LPSGM: A Unified Flexible Large PSG Model for Sleep Staging and Mental Disorder Diagnosis

Guifeng Deng[1,2], Mengfan Niu[1], Yuxi Luo[3], Shuying Rao[1,2], Jing Sun[1,2], Junyi Xie[1], Zhenghe Yu[1], Wenjuan Liu[1], Sha Zhao[4], Gang Pan[4], Xiaojing Li[1,4], Wei Deng[1,4], Wanjun Guo[1,4], Tao Li[1,4*], Haiteng Jiang[1,4,5*]

[1]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, and Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, 310058, China.
[2]College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, 310058, China.
[3]School of Biomedical Engineering, Shenzhen Campus of Sun Yat sen University, Shenzhen, 518100, China.
[4]MOE Frontier Science Center for Brain Science and Brain-machine Integration, State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou 311121, China.
[5]NHC and CAMS Key Laboratory of Medical Neurobiology, Zhejiang University, Hangzhou 310058, China.

*Corresponding authors:
Tao Li Email: litaozjusc@zju.edu.cn
Haiteng Jiang Email: h.jiang@zju.edu.cn

Abstract

We present the Large PSG Model (LPSGM), a unified and flexible framework for sleep staging and disease diagnosis using polysomnography (PSG) data. LPSGM is designed to address the challenges of cross-center generalization in sleep staging and to enable fine-tuning for downstream disease diagnosis tasks. LPSGM introduces a unified training framework for heterogeneous datasets and allows flexible channel input adjustments during inference. The model is first trained on 220,500 hours whole-night PSG from 16 public datasets, achieving robust sleep staging performance. It is then fine-tuned on target center data for various disease classification tasks, including narcolepsy diagnosis, anxiety and depression detection, and the classification of healthy versus depressed individuals. LPSGM outperforms baseline models on both sleep staging and disease diagnosis tasks. Our results demonstrate that LPSGM not only enhances sleep staging accuracy but also improves the diagnosis of sleep-related and psychiatric disorders, showing promise for clinical applications in sleep medicine and psychiatry.

1. Introduction

Sleep accounts for roughly one-third of a person's life and the quality of sleep is fundamental to overall human health [1], [2]. Sleep disorders, such as insomnia, sleep apnea and narcolepsy, are becoming increasingly prevalent. According to the World Health Organization, the global prevalence of sleep disorders is approximately 27%. Around 50 to 70 million adults in the U.S. are affected by chronic sleep or wakefulness disorders. In China, over 300 million people suffer from sleep disorders.

In the current clinical pratice, polysomnography (PSG), which records various physiological signals such as electroencephalogram (EEG), electrooculogram (EOG), and electromyography (EMG) during a full night's sleep, is the gold standard for sleep assessment and diagnosis. Sleep staging based on PSG is fundamental for understanding sleep architecture and evaluating sleep quality. According to the guidelines established by the American Academy of Sleep Medicine (AASM) in 2007, human sleep is a dynamic process that can be divided into five stages: wakefulness (W), non-rapid eye movement Ⅰ-Ⅲ (N1-N3), and rapid eye movement (R).

Traditionally, clinicians visually score sleep stages for every 30-second epoch. However, scoring a full night's PSG, which typically spans 7 to 9 hours, requires about 2 hours of a doctor's time, making it labor-intensive and time-consuming. Additionally, the results are highly susceptible to the subjective judgment of the doctor, which can affect the accuracy of the assessment.

Over the past few decades, numerous computer-based automatic sleep staging methods have emerged, leading to significant advancements in the field. The rapid progress in deep learning, in particular, has resulted in methods that achieve performance comparable to human experts on large-scale datasets. However, the clinical adoption of these automatic sleep staging systems remains limited due to two primary challenges faced by deep learning-based approaches:

1. Data Scarcity: Deep learning-based methods require large-scale labeled datasets to achieve high accuracy and robust generalization. However, acquiring such large amounts of labeled data is prohibitively expensive and often beyond the reach of many clinical sleep centers.

2. Domain Gap: The lack of standardized protocols for PSG leads to significant variability between datasets from different centers. This variability is caused by different populations, signal acquisition conditions (such as equipment, electrode placements, channel settings, sampling rates, and signal-to-noise ratios), and the subjectivity of annotators. Due to the inherent assumption in deep learning that data are independently and identically distributed, this domain gap results in models trained on datasets from one center experiencing significant performance degradation when applied to datasets from different centers.

To address these challenges, we propose LPSGM (Large PSG Model), a unified and flexible framework for both cross-center sleep staging and disease diagnosis using PSG data. Due to the foundational and long-established nature of sleep staging, there are numerous publicly available datasets. However, the variability in montages hinders the full utilization of these public datasets. Most current studies and methods focus on training and testing within individual dataset, typically using data from specific EEG

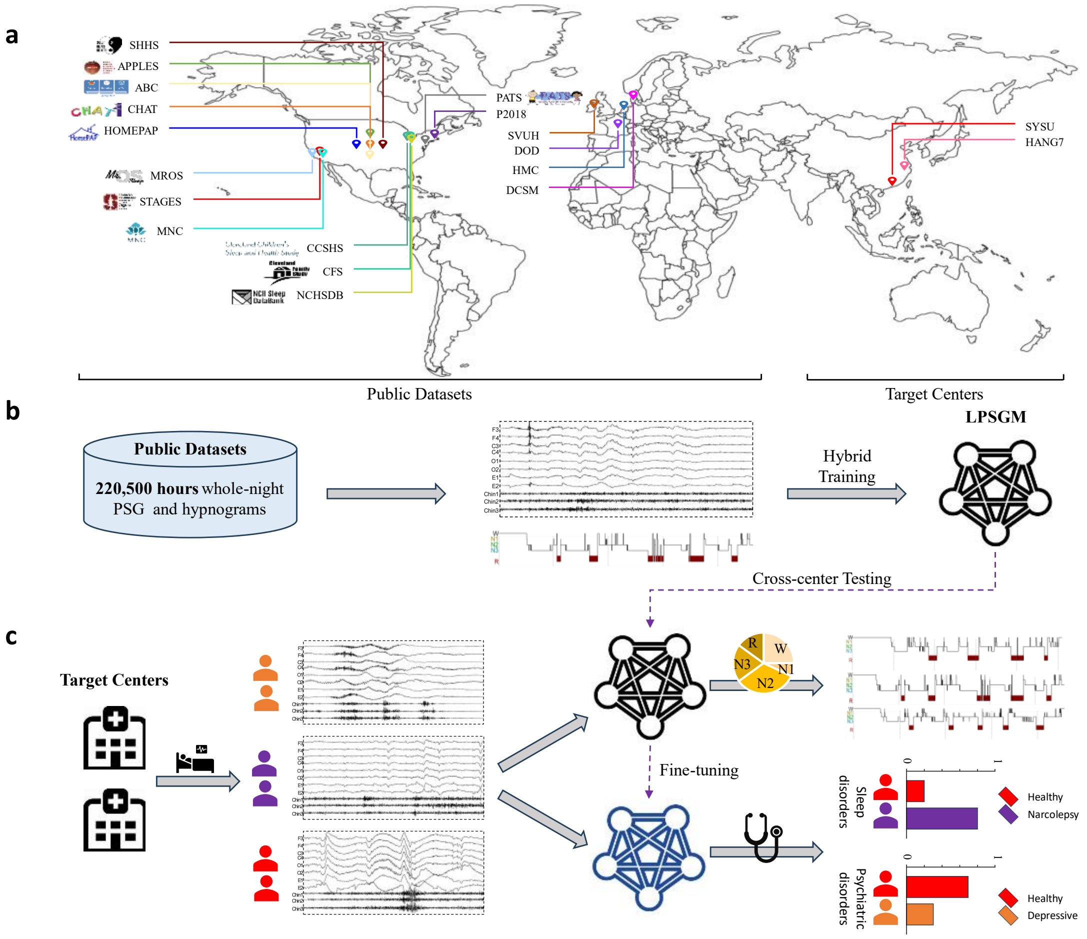


Fig. 1: Overview of the LPSGM framework for sleep staging and disorder diagnosis. (a) Geographic distribution of public and target datasets used in this study. (b) LPSGM training pipeline. (c) Cross-center testing and downstream disorder diagnosis.

channels from one or a few public datasets. No existing method has effectively integrated and utilized these public datasets to their full potential.

Our approach seeks to overcome this limitation by effectively integrating these diverse datasets, thereby enhancing the generalization and robustness of sleep staging models, and enable fine-tuning for downstream disease diagnosis tasks. Our contribution are as follows:

1. Unified Framework for Training: LPSGM utilizes a unified framework during training phase to handle datasets with different channel configurations. Consequently, LPSGM can effectively integrate and leverage multiple large-scale, multi-center public datasets, thereby approximating the true distribution and bridging the domain gap between different centers.

2. Flexibility in Inference: LPSGM demonstrates significant flexibility during the inference phase. By adjusting the number of input channels, LPSGM can balance accuracy and inference speed without altering the model structure. This adaptability makes the model suitable for various application scenarios.

3. Cross-Center Generalization: Our experimental results validate LPSGM's effectiveness in cross-center generalization. We conducted hybrid training on 16 public datasets and tested the model on 2 target center datasets. The results show that LPSGM achieves comparable accuracy to fully supervised training, even when applied to previously unseen datasets, showcasing its plug-and-play capability and potential for practical clinical application.

4. Fine-Tuning for Disease Diagnosis: LPSGM can be fine-tuned for disease diagnosis tasks, such as narcolepsy, depression, and anxiety, based on PSG data. Fine-tuning on disease-specific datasets after pretraining on large-scale public datasets improves diagnostic accuracy, demonstrating LPSGM's versatility in clinical diagnostics beyond sleep staging.

2. Related Work

2.1 Automatic Sleep Staging

Traditional Machine Learning. Early approaches for automatic sleep staging primarily relied on hand-crafted features derived from the time[3], [4], frequency[5], [6], and nonlinear domains[7]. These features were typically extracted based on expert knowledge and were used to train traditional machine learning models such as support vector machines[7], decision trees, random forests[6], [8], and neural networks[9]. While these methods laid the groundwork for automated sleep staging, they suffered from several limitations. Firstly, the feature extraction process was labor-intensive and prone to bias, as it heavily depended on the specific expertise of the researchers. Secondly, the performance of these models was often dataset-specific, limiting their generalizability across different datasets. Lastly, these models lacked the capacity to capture complex patterns in the data, which are crucial for accurate sleep stage classification.

Deep Learning. The development of deep learning brought about a significant shift in automatic sleep staging. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have emerged as the most widely adopted frameworks for sleep staging models, primarily due to their ability to effectively learn hierarchical features from raw physiological signals[10], [11], [12], [13]. CNNs are commonly used to extract spatial features from EEG signals by capturing local patterns, while RNNs are utilized to model the temporal dependencies across consecutive epochs to effectively handle the sequential nature of sleep data. By leveraging large-scale datasets, several deep learning methods have achieved performance levels comparable to those of human experts, typically through a sequence-to-sequence framework. However, a common limitation of these models is their reliance on intra-dataset scheme, where both training and testing data are from the same dataset. This approach overlooks the domain gap that emerge when applying the model to data from different sources, leading to poor generalization to unseen datasets[14], [15].

Transfer Learning and Domain Adaptation. Given the challenges posed by domain gap, recent research has explored transfer learning and domain adaptation techniques[15], [16]. Some methods[14], [17] employ adversarial training to align the feature distributions between the source and target domains, aiming to learn domain-invariant features. Others[18] utilize consistency constraints to enforce the extraction of shared features across domains. These techniques have demonstrated promising results in enhancing model robustness across diverse datasets. However, existing methods fall short in fully exploiting the wealth of available public datasets due to differences in montage setups, signal acquisition protocols, and annotation standards. Additionally, no existing cross-center transfer learning method has achieved comparable performance with models trained directly on target center data.

2.2 EEG-based Large Models

Inspired by the unprecedented success of large language models (LLMs) in natural language processing tasks[19], [20], [21], recent efforts have explored the application of large models to EEG data. From a data perspective, although there is an abundance of EEG datasets available, individual datasets typically have relatively small sample sizes. This highlights the necessity for large models to effectively integrate and harness the potential of this diverse data. However, although the international 10-20 system provides a standard for EEG recording, users opt to collect data with different electrode and channel configurations tailored to specific applications.

To address the challenge of diverse channel configurations across EEG datasets, several strategies have been proposed. Mohsenvand et al.[22] tackled this issue by decomposing all datasets into individual-channel data and utilizing a single-channel model to independently process each channel. Han et al.[23] introduced Graph Convolutional Networks (GCNs) to model the relationships among different channels, allowing the model to adapt to diverse configurations. Gu et al.[24] designed distinct input modules tailored to each dataset's specific configuration, facilitating customized processing. With the rise of Transformer-based models across various domains, Li et al.[25] proposed a method that integrates spatial positional encoding within Transformers to manage different channel inputs. Building on this, Yi et al.[26] further divided the brain into 17 regions, treating each channel and region as distinct tokens embedded within a Transformer. However, both methods rely on frequency-domain features as input, often at the expense of capturing the temporal dynamics intrinsic to EEG signals. Jiang et al. advanced this research by proposing LaBraM[27], which directly processes raw EEG signals. LaBraM segments multi-channel EEG signals into patches, similar to the approach used in ViT[28], and embeds both spatial and temporal information of patches into Transformer through an additive operation.

In this paper, we build upon LaBraM[27] by refining its channel and temporal encoding techniques and introducing padding and masking operations. These enhancements allow the Transformer to more effectively handle batches with varying sequence lengths, thereby increasing the model's flexibility and applicability across diverse PSG datasets.

3. Methods

3.1 Overview

In this section, we detail the whole framework of LPSGM. As illustrated in Fig. 2, LPSGM is a sequence-to-sequence model that takes multi-channel PSG signals as input and output multiple sleep stages. LPSGM consists of three components: epoch encoder, sequence encoder and classifier. The epoch encoder is designed to extract local intra-epoch features and the sequence encoder is designed to capture global inter-epoch features among sequential epochs. Finally, the classifier utilizes the encoded features to classify sleep stages and outputs the predicted probabilities for each sleep stage.

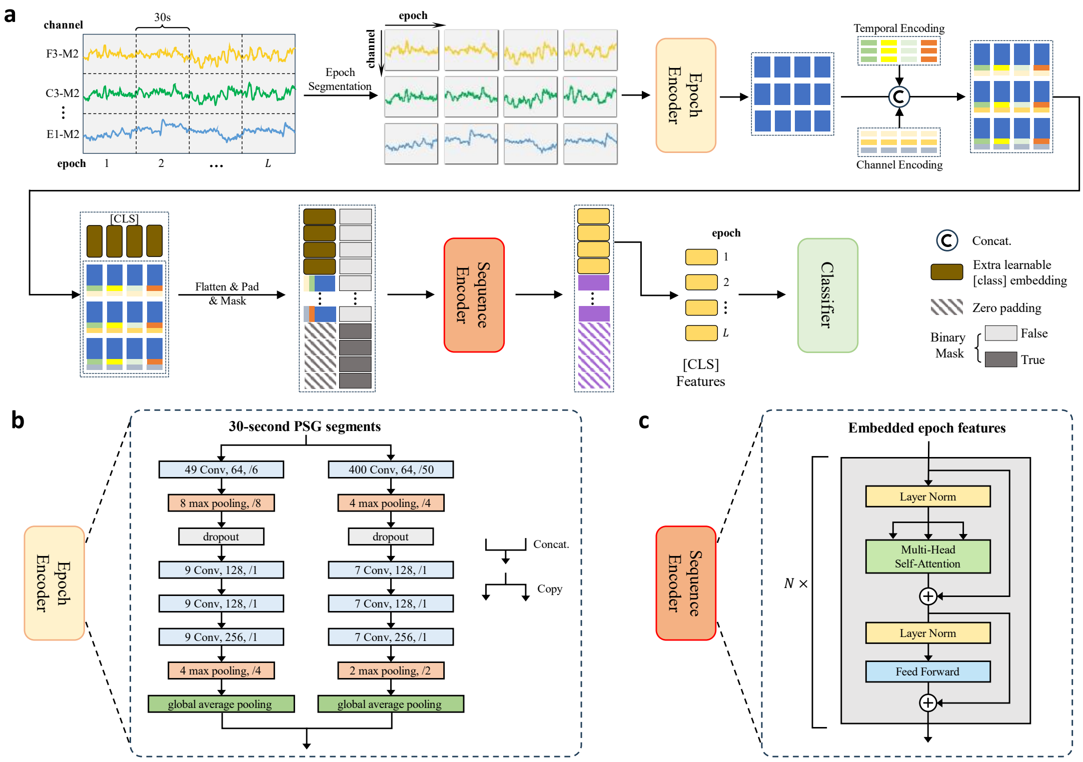


Fig. 2: (a) The overall architecture of LPSGM, which integrates the Epoch Encoder, Sequence Encoder, and Classifier for both sleep staging and disorder diagnosis tasks. (b) The Epoch Encoder uses a dual-branch CNN to extract local features from each 30-second epoch of PSG data. (c) The Sequence Encoder consists of a series of N Transformer blocks to capture temporal dependencies across epochs in the sleep sequence.

3.2 Epoch Segmentation

To process data with varying numbers of channels in a unified model, we first perform epoch segmentation on the input multi-channel PSG signals. Assume the input multi-channel PSG signal $x_i = \{x_i^1, x_i^2, \dots, x_i^L\}$ is a sleep sequence composed of L sleep epochs, where $x_i^l \in \mathbb{R}^{C_i \times n}$ represents the l-th epoch of the sleep sequence with C_i channels, and $n = 30 \times f_s$ denotes the number of samples in a 30-second epoch. We start by segmenting x_i along the temporal dimension and channel dimension to obtain $L \times C_i$ segments $x_i^{l,c} \in R^n$. Each segment represents a single 30-second epoch from one channel. This segmentation process is akin to the patch operation used in other models based on Transformer architectures, where the input data is divided into smaller, fixed-size patches for processing. By performing epoch segmentation, we standardize the input format, facilitating the subsequent encoding and classification steps within the LPSGM framework.

3.3 Epoch Encoder

The segments obtained from the epoch segmentation are single-channel, single-epoch data blocks $x_i^{l,c} \in \mathbb{R}^n$, which are then fed into the epoch encoder. The role of the epoch encoder is to extract time-invariant features from each of segments. Following DeepSleepNet[10], we employ a dual-branch CNN with small and large filter sizes to capture various frequency features. The smaller filter is better to capture high-frequency features while the larger filter is better to capture low-frequency features.

In epoch encoder, each branch is composed of four convolutional layers, two max pooling layers, and one global average pooling layer. Each convolutional layer sequentially performs 1D convolution, batch normalization, and ReLU activation. The max pooling layers downsample inputs using the max operation to reduce feature sizes. The global average pooling layer applies the average operation and aggregates the features along the temporal dimension into a single feature vector. Finally, the outputs of both branches are concatenated along the feature dimention, resulting in a 512-dimensional feature $e_i^{l,c} \in \mathbb{R}^d$ $(d = 512)$. Details on the filter sizes, number of filters, stride sizes, and pooling sizes are provided in Fig. 2b.

3.4 Input Preparation for Sequence Encoder

To ensure the Transformer model can effectively process the input data, we perform several preprocessing steps: channel and temporal encoding, padding and masking, and the insertion of CLS tokens. These steps prepare the input sequences by embedding positional information, standardizing sequence lengths, and providing a unified representation for classification.

Channel & Temporal Encoding

Unlike Recurrent Neural Networks (RNNs) that process sequences sequentially, the Transformer model employs self-attention mechanisms to handle the entire input sequence simultaneously, lacking an inherent mechanism to capture positional information within the sequence. To enable the model to be aware of the temporal and channel information of the segments embedding, we adopt the method of channel and temporal encoding inspired by LaBraM[27]. However, we replace addition operation with concatenation to avoid blending information from different dimensions and better preserve their distinctiveness.

Specifically, we maintain two embedding lists: a channel embedding list $CE =$

$\{ce_1, ce_2, \ldots, ce_{|C|}\}$ and a temporal embedding list $TE = \{te_1, te_2, \ldots, te_L\}$. Each embedding vector in CE and TE has its own specific dimension, denoted as d_{ce} and d_{te}, respectively. For each feature vector $e_i^{l,c}$ produced by epoch encoder, we concatenate the corresponding channel and temporal embeddings based on its channel c and temporal position l within the sequence. This results in an embedded feature vector $\widetilde{e_i^{l,c}} \in \mathbb{R}^{d+d_{ce}+d_{te}}$, described by the following equation:

$$\widetilde{e_i^{l,c}} = e_i^{l,c} \oplus ce_c \oplus te_l, \qquad c = 1,2,\ldots,|C|; \;\; l = 1,2,\ldots,L \tag{1}$$

where $\oplus$ denotes the concatenation operation along the feature dimension, $e_i^{l,c}$ is the original feature vector from the epoch encoder, ce_c is the channel embedding from the channel embedding list CE, and te_l is the temporal embedding from the temporal embedding list TE. The channel and temporal embedding list CE and TE are learnable parameters of the model, optimized during the training process.

Padding & Masking

In cross-dataset hybrid training, the input sequences have a fixed temporal length (L epochs) but vary in the number of channels (C_i). After segmentation, encoding and unfolding, we obtain feature sequences of length $L \times C_i$, as shown in Fig. 3. The variability in sequence lengths presents challenges for batch training due to hardware limitations that prevent the processing of sequences with varying lengths within the same batch.

LaBraM addresses this issue by setting $C \times T$ to a fixed value: for datasets with more channels, the temporal length is shortened, and for datasets with fewer channels, the temporal length is extended. However, this approach has two significant drawbacks. Firstly, some long-term dependencies might only appear in datasets with fewer channels, causing the model to overfit to these datasets and learn dataset-specific features instead of generalizable features. Secondly, the model becomes inflexible and unable to handle datasets with the same temporal length but different numbers of channels.

To overcome these limitations, we drew inspiration from Natural Language Processing (NLP) techniques and designed a padding and masking strategy. As illustrated in Fig. 3b, for each batch of B embedded features with C_i channels and temporal length T_i, we unfold the features into sequences of size $C_i \times T_i$. These sequences are then padded with zero vectors to the maximum sequence length, resulting in a feature sequence of size $B \times (C_{max} \times T_{max})$, where C_{max} and T_{max} are the maximum number of channels and temporal length within the batch. Simultaneously, we generate a binary mask $M \in \{0,1\}^{B\times(C_{max}\times T_{max})}$ to indicate the padded positions, given by:

$$M_{i,j} = \begin{cases} 1, & j > C_i \times T_i \\ 0, & j \le C_i \times T_i \end{cases} \tag{2}$$

In this formulation, $M_{ij} = 1$ indicates that the j-th element of the i-th sequence is padding while $M_{ij} = 0$ indicates valid data. This binary mask serves two primary purposes during training: firstly, as an attention mask, it ensures that the Transformer model focuses solely on valid data and disregards padded elements when computing attention; secondly, during loss computation, it guarantees that only the loss from real data is calculated by excluding the loss from padded elements. Consequently, this prevents the model from learning spurious patterns derived from padding. In LPSGM, we set T_i to a fixed value L, meaning each sequence consists of L 30-second epochs.

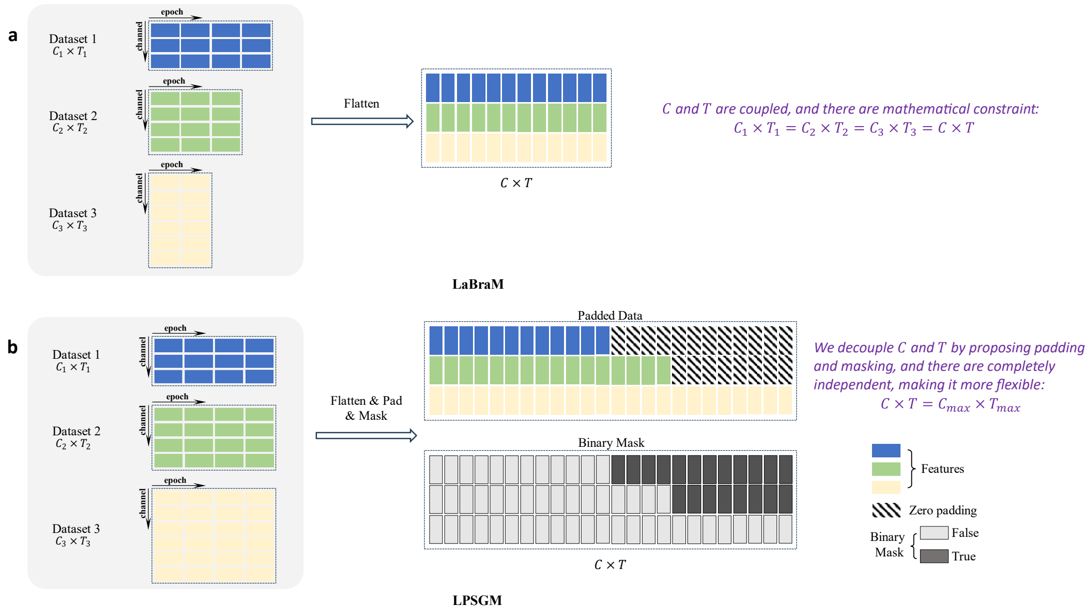


Fig. 3: Illustration of the comparison between (a) LaBraM and (b) LPSGM in handling datasets with varying numbers of channels.

Insertion of CLS token

In order to facilitate the generation of unified representation for sequences of varying length, we introduce the CLS token, a technique commonly used in Transformer-based architectures. Specifically, we prepend L learnable CLS tokens to the beginning of each feature sequence $\tilde{e_i}$. Each CLS token possesses the same dimension as the encoded feature vectors (i.e., $d + d_{ce} + d_{te}$), resulting in an extended sequence $\tilde{e_i}'$ of length $L \times (C_i + 1)$. The CLS tokens are initialized randomly and undergo learning during training.

3.5 Sequence Encoder

The role of the sleep sequence encoder is to encode the padded and masked feature sequences, including the added CLS tokens, in order to extract features from the sleep sequence that consists of multiple channels and epochs. The sequence encoder is based on the Transformer architecture, which comprises N Transformer blocks. Each Transformer block includes a multi-head self-attention layer (MSA) and a feed-forward network (FFN), with layer normalization (LN) applied between them.

Let E_0 represent the input feature sequence and E_{out} represent the output feature sequence. The encoding process for each Transformer block can be described as follows:

$$E'_\ell = \mathrm{MSA}\big(\mathrm{LN}(E_{\ell-1})\big) + E_{\ell-1}, \quad \ell = 1, \ldots, N$$

$$E_\ell = \mathrm{FFN}\big(\mathrm{LN}(E'_\ell)\big) + E'_\ell, \quad \ell = 1, \ldots, N$$

$$E_{out} = \mathrm{LN}(E_N) \tag{3}$$

The output sequence E_{out} retains the same shape as the input sequence. The first L features of the sequence, corresponding to the CLS tokens, are then used as the classification features E_{cls}:

$$E_{cls} = E_{out}[0{:}\,L] \tag{4}$$

3.6 Classifier

We design two classifiers: one for sleep staging and another for disease diagnosis. Each classifier consists of a fully connected layer with a softmax function.

Sleep Staging Classifier

The sleep staging classifier is a five-class classifier responsible for assigning each epoch to one of the five sleep stages. It takes E_{cls} as input and outputs the probabilities for the five sleep stages corresponding to each of the L epochs.

Disease Diagnosis Classifier

The disease diagnosis classifier is a binary classifier designed to predict the presence or absence of a specific disorder, such as narcolepsy or depression, based on the sleep staging sequence. This classifier takes the average feature vector of the sequence, E_{cls}, computed by averaging the features across the L epochs, as input, and outputs the probability indicating the likelihood of the presence or absence of the disorder.

4. Experiments

4.1 Implementation Details

Hybrid Training and Cross-Center Testing on Sleep Staging Task

We conduct hybrid training on source domain consisting of 16 public datasets and evaluate the performance of our approach on 2 target domain datasets. We create a validation set by stratified random sampling of 10% from each public dataset. The model is evaluated on this validation set after every epoch, and the best performing model parameters are saved. The length of the sleep sequence L is set to 20, meaning that the context length of the considered sequence is 10 minutes. The feature dimension d is set to 512. The dimension of channel encoding d_{ce} and temporal encoding d_{te} are set to 32 and 64, respectively. The number of Transformer Block is set to 4, the number of heads h is 8, and the dimension of feed-forward network is 608.

We use the weighted cross-entropy (WCE) function as the loss function for the sleep staging task:

$$\mathcal{L}_{classify} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_k \cdot y_i^k \log \hat{y}_i^k \tag{5}$$

where y_i^k denotes the probability that x_i actually belongs to the k-th stage, and $\hat{y}_i^k$ denotes the probability that x_i is predicted to the k-th stage. The category weight w_k is set to the normalized value of the reciprocal of the k-th stage proportion in the training set.

We implemented LPSGM based on PyTorch. The model is trained using Adam optimizer with default setting, the weight decay is set to 1e-3 and the mini-batch size is set to 256. The training epochs is 50, with a warm-up phase of 15 epochs. The learning rate undergoes linear increase from 0 to 1e-4 during the warm-up phase, followed by decay according to a cosine annealing strategy to 1e-6. During training, we employed a data augmentation technique that randomly drops channels to prevent the model from over-relying on specific channels or too many channels. Specifically, each channel in a sample x_i with C_i channels has a 50% chance of being dropped. The model was trained on one machine with Intel Xeon Gold 6330 CPU and two NVIDIA A800 GPUs.

We evaluated our method using several performance metrics, including accuracy, macro-F1 and Cohen's Kappa.

Fine-tuning for Disorder Diagnosis Tasks

In addition to the sleep staging task, we further fine-tune the model for disorder diagnosis tasks, using the cross-entropy loss function. We investigate three fine-tuning paradigms: partial fine-tuning, full fine-tuning, and joint fine-tuning, and compare them against a model trained from scratch. Partial fine-tuning: In this paradigm, we freeze all modules of the pre-trained sleep staging model except for the classifier, which is modified and fine-tuned specifically for the disorder diagnosis task. Full fine-tuning: In this paradigm, both the classifier and the entire model are fine-tuned for the disorder diagnosis task. Joint fine-tuning: This paradigm simultaneously fine-tunes both the sleep staging task and the disorder diagnosis task in parallel.

The training procedure for all fine-tuning paradigms consists of 10 epochs, whereas training from scratch involves 30 epochs. In all four experimental paradigms, the initial learning rate lr_0 for the disorder diagnosis classifier is set to 1e-3, while for all other modules, the initial learning rate is set to 1e-5. The AdamW optimizer is used, and the learning rate is adjusted using a cosine annealing strategy.

The performance of the disorder diagnosis task is evaluated using multiple metrics: accuracy, F1-score, sensitivity, specificity, and balanced accuracy.

4.2 Baselines for Sleep Staging

To evaluate the effectiveness of our method in sleep staging, we conducted two baseline experiments, each designed to establish the performance boundaries (lower and upper bounds) of our approach. These baselines allow us to compare our method against standard practices in sleep staging, particularly in the context of cross-center data adaptation.

Baseline 1: Direct Application of Models Trained on Other Centers

This baseline represents the lower bound of our method's performance by simulating a scenario where a model trained on datasets from other centers is directly applied to the target center's data without any fine-tuning or domain adaptation. Specifically, we trained the model on the entire dataset from one center and tested it on the dataset from another center. All available channels were used during training and testing.

Baseline 2: Fully Supervised Training on Target Center Data

This baseline establishes the upper bound of our method's performance by simulating an ideal scenario where a model is fully trained from scratch on the target center's data. We conducted rigorous five-fold cross-validation on the target center's data. All available channels were used during training and testing. This scenario serves as a baseline for the maximum achievable performance under fully supervised conditions.

To ensure a fair comparison and prevent potential overfitting of single-center data in the LPSGM model, we used two versions of our model: LPSGM-Large and LPSGM-Small. The LPSGM-Large model, which is the same as the one used in the main experiments, consists of 4 Transformer blocks. In contrast, the LPSGM-Small model reduces the number of Transformer blocks to 1 while keeping all other aspects unchanged. This configuration allows for a more accurate evaluation of baselines and safeguards against biased results caused by overfitting.

5. Results and Analysis

5.1 Cross-Center Classification Performance

The cross-center classification results on HANG7 and SYSU datasets compared with baselines are presented in Table 1. The baseline experiments were conducted using both LPSGM-Large and LPSGM-Small models, revealing consistent superior performance of the LPSGM-Small model over the LPSGM-Large model across Baseline 1 and Baseline 2. Consequently, the LPSGM-Small model was chosen as the official baseline for further comparative analysis. Baseline 1 and Baseline 2 serve as the lower and upper bounds for LPSGM's cross-center classification performance, respectively. To quantify the generalization performance, we calculated the percentage of LPSGM's metrics relative to Baseline 2, which reflects its suitability for seamless deployment across different clinical centers. On the HANG7 dataset, LPSGM achieved 85.68% accuracy, 82.88% macro-F1, and 0.8138 kappa, representing 99.6%, 99.9%, and 99.5% of Baseline 2's performance. On the SYSU dataset, LPSGM reached 84.13% accuracy, 77.88% macro-F1, and 0.7789 kappa, equivalent to 97.1%, 96.7%, and 95.7% of Baseline 2. These results demonstrate that LPSGM effectively generalizes across diverse datasets, achieving performance comparable to models specifically trained on target center data.

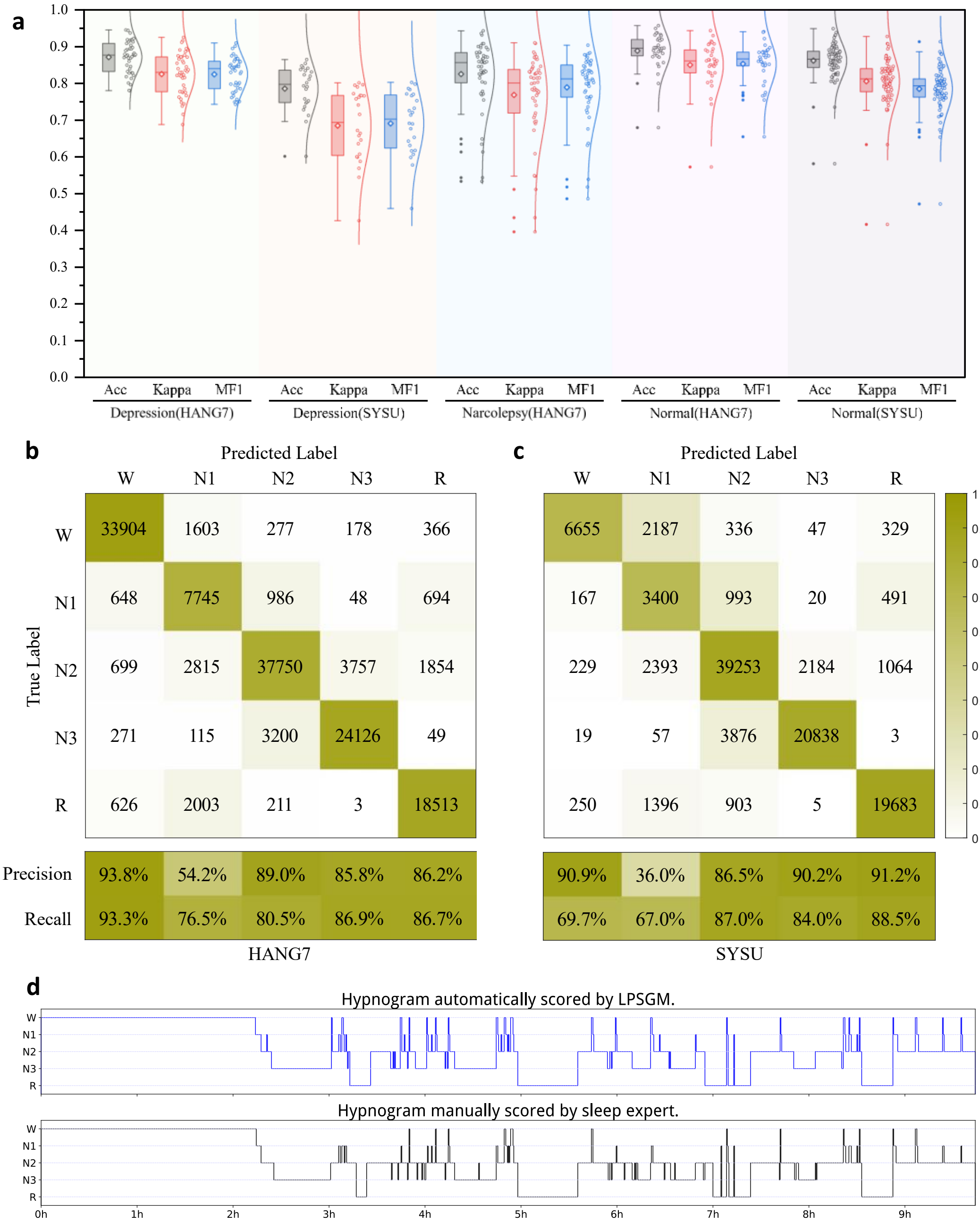
Table 1: Comparison of the sleep staging results between LPSGM and baselines on two target centers.

Center			Acc	MF1	Kappa	Per-class F1				
						W	N1	N2	N3	R
HANG7	Base. 1	LPSGM	0.7757	0.7313	0.7098	0.8420	0.4809	0.7870	0.8085	0.7383
		LPSGM-Small*	0.7840	0.7338	0.7223	0.8657	0.4430	0.7762	0.8176	0.7667
		LPSGM	0.8568	0.8288	0.8138	0.9356	0.6348	0.8455	0.8636	0.8644
	Base. 2	LPSGM	0.8493	0.8180	0.8039	0.9231	0.5996	0.8415	0.8643	0.8612
		LPSGM-Small*	0.8604	0.8300	0.8177	0.9321	0.6344	0.8545	0.8650	0.8641
		Rel. Base. 2	99.6%	99.9%	99.5%	100.4%	100.1%	98.9%	99.8%	100.0%
SYSU	Base. 1	LPSGM	0.7504	0.6821	0.6623	0.6648	0.4154	0.8019	0.7486	0.7497
		LPSGM-Small*	0.7612	0.6866	0.6666	0.6432	0.3808	0.8266	0.7972	0.7851
		LPSGM	0.8413	0.7788	0.7789	0.7888	0.4688	0.8676	0.8703	0.8986
	Base. 2	LPSGM	0.8568	0.7911	0.8012	0.8055	0.4833	0.8795	0.8965	0.8906
		LPSGM-Small*	0.8661	0.8057	0.8138	0.8314	0.5137	0.8862	0.8997	0.8975
		Rel. Base. 2	97.1%	96.7%	95.7%	94.9%	91.3%	97.9%	96.7%	100.1%

On the base of the above results, we further evaluated the performance of LPSGM across different subjects groups in HANG7 and SYSU datasets, as shown in Table 2. Notably, the model achieves its highest performance on normal subjects, with an accuracy of 88.99% and 86.15% on HANG7 and SYSU, respectively. However, a slight decrease in performance is observed when applied to subjects with depression and narcolepsy. On the HANG7 dataset, the model achieves 87.11% accuracy in the depression group and 82.45% in the narcolepsy group. Similarly, on the SYSU dataset, the accuracy drops to 78.36% for the depression group. The reduced performance may be attributed to the more complex or atypical sleep patterns often observed in individuals with these conditions.

Table 2: Performance of LPSGM across different subject groups in HANG7 and SYSU datasets.

Center	Group	Acc	MF1	Kappa
HANG7	Normal	0.8899	0.8624	0.8546
	Depression	0.8711	0.8330	0.8305
	Narcolepsy	0.8245	0.8043	0.7749
SYSU	Normal	0.8615	0.7989	0.8070
	Depression	0.7836	0.7224	0.6980



HANG7

True Label \ Predicted Label	W	N1	N2	N3	R
W	33904	1603	277	178	366
N1	648	7745	986	48	694
N2	699	2815	37750	3757	1854
N3	271	115	3200	24126	49
R	626	2003	211	3	18513
Precision	93.8%	54.2%	89.0%	85.8%	86.2%
Recall	93.3%	76.5%	80.5%	86.9%	86.7%

SYSU

True Label \ Predicted Label	W	N1	N2	N3	R
W	6655	2187	336	47	329
N1	167	3400	993	20	491
N2	229	2393	39253	2184	1064
N3	19	57	3876	20838	3
R	250	1396	903	5	19683
Precision	90.9%	36.0%	86.5%	90.2%	91.2%
Recall	69.7%	67.0%	87.0%	84.0%	88.5%

Fig. 4: Performance of LPSGM on cross-center sleep staging. (a) Boxplot distributions of accuracy (Acc), macro-F1 (MF1), and Cohen's Kappa (Kappa) for different subject groups in the HANG7 and SYSU datasets. Confusion matrix of LPSGM's sleep staging performance on the (b) HANG7 and (c) SYSU dataset. (d) Hypnogram for a full-night PSG recording from the HANG7 dataset. The upper plot shows the sleep stages predicted by LPSGM, while the lower plot shows the manual sleep staging by an expert.

During the inference stage, LPSGM can achieve an optimal balance between accuracy and speed by adjusting the number of input channels without altering the model structure. Fig. 5 demonstrates the trade-off between performance metrics and inference time per recording across different EEG channel configurations (8C, 4C, 2C, 1C) obtained from the HANG7 dataset. In each metric, we consistently observe that reducing the number of channels significantly decreases inference time but also leads to a gradual decline in performance metrics. For instance, employing 8 channels (8C) yields superior performance with an accuracy of 85.68%, macro-F1 of 82.88%, and Kappa of 81.38%, albeit with a longer inference time. Conversely, decreasing the channel count to just one (1C) substantially reduces inference time but results in lower accuracy (82.03%), macro-F1 score (79.21%), and Kappa value (76.76%). These findings indicate that LPSGM can adapt to various computational resource constraints by adjusting the number of input channels accordingly. When high-performance requirements are essential, utilizing more channels is advantageous; however, for faster inference on resource-constrained platforms, fewer channels may be preferred. This adaptability makes LPSGM a versatile solution for deployment across diverse environments with varying resource capabilities.

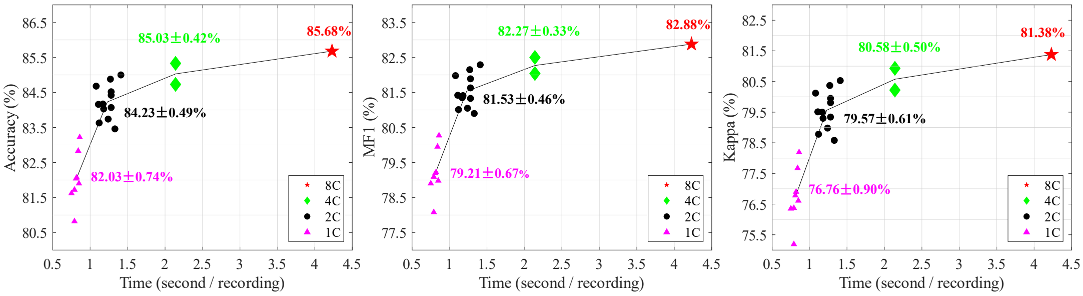


Fig. 5: Trade-off between performance metrics and inference time with different channel configurations on HANG7 dataset.

In addition to cross-center generalization, we investigate the impact of large-scale hybrid pretraining on sleep staging performance using the HANG7 and SYSU datasets. Fig. 6 shows the results of five-fold cross-validation comparing the performance of LPSGM with and without pretraining. The pretrained model, initially trained on large-scale public datasets through hybrid training, significantly outperforms the non-pretrained model when fine-tuned on target center data. On the HANG7 dataset, pretraining led to significant improvements in all metrics, with accuracy increasing by 1.27%, macro-F1 by 1.57%, and Cohen's Kappa by 1.62%. On the SYSU dataset, pretraining also resulted in notable improvements, with accuracy improving by 2.05%, macro-F1 by 1.91%, and Cohen's Kappa by 2.67%. These results highlight the effectiveness of using the pretrained LPSGM model for fine-tuning, leading to faster convergence and better performance than training from scratch.

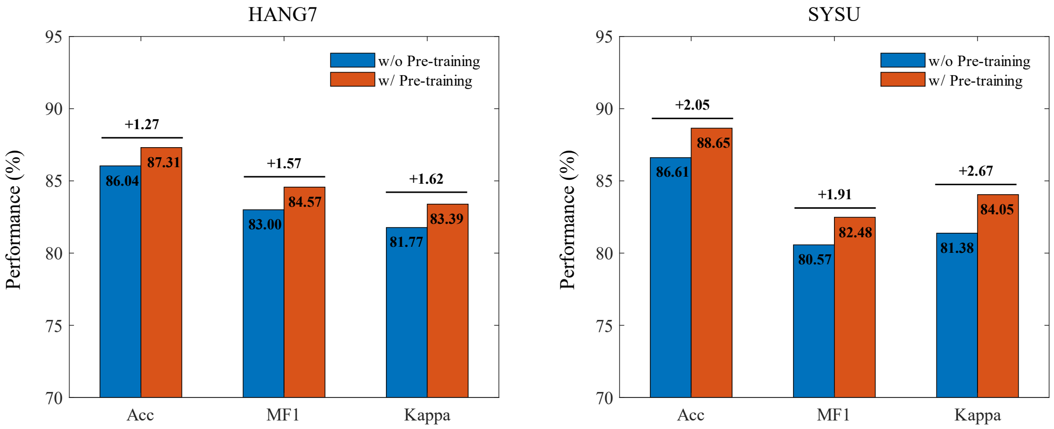


Fig. 6: Comparison of sleep staging performance on the HANG7 and SYSU datasets with and without pretraining.

5.2 Comparison with State-of-the-Art Methods

To the best of our knowledge, LPSGM has been trained on the largest collection of PSG data among all existing sleep staging models, which significantly contributes to its superior generalization ability. We conducted a comparative evaluation of LPSGM against several state-of-the-art methods, ensuring fairness by using datasets of the same scale. These methods are categorized into non-transfer-based and transfer-based approaches based on training and evaluation strategies as described in their original papers. The non-transfer-based methods include: **DeepSleepNet**[10], a classical CNN-BiLSTM network for extracting local features and learning transition rules; **TinySleepNet**[11], a classical model based on CNN and RNN with fewer model parameters; **U-Time**[29], a fully-CNN encoder-decoder architecture for time series segmentation applied to sleep staging. **AttnSleep**[30], composed of a multi-resolution CNN and multi-head self-attention with causal convolutions. The transfer-based methods include: **SleepDG**[18], composed of a combination of CNN and Transformer architectures that incorporates a proposed multi-level feature alignment technique to extract domain-invariant features; **RobustSleepNet**[16] employs attentive channel recombination to handle inputs with varying channel numbers and is trained using datasets from diverse configurations. Notably, except for RobustSleepNet, these methods lack input channel flexibility. Therefore, we limited the input channels across all datasets to overlapping channels C3-M2 and E1-M2. For RobustSleepNet and our LPSGM, experiments were conducted using both 2-channel (2C) and full 8-channel (8C) inputs.

We implemented these methods based on their public code and default hyper-parameters while adhering to our cross-center settings. The comparative results are shown in Table 3. LPSGM (8C) achieves state-of-the-art performance across all metrics, surpassing both LPSGM (2C) and other existing methods. This highlights the advantage of utilizing more EEG channels, as they provide complementary information that enhances sleep staging performance. Even when restricted to only two channels, LPSGM (2C) consistently outperformed other methods, showcasing the superior effectiveness of the Transformer-based channel encoding and attention mechanisms compared to the conventional approach of stacking channels followed by convolution in extracting and integrating multi-channel features. In non-transfer-based methods, a clear positive correlation between model size and performance was observed, suggesting that the model's capacity may be a bottleneck in our large-scale hybrid training setup. With sufficient data, larger models potentially learn more abstract and complex features, which could lead to enhanced performance. Among the transfer-based methods, SleepDG achieves performance closest to LPSGM (2C). Although RobustSleepNet is designed to adapt to different numbers of input channels, its compact model size appears to limit its capacity. This limitation is evident in the noticeable performance discrepancies between the HANG7 and SYSU datasets, suggesting that the 0.2M parameter size may cause the model to learn features that are too dataset-specific and lack broader generalizability.

Table 3: Performance comparison with existing sleep staging methods.

Methods	Model Size	Transfer-based	HANG7			SYSU		
			Acc	MF1	Kappa	Acc	MF1	Kappa
DeepSleepNet	21M	×	0.8223	0.7779	0.7643	0.8053	0.7323	0.7174
TinySleepNet	1.3M	×	0.8132	0.7658	0.7534	0.7795	0.7055	0.6832
U-Time	1.2M	×	0.8061	0.763	0.7439	0.7897	0.7191	0.6998
AttnSleep	0.6M	×	0.7353	0.7079	0.6638	0.7325	0.6644	0.6376
SleepDG	6.5M	√	0.8285	0.7836	0.7725	0.8059	0.7461	0.7331
RobustSleepNet(2C)	0.2M	√	0.7291	0.7087	0.6556	0.8058	0.7397	0.7322
RobustSleepNet(8C)			0.7362	0.714	0.664	0.7942	0.7325	0.7174
LPSGM(2C)	9.9M	√	0.8365	0.8062	0.7865	0.8279	0.7514	0.7525
LPSGM(8C)			**0.8568**	**0.8288**	**0.8138**	**0.8413**	**0.7788**	**0.7789**

5.3 Ablation Study

To verify the effectiveness of the proposed concatenation-based channel and temporal encoding as well as the CLS tokens-based feature fusion methods, we conduct an ablation study. We compare three variants of channel and temporal encoding: no encoding, addition-based encoding and concatenation-based encoding. Additionally, we compare two feature fusion methods: one that averages features across channels and the other that leverages the CLS tokens. As shown in Table 4, the combination of concatenation-based encoding and CLS tokens (ours) achieves the best performance across all metrics.

Among the different channel and temporal encoding methods, the model without encoding yields the worst performance, while the addition-based encoding provides moderate improvement. However, it still fall short of the performance achieved by concatenation-based encoding. This suggests that both channel and temporal information are crucial to the Transformer-based sequence encoder. The inferior performance of addition-based encoding may be attributed to its tendency to conflate spatial and temporal dimensions, leading to information loss. In contrast, concatenation allows the model to better preserve and utilize information from these distinct dimensions, resulting in improved overall accuracy.

The superiority of the CLS tokens-based feature fusion method over the channel averaging approach lies in its ability to address the limitations of simple averaging. Channel averaging tends to obscure the differences in importance between features from different channels, treating them as equally significant. In contrast, the CLS tokens enables a more sophisticated, nonlinear fusion of channel features, allowing the model to adaptively weigh and integrate them based on their individual relevance.

Table 4: Performance comparison of different channel & temporal encoding and feature fusion methods.

Channel & Temporal Encoding			Feature Fusion		HANG7			SYSU		
None	Add	Concat	Average	CLS	Acc	MF1	Kappa	Acc	MF1	Kappa
		√		√	**0.8568**	**0.8288**	**0.8138**	**0.8413**	**0.7788**	**0.7789**
		√	√		0.8416	0.8128	0.7950	0.8267	0.7572	0.7603
	√			√	0.8373	0.8093	0.7898	0.8252	0.7638	0.7601
√				√	0.8043	0.7655	0.7469	0.8045	0.7048	0.7252

Additionally, an experiment was conducted to investigate the impact of varying the number of Transformer blocks (N) used as a sequence encoder. The results are presented in Table 5. As the count of Transformer blocks increased from 1 to 6, a consistent improvement in HANG7 performance was observed; however, SYSU performance initially improved and then declined, reaching its peak at $N = 5$. Considering the increased computational demands and memory consumption associated with augmenting the number of Transformer blocks, we selected $N = 4$ as our default experimental configuration.

Table 5: Performance comparison of different number of Transformer block.

Transformer Block (N)	Model Size	HANG7			SYSU		
		Acc	MF1	Kappa	Acc	MF1	Kappa
1	3.2M	0.8372	0.8078	0.7894	0.8026	0.7348	0.7281
2	5.4M	0.8452	0.8176	0.7997	0.8162	0.7446	0.7467
3	7.6M	0.8504	0.8243	0.8059	0.8206	0.7479	0.7479
4	9.9M	0.8568	0.8288	0.8138	0.8413	0.7788	0.7789
5	12.1M	0.8546	0.8270	0.8114	**0.8416**	**0.7807**	**0.7807**
6	14.3M	**0.8604**	**0.8323**	**0.8188**	0.8380	0.7793	0.7761

5.4 GradCAM Visualizations and Model Interpretability

To gain insights into the decision-making process of the LPSGM, we employed Gradient-weighted Class Activation Mapping (GradCAM) [31] to visualize the features that contributed most significantly to the model's decision-making process. GradCAM is a popular visualization technique used to generate heatmaps of the regions in the

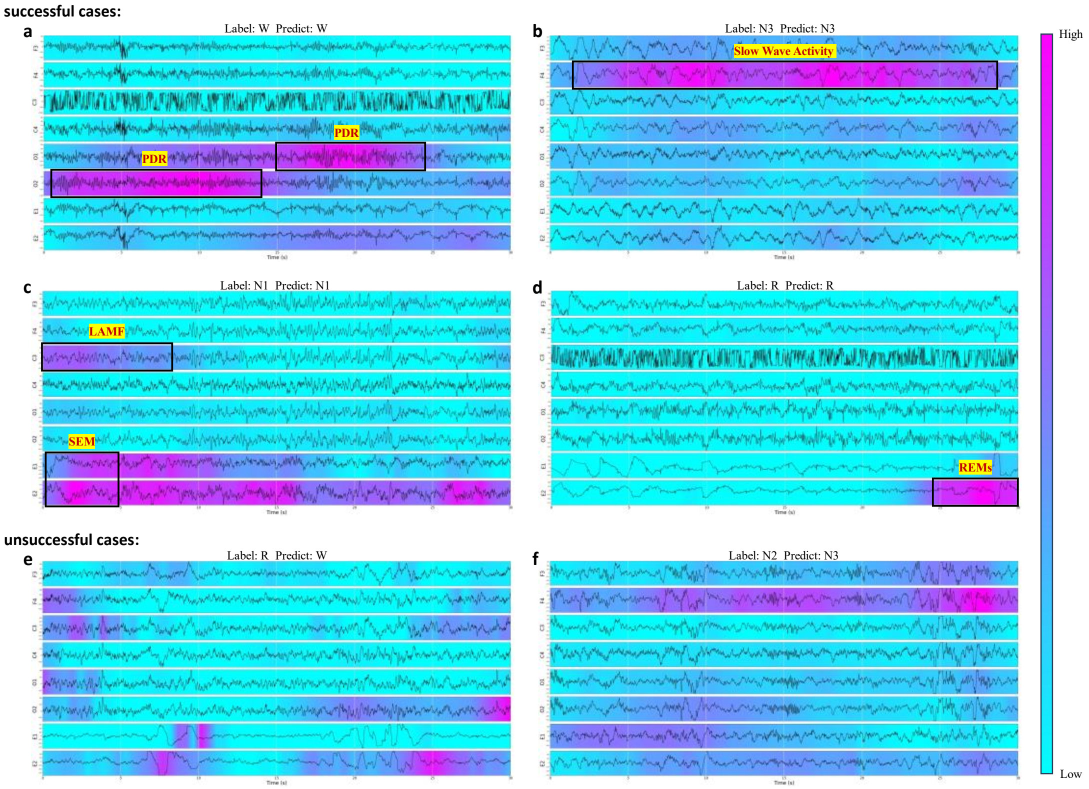


Fig. 7: Visualization of model interpretability with GradCAM for sleep staging. The figure illustrates GradCAM visualizations of the LPSGM model's predictions for six 30-second epochs of EEG data. The eight EEG channels shown from top to bottom are F3, F4, C3, C4, O1, O2, E1, and E2. Panels (a)-(d) show correctly predicted epochs, while panels (e)-(f) represent misclassified epochs. The highlighted regions represent areas of importance as identified by GradCAM, indicating features the model considered for each sleep stage. Black boxes indicate areas that physicians identified as aligning with the sleep staging criteria according to the AASM guidelines. (a) Posterior Dominant Rhythm (PDR) occupying more than 50% of the epoch, indicative of Wake. (b) Presence of slow wave activity in the frontal channels covering more than 20% of the epoch, indicative of N3. (c) Absence of PDR with Low-Amplitude Mixed-Frequency (LAMF) EEG and Slow Eye Movements (SEM), indicative of N1. (d) Presence of Rapid Eye Movements (REMs), consistent with R. Panels (e)-(f) demonstrate misclassifications, where the highlighted regions do not align with clinically relevant features.

input that were most influential in a given model's prediction. By applying GradCAM to the final convolutional layer output of the Epoch Encoder's dual-branch CNN, we generated activation maps that were subsequently resized to match the resolution of the original EEG input. The activation maps from each branch were then combined to produce a comprehensive visualization of the model's focus during sleep stage classification.

Fig. 7 presents examples of GradCAM visualizations for six 30-second epochs of

EEG data, with panels (a)-(d) showing successful predictions and panels (e)-(f) representing misclassifications. In panels (a)-(d), the highlighted regions correspond to clinically relevant features that align with the American Academy of Sleep Medicine (AASM) guidelines. For instance, in panel (a), the model correctly identifies the Posterior Dominant Rhythm (PDR) in the occipital channels, which is characteristic of Wake. In panel (b), slow wave activity in the frontal channels is highlighted, consistent with the N3 stage. Panel (c) demonstrates the model's ability to classify an N1 epoch, where the absence of PDR is coupled with Low-Amplitude Mixed-Frequency (LAMF) EEG and Slow Eye Movements (SEM), consistent with the N1 sleep stage. Panel (d) highlights Rapid Eye Movements (REMs), which is crucial for correctly identifying REM sleep. However, in panels (e) and (f), where misclassifications occur, the highlighted regions do not coincide with clinically relevant features. These discrepancies suggest that the model' s attention may have been misdirected or that the features needed for accurate classification were insufficient, highlighting areas for improvement in feature extraction and attention mechanisms.

Overall, the GradCAM visualizations provide important insights into the interpretability of the LPSGM. They demonstrate that the model is able to focus on clinically meaningful EEG features, but also highlight the need for refinement in handling ambiguous or complex EEG patterns.

5.5 Performance of LPSGM on Disease Diagnosis

In addition to sleep staging, we fine-tuned the LPSGM model on three downstream disease diagnosis tasks, focusing on both sleep disorder and psychiatric disorder. These tasks included diagnosing narcolepsy in the HANG7 dataset, identifying anxiety and depression tendencies in the HANG7 dataset, and classifying depressed versus healthy individuals in the SYSU dataset. The results of these tasks are presented in Tables 6, 7, and 8, respectively.

As shown in Tables 6-8, the fine-tuned LPSGM model consistently outperforms the model trained from scratch across all three disease diagnosis tasks. The joint fine-tune approach, which combines sleep staging and disease diagnosis tasks, achieved the best performance in terms of classification accuracy. These results indicate that integrating sleep staging features with disease diagnosis tasks improves the model' s ability to diagnose sleep-related and psychiatric disorders.

Table 6: Performance of Narcolepsy Dignosis on HANG7 Dataset.

Methods	Sample-wise					Subject-wise				
	Acc	F1	Sens	Spec	BAcc	Acc	F1	Sens	Spec	BAcc
Train from Scratch	0.6940	0.7646	0.7989	0.5274	0.6631	0.7500	0.8031	0.8418	0.6143	0.7281
Partial Fine-tune	0.7484	0.8054	0.8408	0.5940	0.7174	0.8088	0.8496	0.9000	0.6619	0.7810
Full Fine-tune	0.7685	0.8082	0.7877	0.7332	0.7605	0.8684	0.8929	0.9000	0.8143	0.8571
Joint Fine-tune	0.7919	0.8265	0.7955	0.7837	0.7896	0.8801	0.9017	0.9000	0.8429	0.8714

Table 7: Performance of Anxiety and Depression Tendencies Diagnosis on HANG7 Dataset.

Methods	Sample-wise					Subject-wise				
	Acc	F1	Sens	Spec	BAcc	Acc	F1	Sens	Spec	BAcc
Train from Scratch	0.6502	0.7471	0.8855	0.3300	0.6078	0.6433	0.7503	0.9333	0.2810	0.6071
Partial Fine-tune	0.6660	0.7267	0.7674	0.5223	0.6449	0.7358	0.8045	0.9306	0.4810	0.7058
Full Fine-tune	0.6696	0.6856	0.6396	0.7097	0.6747	0.7892	0.7768	0.6750	0.9381	0.8065
Joint Fine-tune	0.7002	0.7423	0.7513	0.6304	0.6908	0.8017	0.8332	0.8639	0.7286	0.7962

Table 8: Performance of Depressive Disorder Dignosis on SYSU Dataset.

Methods	Sample-wise					Subject-wise				
	Acc	F1	Sens	Spec	BAcc	Acc	F1	Sens	Spec	BAcc
Train from Scratch	0.8997	0.8487	0.9774	0.8721	0.9248	0.9730	1.0000	1.0000	0.9500	0.9750
Partial Fine-tune	0.9207	0.8785	0.9772	0.9007	0.9389	0.9730	1.0000	1.0000	0.9500	0.9750
Full Fine-tune	0.9227	0.8621	0.8889	0.9348	0.9118	0.9730	1.0000	1.0000	0.9500	0.9750
Joint Fine-tune	0.9683	0.9501	0.9998	0.9572	0.9712	1.0000	1.0000	1.0000	1.0000	1.0000

6. Conclusion

This paper introduces LPSGM, a unified and flexible model for sleep staging and disease diagnosis using PSG. By hybrid training on large-scale heterogeneous datasets, LPSGM significantly improves the generalizability and robustness of sleep staging models across clinical centers. Our experiments demonstrate that LPSGM achieves performance on par with models trained specifically on target datasets in cross-center testing. Moreover, LPSGM's flexible inference framework allows it to adjust input channel configurations, striking a balance between accuracy and inference speed. In addition to sleep staging, LPSGM can be fine-tuned for disease diagnosis tasks, including narcolepsy diagnosis, anxiety and depression detection, and depressive disorder diagnosis. Fine-tuning the pretrained model on disease-specific datasets further enhances its diagnostic accuracy, demonstrating the potential of LPSGM as a multi-task solution for both sleep disorders and psychiatric conditions.

While this work primarily focuses on sleep staging and disease diagnosis, future research will expand the model’s capabilities by integrating additional physiological signals to improve diagnostic accuracy. We also aim to extend LPSGM to other complex disease diagnosis tasks, incorporating it into a comprehensive framework for sleep disorder diagnosis and mental health screening. Additionally, efforts will be made to adapt LPSGM for use with portable EEG devices, enabling home-based monitoring.

Data Availability

The APPLES[32], [33], STAGES[32], ABC[32], [34], HOMEPAP[32], [35], SHHS[32], [36], PATS[32], [37], [38], CHAT[32], [39], CCSHS[32], [40], CFS[32], [41], MNC[32], [42], NCHSDB[32], [43] and MROS[32], [44] datasets are provided by the National Sleep Research Resource with appropriate deidentification. Permission and access for these datasets can be obtained via the online portal: https://www.sleepdata.org. The SVUH[45], HMC[46], P2018[47] and CAP[48] datasets are available from PhysioNet at https://physionet.org/content/ucddb/1.0.0/, https://physionet.org/content/hmc-sleep-staging/1.1/, https://physionet.org/content/challenge-2018/1.0.0/ and https://physionet.org/content/capslpdb/1.0.0/. The ISRUC[49] dataset can be accessed from https://sleeptight.isr.uc.pt/. DOD-H and DOD-O datasets[50] can be downloaded at https://github.com/Dreem-Organization/dreem-learning-open. Access to the HANG7 and SYSU dataset is governed by data-use agreements, and it is therefore not publicly available.

Funding

This work was supported by STI2030-Major Projects (2022ZD0212400), National Natural Science Foundation of China (82371453), Key R&D Program of Zhejiang （2024SSYS0017), Hangzhou Biomedical and Health Industry Special Projects for Science and Technology (2021WJCY240), Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2023-PT310-01) and Fundamental Research Funds for the Central Universities（2023ZFJH01-01, 2024ZFJH01-01）

References

[1] P. Maquet, “The Role of Sleep in Learning and Memory,” Science, vol. 294, no. 5544, pp. 1048–1052, Nov. 2001, doi: 10.1126/science.1062856.

[2] M. R. Irwin, “Why Sleep Is Important for Health: A Psychoneuroimmunology Perspective,” Annu. Rev. Psychol., vol. 66, no. 1, pp. 143–172, Jan. 2015, doi: 10.1146/annurev-psych-010213-115205.

[3] A. R. Hassan and M. I. H. Bhuiyan, “Automatic sleep stage classification,” in 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT), Khulna, Bangladesh: IEEE, Dec. 2015, pp. 211–216. doi: 10.1109/EICT.2015.7391948.

[4] Md. A. Rahman, Md. A. Hossain, Md. R. Kabir, M. H. Sani, Abdullah-Al-Mamun, and Md. A. Awal, “Optimization of Sleep Stage Classification using Single-Channel EEG Signals,” in 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh: IEEE, Dec. 2019, pp. 1–6. doi: 10.1109/EICT48899.2019.9068825.

[5] A. R. Hassan and M. I. H. Bhuiyan, “A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features,” Journal of Neuroscience Methods, vol. 271, pp. 107–118, Sep. 2016, doi: 10.1016/j.jneumeth.2016.07.012.

[6] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, “Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier,” Computer Methods and Programs in

Biomedicine, vol. 108, no. 1, pp. 10 – 19, Oct. 2012, doi: 10.1016/j.cmpb.2011.11.005.

[7] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and Classification Algorithms," J Med Syst, vol. 38, no. 3, p. 18, Mar. 2014, doi: 10.1007/s10916-014-0018-0.

[8] S. Zhao, F. Long, X. Wei, X. Ni, H. Wang, and B. Wei, "Evaluation of a Single-Channel EEG-Based Sleep Staging Algorithm," International Journal of Environmental Research and Public Health, vol. 19, no. 5, Art. no. 5, Jan. 2022, doi: 10.3390/ijerph19052845.

[9] L. Bahatti, " Sleep ' s depth detection using electroencephalogram signal processing and neural network classification," Journal of Medical Artificial Intelligence, Accessed: Dec. 28, 2023. [Online]. Available: https://www.academia.edu/109044011/Sleep_s_depth_detection_using_electroencephalogram_signal_processing_and_neural_network_classification

[10] 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, doi: 10.1109/TNSRE.2017.2721116.

[11] A. Supratak and Y. Guo, "TinySleepNet: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada: IEEE, Jul. 2020, pp. 641 – 644. doi: 10.1109/EMBC44109.2020.9176741.

[12] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. Bianchi, "Expert-level sleep scoring with deep neural networks," Journal of the American Medical Informatics Association : JAMIA, vol. 25, Nov. 2018, doi: 10.1093/jamia/ocy131.

[13] H. Phan, O. Y. Chen, 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., pp. 1 – 1, 2021, doi: 10.1109/TPAMI.2021.3070057.

[14] C. Yoo, H. W. Lee, and J.-W. Kang, "Transferring Structured Knowledge in Unsupervised Domain Adaptation of a Sleep Staging Network," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 3, pp. 1273 – 1284, Mar. 2022, doi: 10.1109/JBHI.2021.3103614.

[15] J. Fan et al., "Unsupervised Domain Adaptation by Statistics Alignment for Deep Sleep Staging Networks," IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, vol. 30, 2022.

[16] A. Guillot and V. Thorey, "RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 1441 – 1451, 2021, doi: 10.1109/TNSRE.2021.3098968.

[17] Z. Deng, L. Zhang, K. Vodrahalli, K. Kawaguchi, and J. Y. Zou, "Adversarial Training Helps Transfer Learning via Better Representations," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2021, pp. 25179 – 25191. Accessed: Feb. 25, 2024. [Online]. Available:

https://proceedings.neurips.cc/paper/2021/hash/d3aeec875c479e55d1cdeea161842ec6-Abstract.html

[18] J. Wang, S. Zhao, H. Jiang, S. Li, T. Li, and G. Pan, "Generalizable Sleep Staging via Multi-Level Domain Alignment," Jan. 12, 2024, arXiv: arXiv:2401.05363. Accessed: Jan. 22, 2024. [Online]. Available: http://arxiv.org/abs/2401.05363

[19] OpenAI et al., "GPT-4 Technical Report," Mar. 04, 2024, arXiv: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[20] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 29, 2019, arXiv: arXiv:1910.13461. Accessed: Aug. 30, 2024. [Online]. Available: http://arxiv.org/abs/1910.13461

[21] "The Llama3 Herd of Models".

[22] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive Representation Learning for Electroencephalogram Classification," in Proceedings of the Machine Learning for Health NeurIPS Workshop, PMLR, Nov. 2020, pp. 238–253. Accessed: Apr. 19, 2024. [Online]. Available: https://proceedings.mlr.press/v136/mohsenvand20a.html

[23] J. Han, X. Wei, and A. A. Faisal, "EEG Decoding for Datasets with Heterogenous Electrode Configurations using Transfer Learning Graph Neural Networks," Jun. 20, 2023, arXiv: arXiv:2306.13109. Accessed: Apr. 19, 2024. [Online]. Available: http://arxiv.org/abs/2306.13109

[24] X. Gu, J. Han, G.-Z. Yang, and B. Lo, "Generalizable Movement Intention Recognition with Multiple Heterogeneous EEG Datasets," in 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom: IEEE, May 2023, pp. 9858–9864. doi: 10.1109/ICRA48891.2023.10160462.

[25] R. Li, Y. Wang, W.-L. Zheng, and B.-L. Lu, "A Multi-view Spectral-Spatial-Temporal Masked Autoencoder for Decoding Emotions with Self-supervised Learning," in Proceedings of the 30th ACM International Conference on Multimedia, Lisboa Portugal: ACM, Oct. 2022, pp. 6–14. doi: 10.1145/3503161.3548243.

[26] K. Yi, K. Ren, Y. Wang, and D. Li, "Learning Topology-Agnostic EEG Representations with Geometry-Aware Modeling".

[27] "Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI," presented at the The Twelfth International Conference on Learning Representations, Oct. 2023. Accessed: Jan. 17, 2024. [Online]. Available: https://openreview.net/forum?id=QzTpTRVtrP

[28] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 03, 2021, arXiv: arXiv:2010.11929. Accessed: Nov. 28, 2023. [Online]. Available: http://arxiv.org/abs/2010.11929

[29] M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging," Oct. 24, 2019, arXiv: arXiv:1910.11162. Accessed: Dec. 30, 2023. [Online]. Available: http://arxiv.org/abs/1910.11162

[30] 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, doi: 10.1109/TNSRE.2021.3076234.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Int J Comput Vis, vol. 128, no. 2, pp. 336 – 359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[32] Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. 2018 Oct 1;25(10):1351-1358. doi: 10.1093/jamia/ocy064. PMID: 29860441; PMCID: PMC6188513.

[33] Quan SF, Chan CS, Dement WC, Gevins A, Goodwin JL, Gottlieb DJ, Green S, Guilleminault C, Hirshkowitz M, Hyde PR, Kay GG, Leary EB, Nichols DA, Schweitzer PK, Simon RD, Walsh JK, Kushida CA. The association between obstructive sleep apnea and neurocognitive performance--the Apnea Positive Pressure Long-term Efficacy Study (APPLES). Sleep. 2011 Mar 1;34(3):303-314B. doi: 10.1093/sleep/34.3.303. PMID: 21358847; PMCID: PMC3041706.

[34] Bakker JP, Tavakkoli A, Rueschman M, Wang W, Andrews R, Malhotra A, Owens RL, Anand A, Dudley KA, Patel SR. Gastric Banding Surgery versus Continuous Positive Airway Pressure for Obstructive Sleep Apnea: A Randomized Controlled Trial. Am J Respir Crit Care Med. 2018 Apr 15;197(8):1080-1083. doi: 10.1164/rccm.201708-1637LE. PMID: 29035093; PMCID: PMC5909166.

[35] Rosen CL, Auckley D, Benca R, Foldvary-Schaefer N, Iber C, Kapur V, Rueschman M, Zee P, Redline S. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. Sleep. 2012 Jun 1;35(6):757-67. doi: 10.5665/sleep.1870. PMID: 22654195; PMCID: PMC3353048.

[36] Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec;20(12):1077-85. PMID: 9493915.

[37] Wang R, Bakker JP, Chervin RD, Garetz SL, Hassan F, Ishman SL, Mitchell RB, Morrical MG, Naqvi SK, Radcliffe J, Riggan EI, Rosen CL, Ross K, Rueschman M, Tapia IE, Taylor HG, Zopf DA, Redline S. Pediatric Adenotonsillectomy Trial for Snoring (PATS): protocol for a randomised controlled trial to evaluate the effect of adenotonsillectomy in treating mild obstructive sleep-disordered breathing. BMJ Open. 2020 Mar 15;10(3):e033889. doi: 10.1136/bmjopen-2019-033889. PMID: 32179560; PMCID: PMC7073822.

[38] Redline S, Cook K, Chervin RD, Ishman S, Baldassari CM, Mitchell RB, Tapia IE, Amin R, Hassan F, Ibrahim S, Ross K, Elden LM, Kirkham EM, Zopf D, Shah J, Otteson T, Naqvi K, Owens J, Young L, Furth S, Connolly H, Clark CAC, Bakker JP, Garetz S, Radcliffe J, Taylor HG, Rosen CL, Wang R; Pediatric Adenotonsillectomy Trial for Snoring (PATS) Study Team. Adenotonsillectomy for Snoring and Mild Sleep Apnea in Children: A Randomized Clinical Trial. JAMA.

2023 Dec 5;330(21):2084-2095. doi: 10.1001/jama.2023.22114. PMID: 38051326; PMCID: PMC10698619.

[39] Marcus CL, Moore RH, Rosen CL, Giordani B, Garetz SL, Taylor HG, Mitchell RB, Amin R, Katz ES, Arens R, Paruthi S, Muzumdar H, Gozal D, Thomas NH, Ware J, Beebe D, Snyder K, Elden L, Sprecher RC, Willging P, Jones D, Bent JP, Hoban T, Chervin RD, Ellenberg SS, Redline S; Childhood Adenotonsillectomy Trial (CHAT). A randomized trial of adenotonsillectomy for childhood sleep apnea. N Engl J Med. 2013 Jun 20;368(25):2366-76. doi: 10.1056/NEJMoa1215881. Epub 2013 May 21. PMID: 23692173; PMCID: PMC3756808.

[40] Rosen CL, Larkin EK, Kirchner HL, Emancipator JL, Bivins SF, Surovec SA, Martin RJ, Redline S. Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: association with race and prematurity. J Pediatr. 2003 Apr;142(4):383-9. doi: 10.1067/mpd.2003.28. PMID: 12712055.

[41] Redline S, Tishler PV, Tosteson TD, Williamson J, Kump K, Browner I, Ferrette V, Krejci P. The familial aggregation of obstructive sleep apnea. Am J Respir Crit Care Med. 1995 Mar;151(3 Pt 1):682-7. doi: 10.1164/ajrccm/151.3_Pt_1.682. PMID: 7881656.

[42] Stephansen JB, Olesen AN, Olsen M, Ambati A, Leary EB, Moore HE, Carrillo O, Lin L, Han F, Yan H, Sun YL, Dauvilliers Y, Scholz S, Barateau L, Hogl B, Stefani A, Hong SC, Kim TW, Pizza F, Plazzi G, Vandi S, Antelmi E, Perrin D, Kuna ST, Schweitzer PK, Kushida C, Peppard PE, Sorensen HBD, Jennum P, Mignot E. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat Commun. 2018 Dec 6;9(1):5229. doi: 10.1038/s41467-018-07229-3. PMID: 30523329; PMCID: PMC6283836.

[43] Lee H, Li B, DeForte S, Splaingard ML, Huang Y, Chi Y, Linwood SL. A large collection of real-world pediatric sleep studies. Sci Data. 2022 Jul 19;9(1):421. doi: 10.1038/s41597-022-01545-6. PMID: 35853958; PMCID: PMC9296671.

[44] Blackwell T, Yaffe K, Ancoli-Israel S, Redline S, Ensrud KE, Stefanick ML, Laffan A, Stone KL; Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the Osteoporotic Fractures in Men Sleep Study. J Am Geriatr Soc. 2011 Dec;59(12):2217-25. doi: 10.1111/j.1532-5415.2011.03731.x. Epub 2011 Nov 7. PMID: 22188071; PMCID: PMC3245643.

[45] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215 – e220.

[46] Alvarez-Estevez, Diego, and Roselyne Rijsman. "Haaglanden Medisch Centrum sleep staging database" (version 1.1). PhysioNet (2022), https://doi.org/10.13026/t79q-fr32.

[47] Ghassemi MM, Moody BE, Lehman LW, Song C, Li Q, Sun H, Mark RG, Westover MB, Clifford GD. You snooze, you win: the physionet/computing in cardiology challenge 2018. In 2018 Computing in Cardiology Conference (CinC) 2018 Sep 23 (Vol. 45, pp. 1-4). IEEE. doi: 10.22489/CinC.2018.049.

[48] MG Terzano, L Parrino, A Sherieri, R Chervin, S Chokroverty, C Guilleminault, M Hirshkowitz, M Mahowald, H Moldofsky, A Rosa, R Thomas, A Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 2001 Nov; 2(6):537-553.

[49] Khalighi, Sirvan, et al. "ISRUC-Sleep: A comprehensive public dataset for sleep researchers." Computer methods and programs in biomedicine 124 (2016): 180-192.

[50] Guillot, Antoine, et al. "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging." IEEE transactions on neural systems and rehabilitation engineering 28.9 (2020): 1955-1965.

LPSGM: A Unified Flexible Large PSG Model for Sleep Staging and Mental Disorder Diagnosis

Supplementary Material

Datasets

To the best of our knowledge, this work represents the most extensive use of data for training a large-scale sleep staging model to date. We aggregated approximately 24,000 full-night PSG recordings, encompassing roughly 220,000 hours of sleep data, from 16 publicly available datasets as source domains for training. A summary of the public datasets is provided below.

(1) The Apnea Positive Pressure Long-term Efficacy Study (APPLES) is a multi-center dataset that includes overnight PSG recordings from 1,104 patients with obstructive sleep apnea syndrome (OSAS). The dataset includes four EEG and two EOG channels, and was scored using Rechtschaffen and Kales (R&K) criteria.

(2) The Danish Center for Sleep Medicine (DCSM) dataset consists of 255 randomly selected and fully anonymized overnight lab-based PSG recordings from patients visiting the DCSM for the diagnosis of non-specific sleep related disorders. The dataset includes six EEG and two EOG channels, and was scored according to the AASM criteria.

(3) The Dreem Open Dataset (DOD) consists of two subsets, DOD-H and DOD-O. DOD-H comes from French Armed Forces Biomedical Research Institute' s (IRBA), and contains PSG recordings from 25 healthy volunteers. DOD-O comes from the Stanford Sleep Medicine Center, and contains PSG recordings from 56 OSAS patients. Both datasets contain twelve and eight EEG channels, respectively. For experimentation, we specifically selected the three and five channels that overlap with the AASM criteria. Both contain two EOG channels and were scored according to the AASM criteria.

(4) Haaglanden Medisch Centrum (HMC) dataset consists of 151 randomly selected whole-night PSG of different sleep disorders. The dataset includes four EEG and two EOG channels, and was scored according to AASM criteria.

(5) The Institute of Systems and Robotics, University of Coimbra (ISRUC) dataset consists of 126 PSG recordings from the Sleep Medicine Center of the Hospital of the University of Coimbra, Portugal. The dataset comprises three groups of data. Data in group one concerning 100 subjects, with one recording session per subject. Data in group two is gathered from 8 subjects and two recording sessions were performed per subject. Data in group three is collected from one recording

session related to 10 healthy subjects. The dataset includes six EEG and two EOG channels and was scored according to the AASM criteria.

(6) The St. Vincent's University Hospital (SVUH) dataset contains 25 full overnight PSG with suspected sleep-disordered breathing. The dataset contains two EEG and two EOG channels and was scored according to R&K criteria.

(7) P2018 (You Snooze You Win: The PhysioNet/Computing in Cardiology Challenge 2018) dataset was contributed by the Massachusetts General Hospital’s (MGH) Computational Clinical Neurophysiology Laboratory (CCNL), and the Clinical Data Animation Laboratory (CDAC). The dataset consists of 994 training examples and 989 test examples, with only the training data having labels publicly available. We only use the training data, which includes 6 EEG and 1 EOG channels, and uses the AASM criteria for sleep staging.

(8) The Stanford Technology Analytics and Genomics in Sleep (STAGES) dataset was collected on 1500 patients evaluated for sleep disorders from six centers. The dataset contains six EEG and two EOG channels and is annotated based on AASM criteria.

(9) The Apnea, Bariatric surgery, and CPAP (ABC) dataset includes 80 patients with severe OSAS, with six EEG and two EOG channels, being scored based on AASM criteria.

(10) The Nationwide Children's Hospital Sleep DataBank (NCHSDB) has 3,984 pediatric sleep studies on 3,673 unique patients conducted at NCH in Columbus, Ohio, USA between 2017 and 2019. The dataset includes six EEG and two EOG channels and using AASM criteria.

(11) The Home Positive Airway Pressure (HOMEPAP) study was a multi-center dataset that enrolled 373 patients with suspected moderate and severe OSAS. Subjects were randomized to lab-based and home-based management. We only use the lab-based subset as it includes the channels we need. The lab-based subset includes six EEG and two EOG channels with AASM criteria annotation.

(12) The Childhood Adenotonsillectomy Trial (CHAT) is a multi-center dataset that enrolled 1447 children with mild to moderate OSAS. The dataset includes six EEG and two EOG channels with AASM criteria annotation.

(13) The Cleveland Children's Sleep and Health Study (CCSHS) dataset consists of 515 PSG recordings from 907 children aged between 8 and 11 years old. The dataset includes two EEG and two EOG channels, and was scored according to AASM criteria.

(14) The Cleveland Family Study (CFS) is a family-based study of sleep apnea worldwide, compring 730 overnight PSG from 2284 indivisuals. The dataset includes two EEG and two EOG channels with AASM criteria.

(15) The MROS dataset enrolled 5994 men 65 years or older at six clinical centers. The dataset consists of 3929 PSG recordings with two EEG and two EOG channals, and was scored according to AASM criteria.

(16) The Sleep Heart Health Study (SHHS) is a multi-center cohort study implemented by the National Heart Lung & Blood Institute, consisting of two subset. SHHS-1 and SHHS-2 contain 5793 and 2651 overnight PSG, respectively.

The dataset contains two EEG and EOG channels and was scored according to R&K criteria.

We test the model on two private datasets from different clinical centers to evaluate the cross-center generalization performance of our methods. The discription of each dataset is provided below.

(1) The HANG7 dataset was acquired from Affiliated Mental Health Center & Hangzhou Seventh People’s Hospital, Zhejiang University School of Medicine. It was performed at the Zhejiang University with Institutional Review Board approval and written consent was obtained from all the subjects or their caregivers. It comprises PSG recordings from 127 subjects, including 33 healthy individuals, 51 patients diagnosed with narcolepsy, and 43 patients diagnosed with depression. The dataset includes six EEG and two EOG channels sampled at a frequency of 512 Hz. The PSG were scored by experienced clinicians according to AASM criteria.

(2) The SYSU dataset includes two groups: healthy individuals and patients with depression. All subjects were equipped with the same device to collect overnight PSG signals. This study received approval from the Ethics Committee of Guangdong 999 Brain Hospital (approval number: 2020-010-059). The experiments involving healthy individuals were conducted at the sleep laboratory of Sun Yat-sen University, encompassing 80 PSG recordings from 20 subjects sampled at 500 Hz. The depression dataset comprised 24 PSG recordings from 24 depressed patients, who were diagnosed by two experienced psychiatrists based on the criteria of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The scores of 24-item Hamilton Depression Scale (HAMD-24) and the self-rating Depression Scale (SDS) were 22.6±6.20 and 65.57±9.53, respectively. All patients were without drug abuse, suicide risk, pregnancy, present or history of head injuries, seizures, or epilepsy. The EEG signals were sampled at 256 Hz. Both groups included six EEG and two EOG channels and were scored according to the AASM scoring manual by two well-trained sleep technologists.

Preprocessing

For all PSG datasets, we selected eight EEG/EOG channels recommended by the AASM criteria for sleep staging. No other channels or data types present in the datasets were utilized. The chosen eight channels included F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1, E1-M2, and E2-M1.

All the PSG recordings were initially subjected to a fourth-order bandpass filter (0.3 Hz to 35 Hz) and subsequently resampled at a rate of 100 Hz. Finally, Z-score normalization was applied individually to each channel of every PSG recording:

$$x[c] = \frac{x[c] - \text{mean}(x[c])}{\text{std}(x[c])}, c \in C \tag{11}$$

where x represents a single PSG recording, C denotes the set of channels for that recording. The samples were clamped to the range $[-10,10]$ after Z-score normalization to minimize the impact of outliers.

We adhered to the current AASM sleep staging standards. For datasets originally labeled according to the R&K standards, we followed the conventional approach of merging stages N3 and N4 into a single N3 stage. Additionally, we removed any sleep epochs without labels, which typically indicated sensor detachment, sleep interruptions, or other anomalies. After removing such segments, the data was divided into two distinct segments at that specific point. Table 2 shows the class distribution across the datasets.

Acknowledgements

The Apnea Positive Pressure Long-term Efficacy Study (APPLES) was supported by the National Heart, Lung, and Blood Institute (U01HL68060).

This research has been conducted using the STAGES - Stanford Technology, Analytics and Genomics in Sleep Resource funded by the Klarman Family Foundation. The investigators of the STAGES study contributed to the design and implementation of the STAGES cohort and/or provided data and/or collected biospecimens, but did not necessarily participate in the analysis or writing of this report. The full list of STAGES investigators can be found at the project website.

The Apnea, Bariatric surgery, and CPAP study (ABC Study) was supported by National Institutes of Health grants R01HL106410 and K24HL127307. Philips Respironics donated the CPAP machines and supplies used in the perioperative period for patients undergoing bariatric surgery.

The Home Positive Airway Pressure study (HomePAP) was supported by the American Sleep Medicine Foundation 38-PM-07 Grant: Portable Monitoring for the Diagnosis and Management of OSA.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

The Pediatric Adenotonsillectomy Trial for Snoring (PATS) study was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (1U01HL125307, 1U01HL125295).

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989).

The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39).

The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380).

The Mignot Nature Communications research was mostly supported by a grant from Jazz Pharmaceuticals to E.M. Additional funding came from: NIH grant R01HL62252 (to P.E.P.); Ministry of Science and Technology 2015CB856405 and National

Foundation of Science of China 81420108002,81670087 (to F.H.); H. Lundbeck A/S, Lundbeck Foundation, Technical University of Denmark and Center for Healthy Aging, University of Copenhagen (to P.J. and H.B.D.S). Additional support was provided by the Klarman Family, Otto Mønsted, Stibo, Vera & Carl Johan Michaelsens, Knud Højgaards, Reinholdt W. Jorck and Hustrus and Augustinus Foundations (to A.N.O.).

NCH Sleep DataBank was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB025018.

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839.

The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

Table S1. Overview of datasets used in our experiments.

Datasets		Recordings	Annotation	EEG						EOG		Channels	Sample Rate	
				F3-M2	F4-M1	C3-M2	C4-M1	O1-M2	O2-M1	E1-M2	E2-M1		EEG	EOG
APPLES		1067	R&K			√	√	√	√	√	√	6	100	100
DCSM		255	AASM	√	√	√	√	√	√	√	√	8	256	256
DOD	DOD-H	25	AASM	√	√	√				√	√	5	250	250
	DOD-O	56	AASM	√		√	√	√	√	√	√	7	250	250
HMC		151	AASM		√	√	√		√	√	√	6	256	256
ISRUC		126	AASM	√	√	√	√	√	√	√	√	8	200	200
SVUH		25	R&K			√	√			√	√	4	128	64
P2018		994	AASM	√	√	√	√	√	√	√		7	200	200
STAGES	BOGN	85	AASM	√	√	√	√	√	√	√	√	8	200	200
	STNF	525		√	√	√	√	√	√	√	√	8	256	256
	GSDV	288		√	√	√	√	√	√	√	√	8	200	200
	MSTR	286		√	√	√	√	√	√	√	√	8	256	256
	GSBB	38		√	√	√	√	√	√	√	√	8	200	200
	GSLH	51		√	√	√	√	√	√	√	√	8	200	200
	GSSA	34		√	√	√	√	√	√	√	√	8	200	200
	GSSW	131		√	√	√	√	√	√	√	√	8	200	200
	MSMI	61		√	√	√	√	√	√	√	√	8	200	200
	MSNF	35		√	√	√	√	√	√	√	√	8	200	200
	MSQW	145		√	√	√	√	√	√	√	√	8	200	200
	MSTH	31		√	√	√	√	√	√	√	√	8	256	256
	STLK	156		√	√	√	√	√	√	√	√	8	500	500

ABC		132	AASM	√	√	√	√	√	√	√	√	8	256	256
NCHSDB		3947	AASM	√	√	√	√	√	√	√	√	8	256/400/512	256/400/512
HOMEPAP		245	AASM	√	√	√	√	√	√	√	√	8	200/256	200/256
CHAT		1638	AASM	√	√	√	√	√	√	√	√	8	200	200
CCSHS		515	AASM			√	√			√	√	4	128	128
CFS		730	AASM			√	√			√	√	4	128	128
MROS		3929	AASM			√	√			√	√	4	256	256
SHHS	SHHS-1	5793	R&K			√	√			√	√	4	125	50
	SHHS-2	2651				√	√			√	√	4	128	32
HANG7		127	AASM	√	√	√	√	√	√	√	√	8	512	512
SYSU		104	AASM	√	√	√	√	√	√	√	√	8	500/256	500/256

Table S2. Number of sleep stages of the datasets after preprocessing.

Datasets		Total	Epochs					Ratio %				
			W	N1	N2	N3	R	W	N1	N2	N3	R
APPLES		1049110	256128	147217	481003	24295	140467	24.4	14	45.8	2.3	13.4
DCSM		304266	79636	21140	113027	43637	46826	26.2	6.9	37.1	14.3	15.4
DOD	DOD-H	24662	3037	1505	11879	3514	4727	12.3	6.1	48.2	14.2	19.2
	DOD-O	54197	10660	2898	26650	5683	8306	19.7	5.3	49.2	10.5	15.3
HMC		137243	23686	15548	50083	26671	21255	17.3	11.3	36.5	19.4	15.5
ISRUC		107784	23198	14254	34661	21489	14182	21.5	13.2	32.2	19.9	13.2
SVUH		20774	4707	3016	3403	7658	1990	22.7	14.5	16.4	36.9	9.5

P2018		892262	157945	136978	377870	102592	116877	17.7	15.4	42.3	11.5	13.1
STAGES	BOGN	75930	23185	4623	29819	8236	10067	30.5	6.1	39.3	10.8	13.3
	STNF	592027	202352	57652	171620	52147	108256	34.2	9.7	29	8.8	18.3
	GSDV	218608	54699	13363	114794	10194	25558	25	6.1	52.5	4.7	11.7
	MSTR	221466	42522	23789	103118	24201	27836	19.2	10.7	46.6	10.9	12.6
	GSBB	29843	7954	2469	14423	1668	3329	26.7	8.3	48.3	5.6	11.2
	GSLH	32118	8540	2684	16485	1597	2812	26.6	8.4	51.3	5	8.8
	GSSA	27751	6998	920	15233	881	3719	25.2	3.3	54.9	3.2	13.4
	GSSW	90341	25790	5944	44213	3549	10845	28.5	6.6	48.9	3.9	12
	MSMI	45892	8035	4227	22959	4847	5824	17.5	9.2	50	10.6	12.7
	MSNF	27061	4905	1385	13454	4012	3305	18.1	5.1	49.7	14.8	12.2
	MSQW	113085	25039	13622	53319	8125	12980	22.1	12	47.1	7.2	11.5
	MSTH	23682	5036	2146	12033	1755	2712	21.3	9.1	50.8	7.4	11.5
	STLK	154691	29429	11937	76878	11426	25021	19	7.7	49.7	7.4	16.2
ABC		133000	30938	19296	52334	11761	18671	23.3	14.5	39.3	8.8	14
NCHSDB		3661376	665063	128183	1382551	874762	610817	18.2	3.5	37.8	23.9	16.7
HOMEPAP		229604	63395	24759	86718	22645	32087	27.6	10.8	37.8	9.9	14
CHAT		1957293	469804	119436	628932	464993	274128	24	6.1	32.1	23.8	14
CCSHS		691401	212027	19221	249698	110191	100264	30.7	2.8	36.1	15.9	14.5
CFS		866204	321333	26394	306264	111937	100276	37.1	3	35.4	12.9	11.6
MROS		5373725	2609224	218233	1735010	278620	532638	48.6	4.1	32.3	5.2	9.9
SHHS	SHHS-1	5863207	1691288	217583	2397460	739403	817473	28.8	3.7	40.9	12.6	13.9
	SHHS-2	3192507	1208326	111456	1147780	313790	411155	13.9	3.5	36	9.8	12.9
HANG7		142450	36337	10121	46875	27761	21356	25.5	7.1	32.9	19.5	15
SYSU		106778	9554	5071	45123	24793	22237	8.9	4.7	42.3	23.2	20.8