, "A dual-stream deep neural network integrated with adaptive boosting for sleep staging," *Biomed. Signal Process. Control*, vol. 79, 2023, Art. no. 104150.

[52] C. Zhao, J. Li, and Y. Guo, "Sequence signal reconstruction based multi-task deep learning for sleep staging on single-channel EEG," *Biomed. Signal Process. Control*, vol. 88, 2024, Art. no. 105615.

[53] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, 2018.

[54] 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," *Biomed. Signal Process. control*, vol. 61, 2020, Art. no. 102037.

[55] 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.

[56] E. Eldele et al., "Time-series representation learning via temporal and contextual contrasting," 2021, *arXiv:2106.14112*.

[57] E. Eldele et al., "Self-supervised contrastive representation learning for semi-supervised time-series classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15604–15618, Dec. 2023.

[58] E. Seong, H. Lee, and D.-K. Chae, "Self-supervised framework based on subject-wise clustering for human subject time series data," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 22341–22349.