A Unified Flexible Large Polysomnography Model for Sleep Staging and Mental Disorder Diagnosis

Guifeng Deng[1,2], Mengfan Niu[1], Yuxi Luo[3], Shuying Rao[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

Sleep quality is vital to human health, yet automated sleep staging faces challenges in cross-center generalization due to data scarcity and domain gaps. Traditional scoring is labor-intensive, while deep learning models often fail to generalize across datasets. Here, we present LPSGM, a unified and flexible large polysomnography (PSG) model designed to enhance cross-center generalization in sleep staging and enable fine-tuning for disease diagnosis. Trained on 220,500 hours of PSG data from 16 public datasets, LPSGM integrates domain-adaptive learning and supports variable-channel configurations, achieving performance comparable to models trained directly on target-center data. In a prospective clinical study, LPSGM matches expert-level accuracy with lower variability. When fine-tuned, it attains 88.01% accuracy in narcolepsy detection and 100% in depression detection. These results establish LPSGM as a scalable, plug-and-play solution for automated PSG analysis, bridging the gap between sleep staging and clinical deployment.

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. One study [51] reports inter-scorer reliability for sleep stage scoring to be 82.6% on average, a result consistently found by others[52-55].

Over the past few decades, numerous computer-based automatic sleep staging methods have emerged, leading to significant advancements in the field. Early automatic sleep staging methods relied on hand-crafted features to train traditional machine learning models[3-9], which lacked the capacity to capture complex patterns. Deep learning-based methods have achieved performance comparable to human

experts on large-scale datasets, revolutionizing automatic sleep staging. By leveraging CNNs and RNNs, these methods can effectively learn hierarchical features from raw physiological signals, enabling the extraction of sophisticated patterns from EEG data[10-13]. However, most of these methods focus on intra-dataset learning, limiting their effectiveness when applied to external datasets due to 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. Others[18] utilize consistency constraints to enforce the extraction of shared features across domains. Although these approaches have shown promising results in improving out of distribution (OOD) performance, none of these transfer learning methods have achieved performance comparable to models trained directly on data from the target center.

Nowadays, the clinical adoption of these automatic sleep staging systems remains limited, primarily due to the challenges of domain gap and data scarcity. On the one hand, the lack of standardized PSG protocols introduces significant variability across datasets from different centers, stemming from differences in population characteristics, signal acquisition conditions (such as equipment, electrode placement, channel configurations, sampling rates, and signal-to-noise ratios), and the subjectivity of annotators. As deep learning methods inherently assume that data are independently and identically distributed, models trained on data from a single center often exhibit considerable performance degradation when applied to data from other centers. On the other hand, deep learning-based approaches require large-scale labeled datasets to achieve high accuracy and robust generalization. However, obtaining such extensive labeled data is prohibitively expensive and often beyond the capacity of many clinical sleep centers, further restricting the widespread implementation of these systems.

A large amount of public datasets are available because of the foundational and long-established nature of the sleep staging. However, the variability in montages hinders the full utilization of these public datasets. Current studies focus on training and testing within individual dataset, typically using data from specific EEG channels from one or a few public datasets. Inspired by the success of large language model (LLM)[19-21] and EEG foundation model [24-27], we propose to address the cross-center sleep staging by effectively integrating these diverse datasets, thereby enhancing the generalization and robustness of models.

Furthermore, studies have shown that sleep characteristics are closely associated with various mental disorder like narcolepsy and depression. Narcolepsy is marked by early REM onset and rapid transitions between wakefulness and REM sleep [42, 56, 57]. Depressive disorder exhibits disrupted sleep continuity, characterized by prolonged sleep latency and frequent awakenings, along with reduced slow wave sleep (SWS, measured as decline delta power), disinhibition REM sleep and diminished sleep spindle activity[58-61]. These findings support the feasibility of PSG-based diagnosis for mental disorders like narcolepsy and depression.

In this work, we present LPSGM, a unified and flexible large PSG model designed to address the challenges of cross-center sleep staging and to enable fine-tuning for

PSG-based disease diagnosis. LPSGM introduces a unified training framework for heterogeneous datasets and allows flexible channel input adjustments during inference. By hybrid training on 220,500 hours PSG from 16 public datasets, LPSGM achieves comparable accuracy to fully supervised training on two target centers. Prospective study demonstrates that LPSGM achieves performance on par with human experts in sleep staging and can be seamlessly applied in clinical settings as a plug-and-play solution. Furthermore, fine-tuning on disease-specific datasets demonstrating LPSGM's versatility in clinical diagnostics beyond sleep staging.

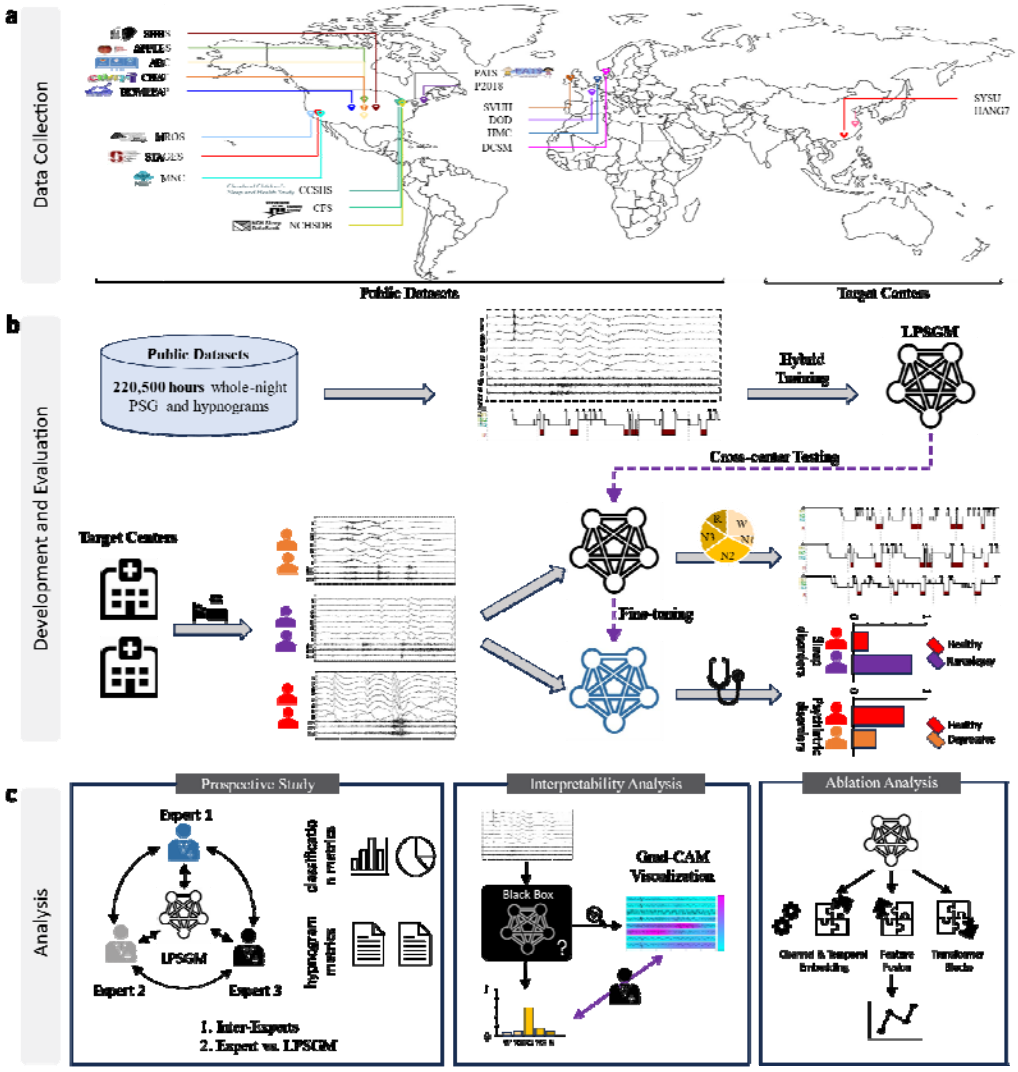

Fig. 1: Overview of the LPSGM framework for sleep staging and disorder diagnosis. Panel (a) illustrates the data collection process, which integrates 220,500 hours of polysomnography (PSG) data from 16 publicly available datasets and two target-center datasets, covering a diverse range of geographic regions and recording conditions. Panel (b) presents the development and evaluation pipeline, where LPSGM is trained on large-scale public datasets and evaluated for cross-center sleep staging on target-center data. The model is then fine-tuned for downstream applications, including narcolepsy diagnosis and depression detection. Panel (c)

summarizes the research analyses conducted in this study, including a prospective clinical study to validate real-world performance, an interpretability analysis to examine model decision-making, and an ablation study to assess the contribution of key model components.

Results

Cross-Center Classification Performance of Sleep Staging

The cross-center classification results on HANG7 and SYSU datasets compared with baselines are presented in Table 1. To establish performance boundaries, we conducted two baseline experiments: (1) Baseline 1, where a model trained on one center's dataset was directly applied to another center's dataset without adaptation, representing the lower bound of performance; and (2) Baseline 2, where a model was fully trained from scratch on the target center's dataset using five-fold cross-validation, establishing the upper bound of achievable performance. Both baselines were evaluated using LPSGM and LPSGM-Small models, with the latter consistently outperforming the former. Therefore, LPSGM-Small was chosen as the official baseline for further comparisons.

To quantify the generalization performance, we computed LPSGM's performance as a percentage relative to Baseline 2, indicating its adaptability across different clinical centers. On the HANG7 dataset, LPSGM achieved 85.68% accuracy, 82.88% macro-F1, and 0.8138 kappa, reaching 99.6%, 99.9%, and 99.5% of Baseline 2's performance. On the SYSU dataset, LPSGM obtained 84.13% accuracy, 77.88% macro-F1, and 0.7789 kappa, corresponding 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*	0.7940	0.7438	0.7323	0.8757	0.4530	0.7862	0.8276	0.7767
	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%

a. Base. 1 is trained on one center and directly tested on another without adaptation, representing the lower bound of performance. Base. 2 is fully trained on the target center using five-fold cross-validation, serving as the upper bound.
b. LPSGM-Small is a smaller variant of LPSGM, with one Transformer block (Fig. 5c) instead of four.
c. Rel. Base. 2 measures LPSGM's performance relative to Base. 2, indicating its generalization capability.

We analyzed the performance of LPSGM across different subject groups in the HANG7 and SYSU datasets, as illustrated in Fig. 2b. The model demonstrated the highest classification performance in normal subjects, achieving an accuracy of 88.99% on HANG7 and 86.15% on SYSU. However, a slight performance decline was observed in subjects with sleep-related or neurological disorders. Specifically, in the HANG7 dataset, the accuracy dropped to 87.11% for the hypersomnia group and further to 82.45% for the narcolepsy group. Similarly, in the SYSU dataset, the accuracy decreased to 78.36% for the depression group. This performance degradation may be attributed to the atypical and complex sleep patterns often exhibited by individuals with these conditions.

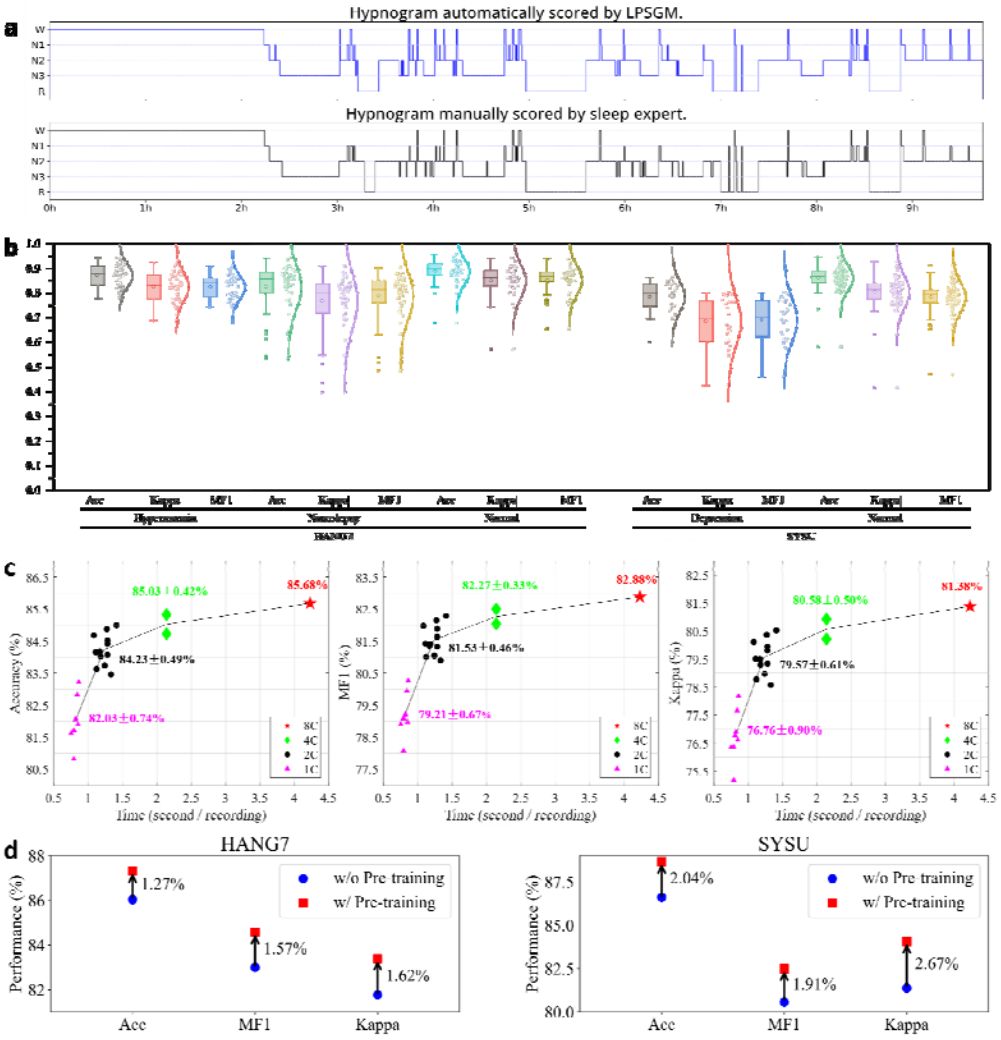

Fig. 2: Performance of LPSGM on sleep staging. (a) Hypnogram of a full-night PSG recording from the HANG7 dataset. The top panel presents sleep stages predicted by LPSGM, while the bottom panel displays the reference hypnogram manually scored by a sleep expert. (b) Boxplot distributions of accuracy (Acc), macro-F1 (MF1), and Cohen's Kappa (Kappa) for different subject groups in the HANG7 and SYSU datasets. (c) Performance-speed trade-off of LPSGM with different channel configurations on the HANG7 dataset. The x-axis represents the inference time required for processing a full-night PSG recording under different channel configurations (8C, 4C, 2C, and 1C). The y-axis indicates different performance metrics. (d) Five-fold cross-validation results on the HANG7 and SYSU datasets with and without large-scale hybrid pre-training. The blue circles represent models trained solely on the target center data (w/o pre-training), while the red squares indicate models pre-trained on public datasets before fine-tuning on the target center data (w/ pre-training).

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. 2c 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.

In addition to cross-center generalization, we investigate the impact of large-scale hybrid pretraining on sleep staging performance. Fig. 2d 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 SYSU dataset, pretraining led to significant improvements in all metrics, with accuracy increasing by 2.05%, macro-F1 by 1.91%, and Cohen's Kappa by 2.67%. On the HANG7 dataset, pretraining also resulted in notable improvements, with accuracy improving by 1.27%, macro-F1 by 1.57%, and Cohen's Kappa by 1.62%. 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.

Comparison with State-of-the-Art Sleep Staging 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 2. 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 dataset-specific and lack broader generalizability.

Table 2: 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**

a. 2C and 8C refer to the number of EEG channels used for model training. 2C includes the common channels across all datasets, C3-M2 and E1-M2. 8C consists of the full set of available channels, including C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, O2-M1, E1-M2, and E2-M1.

b. RobustSleepNet and LPSGM support flexible input configurations, allowing both 2C and 8C inputs. Other models are restricted to the 2C configuration.

Performance of Mental Disorder Diagnosis

In addition to sleep staging, we fine-tuned the LPSGM model on two disorder diagnosis tasks: narcolepsy diagnosis using the HANG7 dataset and depression detection using the SYSU dataset. Three fine-tuning strategies were applied and compared with models trained from scratch: partial fine-tuning updated only the classifier (Fig. 5a) while keeping other parameters frozen; full fine-tuning optimized the entire model for disorder diagnosis; and joint fine-tuning optimized the model for both sleep staging and disorder diagnosis simultaneously. The classification results for both disorders are shown in Fig. 3. Subject-wise diagnoses were determined by aggregating sample-wise predictions using majority voting. All results were obtained using five-fold cross-validation, ensuring no subject overlap between training and test sets. For narcolepsy diagnosis, the fine-tuned LPSGM achieved 88.01% accuracy, 90% sensitivity, and 84.29% specificity. For depression detection, it reached 100% accuracy, sensitivity, and specificity. Fine-tuning significantly improved PSG-based disorder diagnosis compared to training from scratch, with joint fine-tuning yielding the best overall performance. These findings demonstrate that integrating sleep staging with disease diagnosis tasks enhances the model's ability to detect sleep-related and psychiatric disorders.

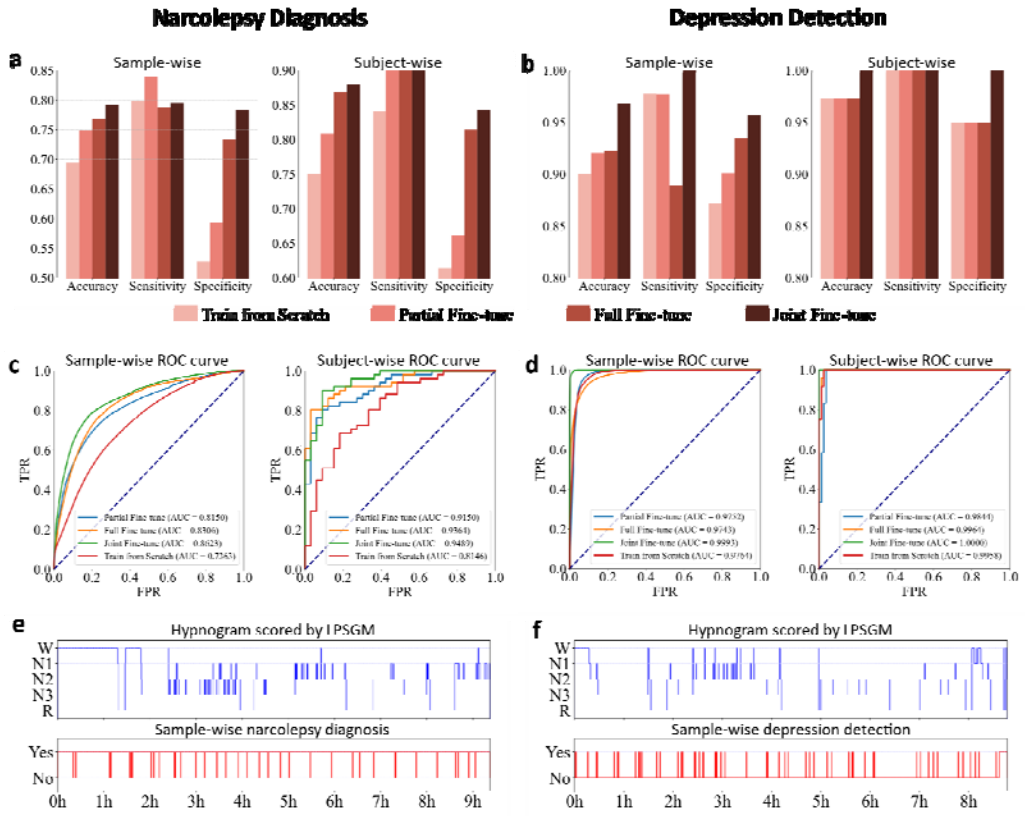

Fig. 3: Performance of fine-tuned LPSGM on narcolepsy diagnosis (left panels: a, c, e) and depression detection (right panels: b, d, f). The model was fine-tuned using three different strategies and compared against models trained from scratch. Partial

fine-tune, where only the classifier (Fig. 5a) is fine-tuned for the disorder diagnosis; full fine-tune, where the entire model is fine-tuned for disorder diagnosis; and joint fine-tune, where the entire model is fine-tuned simultaneously for sleep staging and disorder diagnosis. Panels (a) and (b) display classification performance (accuracy, sensitivity, and specificity) evaluated at both sample-wise and subject-wise levels. Panels (c) and (d) present receiver operating characteristic (ROC) curves for sample-wise and subject-wise classification. Panels (e) and (f) illustrate sample-wise predictions of a full-night PSG. The blue line shows hypnograms scored by LPSGM. The red line represents sample-wise disorder diagnosis, where "Yes" denotes samples classified as positive for the disorder, and "No" denotes negative classifications. The subject-wise diagnosis is determined via majority voting.

Prospective Study on Sleep Staging

To evaluate LPSGM's clinical applicability in sleep staging, a prospective study was conducted at HANG7 center, following the paradigm illustrated in Fig. 1c. Over a one-month period, 20 full-night PSG recordings were randomly selected and independently scored by three experts. LPSGM was then applied to the same PSG data for automatic sleep staging. Inter-expert and model vs. experts comparison were conducted using classification metrics and hypnogram metrics (see Supplementary Table 1 for definitions). For inter-expert comparison, each expert's annotations were evaluated against the consensus formed by the other two experts, and performance metrics were averaged across all experts to obtain an overall score. For model vs. expert comparison, the overall score was evaluated by comparing LPSGM's predictions against the consensus formed by all experts.

The results of classification metrics are summarized in Table 3. An extended definition of accuracy and Cohen's kappa based on probability distributions was used, and the consensus was determined by majority voting for each epoch. In cases where no clear majority was reached, probabilities were evenly distributed among the majority classes. LPSGM achieved accuracy of 89.2% and kappa of 0.845, closely matching experts' performance (overall accuracy of 88.7%, kappa of 0.85). The statistical significance of differences between LPSGM and experts was assessed using *t*-test, yielding *p*-values of 0.74 for accuracy and 0.60 for kappa, indicating no significant difference between LPSGM and experts. Notably, LPSGM exhibited lower standard deviation than individual experts, suggesting greater stability across PSG recordings.

Hypnogram metrics quantify sleep architecture from stage transitions and are critical components of clinical sleep reports. Table 4 categorizes these metrics into three major groups: sleep latency, measuring the time taken to enter specific sleep stages; sleep duration, indicating the total time spent in different sleep states; and sleep stage distributions, reflecting the proportion of each stage relative to the total sleep period. For each scorer (LPSGM or expert), MAE was calculated by averaging the absolute differences between their results and those of all other scorers contributing to the consensus. In the model vs. experts comparison, MAE represents

the mean absolute difference between LPSGM and each of the three experts. In the inter-expert comparison, MAE reflects the average absolute difference between an individual expert and the other two experts contributing to the consensus. *T*-test comparing model and experts revealed significant differences in N2 sleep latency ($p < 0.05$), as well as N1 duration and N1 percentage of total sleep time (TST) ($p < 0.01$). These discrepancies suggest that while LPSGM effectively captures global sleep structure, further refinements in detecting N1 and N2 could improve its clinical applicability.

Table 3: Comparison of classification metrics between LPSGM and experts.

		Overall	Expert 1	Expert 2	Expert 3
Inter-Expert	Accuracy	88.7±9.7	89.1±10.3	88.5±9.2	88.5±9.4
	Kappa	85.0±10.2	87.2±5.0	83.9±11.7	84.0±12.0
Model vs. Experts	Accuracy (With Expert)	89.2±4.6	88.8±3.7	87.5±8.8	89.1±3.3
	Accuracy (Excluding Expert)	—	88.3±4.8	89.1±3.4	87.4±8.8
	Kappa (With Expert)	84.5±6.6	84.3±5.1	82.7±11.0	84.4±4.5
	Kappa (Excluding Expert)	—	83.3±6.8	84.5±4.6	82.4±10.9

a. T-test results showed no significant difference between LPSGM and experts ($p = 0.74$ for accuracy, $p = 0.60$ for kappa).
b. Consensus was obtained by majority voting; when ties occurred, probabilities were evenly distributed. Accuracy and kappa were computed using an extended probabilistic definition.
c. "Inter-Expert" compares each expert with the consensus of the other two. "Model vs. Experts (With Expert)" compares LPSGM with an individual expert, while "Model vs. Experts (Excluding Expert)" compares LPSGM with the consensus of the other two experts.

Table 4: Mean absolute error (MAE) comparison of hypnogram metrics between LPSGM and experts.

Hypnogram Metrics		T-test *($p < 0.05$) **($p < 0.01$)	Model vs. Experts	Inter-Expert			
				Overall	Expert 1	Expert 2	Expert 3
Sleep Latency (min)							
	N1	—	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	N2	*	13.4±19.6	2.9±5.5	1.8±2.9	3.1±5.8	3.9±8.4
	N3	—	22.8±31.2	7.7±14.7	8.7±21.3	6.2±11.6	8.1±12.8
	REM	—	16.7±21.1	11.6±18.7	11.5±22.2	9.1±14.1	14.1±22.9
Sleep Duration (min)							
	TST	—	14.3±34.8	16.1±35.0	15.2±32.3	10.0±20.9	23.0±52.8
	REM	—	8.4±7.1	7.9±9.7	6.9±9.4	8.8±13.7	8.1±9.8
	NREM	—	11.2±10.9	13.8±22.4	13.2±17.6	15.6±34.1	12.6±18.0
	SWS	—	12.8±11.5	13.8±8.2	10.0±7.8	19.7±12.6	11.8±7.8
Sleep Stages							
W	Episodes (#)	—	5.0±4.4	4.7±3.2	5.4±4.2	4.0±3.1	4.9±5.3

(SPT)	Duration (min)	—	19.0±36.9	16.1±35.8	15.4±32.8	10.2±21.3	22.6±54.0
R	Duration (min)	—	8.4±7.1	7.9±9.7	6.9±9.4	8.8±13.7	8.1±9.8
	TST (%)	—	2.0±1.4	1.8±1.9	2.0±2.5	2.0±2.4	1.5±1.8
N1	Duration (min)	**	22.3±10.6	6.9±4.3	7.9±6.5	7.3±6.9	5.6±4.6
	TST (%)	**	6.2±2.8	1.9±1.2	2.2±1.7	1.8±1.9	1.6±1.3
N2	Duration (min)	—	25.9±17.2	17.5±15.0	15.4±12.9	22.7±23.7	14.3±14.0
	TST (%)	—	6.7±4.8	4.6±3.4	4.6±5.3	5.6±3.7	3.6±2.9
N3	Duration (min)	—	12.8±11.5	13.8±8.2	10.0±7.8	19.7±12.6	11.8±7.8
	TST (%)	—	4.4±3.8	3.7±2.5	3.0±3.1	4.9±3.5	3.4±2.5

a. Bolded text represents hypnogram metric categories. See Supplementary Table 1 for detailed definitions.
b. MAE was computed by averaging the absolute differences between each scorer's results and those of the remaining scorers contributing to the consensus.
c. T-test evaluates the statistical significance of differences between MAE of model and experts (overall).
d. TST (%) represents the proportion of total sleep time spent in a given stage.

Interpretability Analysis

Deep learning models are often considered black boxes, lacking interpretability, which poses challenges for clinical applications where decision transparency is desirable. To analyze LPSGM's decision-making, we applied gradient-weighted class activation mapping (GradCAM) [31], a widely used technique for visualizing model attention. GradCAM generates heatmaps indicating the most influential regions in the input for a given prediction. We applied GradCAM to the final convolutional layer of the Epoch Encoder's dual-branch CNN (Fig. 5b), resized the activation maps to match the EEG input resolution, and combined outputs from both branches to generate a comprehensive visualization of model attention during sleep staging.

Fig. 4 presents GradCAM visualizations for six 30-second EEG epochs, where panels (a)-(d) correspond to correct classifications and panels (e)-(f) represent misclassifications. In correctly predicted cases, highlighted regions align with clinically relevant EEG features, as defined by AASM guidelines. For instance, panel (a) correctly identifies posterior dominant rhythm (PDR) in occipital channels, characteristic of Wake. Panel (b) highlights slow wave activity in frontal regions, consistent with N3 sleep. Similarly, panel (c) captures low-amplitude mixed-frequency (LAMF) EEG and slow eye movements (SEM), indicative of N1 sleep, while panel (d) correctly detects rapid eye movements (REMs), crucial for classifying REM sleep. However, in panels (e)-(f), misclassifications occur, where model attention diverges from key clinical features. In these cases, the highlighted regions do not coincide with clinically relevant features, suggesting the model's attention may have been misdirected or that the features needed for accurate classification were insufficient.

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.

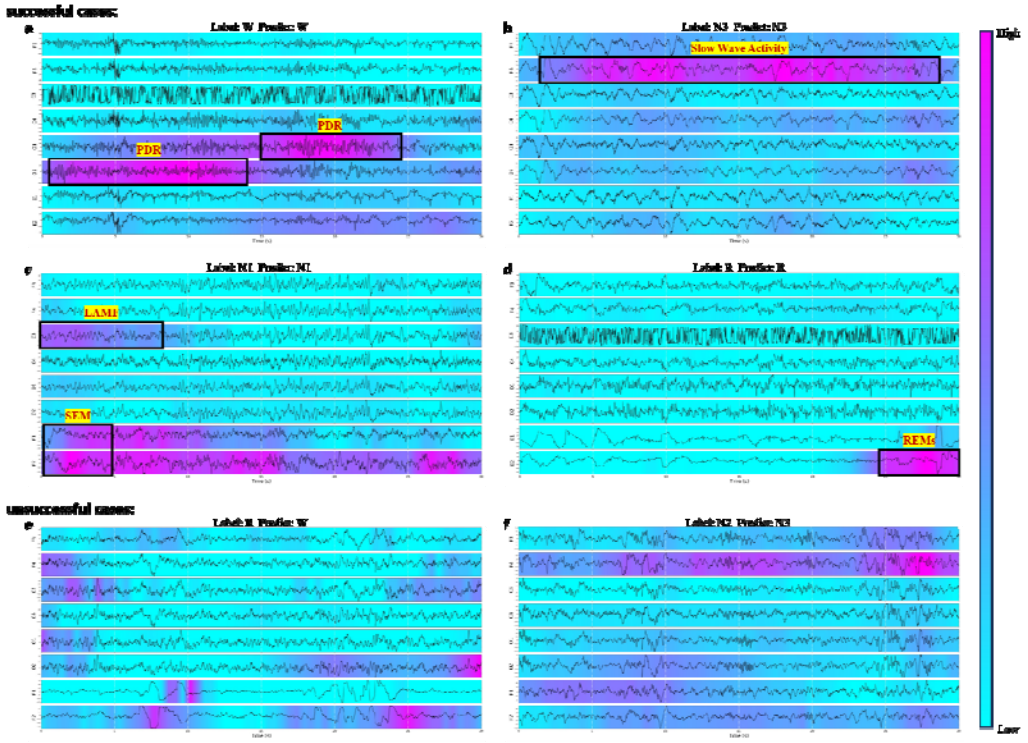

Fig. 4: GradCAM visualizations of LPSGM predictions for sleep staging. The figure presents six 30-second EEG epochs analyzed using GradCAM to highlight model-attended regions across eight EEG channels (F3, F4, C3, C4, O1, O2, E1, E2). Panels (a)-(d) depict correctly classified epochs, while panels (e)-(f) show misclassifications. Highlighted regions indicate features the model deemed relevant for classification. Black boxes mark physician-identified features consistent with AASM sleep staging criteria. (a) Posterior dominant rhythm (PDR) in occipital channels (>50% of the epoch), indicative of Wake. (b) Slow wave activity in frontal channels (>20% of the epoch), characteristic of N3. (c) Absence of PDR, low-amplitude mixed-frequency (LAMF) EEG, and slow eye movements (SEM), consistent with N1. (d) Rapid eye movements (REMs), indicative of REM sleep. (e)-(f) Misclassified epochs where model attention diverges from clinically relevant features.

Ablation Study

To validate the key architectural choices in LPSGM, we conducted an ablation study focusing on two core components: channel & temporal encoding, and feature fusion. These innovations, described in detail in the Methods section, are designed to enhance the model's ability to process heterogeneous PSG datasets. 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 5, 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 5: 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 6. 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 6: 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

Discussion

LPSGM presents a significant advancement in automated sleep staging and mental disorder diagnosis using polysomnography (PSG). By leveraging large-scale, cross-center hybrid training on 220,500 hours of data from 16 public datasets, LPSGM demonstrates superior generalization capabilities, achieving performance on par with models trained directly on target-center data. The flexible channel input adaptation further enhances its applicability across diverse clinical settings, addressing key challenges such as inter-center variability and data scarcity.

Our results indicate that LPSGM achieves high cross-center classification performance in sleep staging, reaching 99.6% and 97.1% of the upper-bound performance on HANG7 and SYSU datasets, respectively. Additionally, prospective validation confirms that LPSGM performs at an expert level, with accuracy and kappa scores statistically indistinguishable from human scorers. These findings underscore its potential as a scalable, plug-and-play solution for sleep staging in both research and clinical practice.

Beyond sleep staging, LPSGM extends its utility to mental disorder diagnosis. Fine-tuning the model for narcolepsy and depression detection yielded promising results, achieving 88.01% accuracy for narcolepsy diagnosis and 100% accuracy for depression screening. These findings support the feasibility of PSG-based biomarkers for psychiatric conditions, offering an objective and scalable alternative to traditional diagnostic methods.

Despite these advancements, several limitations should be acknowledged. First, while LPSGM demonstrates strong generalization, its performance slightly declines in patients with sleep disorders or neurological conditions, likely due to the atypical sleep patterns in these populations. Future work should incorporate more diverse clinical cohorts to further enhance robustness. Second, although LPSGM supports variable-channel configurations, performance degradation occurs when reducing input channels, particularly in single-channel settings. Optimizing feature extraction for minimal-channel PSG recordings could facilitate deployment in resource-constrained settings and home-based monitoring. Third, while our interpretability analysis via GradCAM highlights the model's reliance on clinically relevant features, further improvements in explainability could enhance clinical trust and adoption. Integrating attention-based visualization methods and clinician-in-the-loop feedback mechanisms may help refine decision-making transparency.

Future research directions include expanding LPSGM's capabilities to additional sleep-related and neurological disorders. Additionally, incorporating multi-modal physiological signals, including heart rate variability and respiration, may improve diagnostic accuracy. Finally, adapting LPSGM for use with wearable or portable EEG

devices could enable real-world, at-home sleep monitoring, making PSG-based assessments more accessible.

In conclusion, LPSGM represents a scalable and flexible solution for automated sleep staging and PSG-based disorder diagnosis. By addressing key limitations of existing models and integrating domain-adaptive learning, LPSGM bridges the gap between automated PSG analysis and clinical deployment. Continued refinement and validation in broader clinical settings will further solidify its role in advancing sleep medicine and psychiatric diagnostics.

Methods

Overview

LPSGM is a sequence-to-sequence model for automatic sleep staging from multi-channel PSG signals (Fig. 5a). It comprises three components: an Epoch Encoder, which extracts local intra-epoch features; a Sequence Encoder, which captures global inter-epoch dependencies; and a Classifier, which assigns sleep stages based on the encoded representations. The model processes sequential 30-second PSG epochs and predicts the corresponding sleep stages.

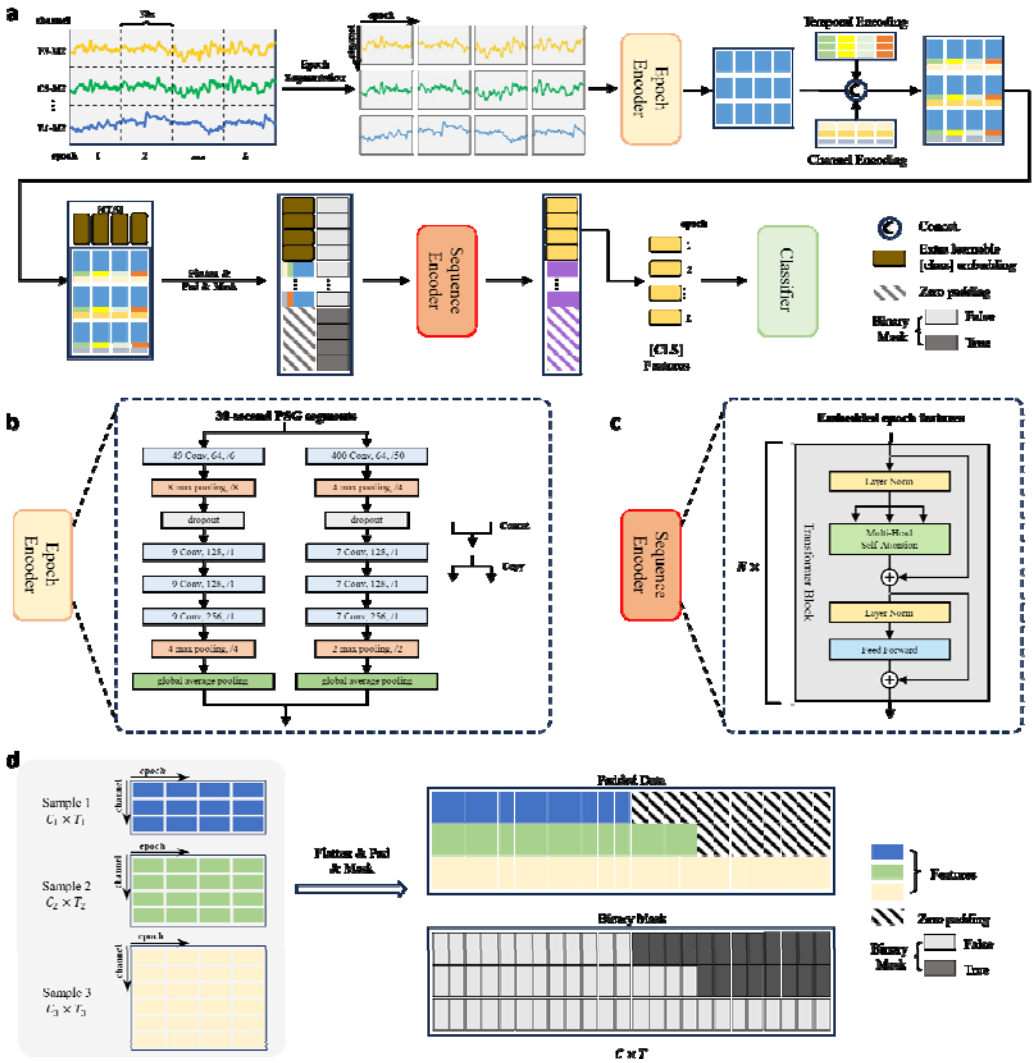

Fig. 5: Overall architecture of LPSGM. (a) LPSGM consists of an Epoch Encoder, Sequence Encoder, and Classifier, designed for both sleep staging and disorder diagnosis. (b) The Epoch Encoder employs a dual-branch CNN to extract local intra-epoch features from each 30-second PSG segment, using small and large convolutional filters to capture high- and low-frequency EEG features, respectively. (c) The Sequence Encoder consists of a series of Transformer blocks to capture temporal dependencies across epochs in the sleep sequence. Each Transformer block consists of multi-head self-attention (MSA), feed-forward networks (FFN), and layer normalization (LN). (d) Padding and masking strategy implemented to handle samples with varying numbers of EEG channels, ensuring compatibility across different PSG datasets.

Epoch Segmentation

To accommodate varying channel configurations, PSG signals are first segmented into 30-second epochs. Given a multi-channel PSG sequence

with L epochs, each epoch $x_i^l \in \mathbb{R}^{C_i \times n}$ contains C_i channels and $n = 30 \times f_s$ samples. Segmentation is performed along both the temporal and channel dimensions, producing $L \times C_i$ single-channel segments $x_i^{l,c} \in R^n$. Each segment represents a single 30-second epoch from one channel. By performing epoch segmentation, we standardize the input format, facilitating the subsequent encoding and classification steps within the LPSGM framework.

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. 5b.

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 $CE = \{ce_1, ce_2, \dots, ce_{|C|}\}$ and a temporal embedding list $TE = \{te_1, te_2, \dots, te_L\}$. 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:

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

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 $L \times C_i$. 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.

To overcome these limitations, we drew inspiration from Natural Language Processing (NLP) techniques and designed a padding and masking strategy. As illustrated in Fig. 5d, 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:

$$M_{i,j} = \begin{cases} 1, & j > C_i \times T_i \\ 0, & j \leq C_i \times T_i \end{cases} \# (2)$$

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.

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 $\widetilde{e_i}$. Each CLS token possesses the same

dimension as the encoded feature vectors (i.e., $d + d_{ce} + d_{te}$), resulting in an extended sequence $\widetilde{e_i}'$ of length $L \times (C_i + 1)$. The CLS tokens are initialized randomly and undergo learning during training.

Sequence Encoder

The role of the 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:

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

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

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

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}:

$$E_{cls} = E_{out}[0{:}\,L] \#(4)$$

Classifier

Two classifiers are used: one for sleep staging and another for disease diagnosis. Each classifier consists of a fully connected layer with a softmax function. 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. The disease diagnosis classifier is a binary classifier designed to predict the presence or absence of a specific disorder. 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.

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:

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

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 512. 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.

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.

Data Availability

See Supplementary Methods and Supplementary Tables 2 and 3 for a detailed description of the data.

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.

Code Availability

The code of our proposed LPSGM is made available in GitHub: https://github.com/Deng-GuiFeng/LPSGM.

References

1. Maquet, P.. The role of sleep in learning and memory. science 294, 1048--1052 (2001).
2. Irwin, M.. Why sleep is important for health: a psychoneuroimmunology perspective. Annual review of psychology 66, 143--172 (2015).
3. Hassan, A., Bhuiyan, M.. Automatic sleep stage classification. 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT) 211--216 (2015).
4. Rahman, M., Hossain, M., Kabir, M., Sani, M., Awal, M., others. Optimization of sleep stage classification using single-channel eeg signals. 2019 4th International Conference on Electrical Information and Communication Technology (EICT) 1--6 (2019).
5. Hassan, A., Bhuiyan, M.. A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. Journal of neuroscience methods 271, 107--118 (2016).
6. Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., Dickhaus, H.. 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 108, 10--19 (2012).
7. {\c{S. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. Journal of medical systems 38, 1--21 (2014).
8. Zhao, S., Long, F., Wei, X., Ni, X., Wang, H., Wei, B.. Evaluation of a single-channel EEG-based sleep staging algorithm. International Journal of Environmental Research and Public Health 19, 2845 (2022).
9. Touil, M., Bahatti, L., El Magri, A.. Sleep's depth detection using electroencephalogram signal processing and neural network classification. Journal of Medical Artificial Intelligence 5, (2022).

10. Supratak, A., Dong, H., Wu, C., Guo, Y.. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. IEEE transactions on neural systems and rehabilitation engineering 25, 1998--2008 (2017).
11. Supratak, A., Guo, Y.. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine \& Biology Society (EMBC) 641--644 (2020).
12. Biswal, S., Sun, H., Goparaju, B., Westover, M., Sun, J., Bianchi, M.. Expert-level sleep scoring with deep neural networks. Journal of the American Medical Informatics Association 25, 1643--1650 (2018).
13. Phan, H., Ch{\'e. XSleepNet: Multi-view sequential model for automatic sleep staging. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 5903--5915 (2021).
14. Yoo, C., Lee, H., Kang, J.. Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network. IEEE journal of biomedical and health informatics 26, 1273--1284 (2021).
15. Fan, J., Zhu, H., Jiang, X., Meng, L., Chen, C., Fu, C., Yu, H., Dai, C., Chen, W.. Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. IEEE Transactions on Neural Systems and Rehabilitation Engineering 30, 205--216 (2022).
16. Guillot, A., Thorey, V.. RobustSleepNet: Transfer learning for automated sleep staging at scale. IEEE Transactions on Neural Systems and Rehabilitation Engineering 29, 1441--1451 (2021).
17. Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., Zou, J.. Adversarial training helps transfer learning via better representations. Advances in Neural Information Processing Systems 34, 25179--25191 (2021).
18. Wang, J., Zhao, S., Jiang, H., Li, S., Li, T., Pan, G.. Generalizable sleep staging via multi-level domain alignment. Proceedings of the AAAI Conference on Artificial Intelligence 265--273 (2024).
19. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., others. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 , (2023).
20. Lewis, M.. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 , (2019).
21. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., others. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 , (2024).
22. Mohsenvand, M., Izadi, M., Maes, P.. Contrastive representation learning for electroencephalogram classification. Machine Learning for Health 238--253 (2020).
23. Han, J., Wei, X., Faisal, A.. EEG decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks. Journal of Neural Engineering 20, 066027 (2023).

24. Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., Pan, G.. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. arXiv preprint arXiv:2412.07236 , (2024).
25. Li, R., Wang, Y., Zheng, W., Lu, B.. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. Proceedings of the 30th ACM International Conference on Multimedia 6--14 (2022).
26. Yi, K., Wang, Y., Ren, K., Li, D.. Learning topology-agnostic eeg representations with geometry-aware modeling. Advances in Neural Information Processing Systems 36, (2024).
27. Jiang, W., Zhao, L., Lu, B.. Large brain model for learning generic representations with tremendous EEG data in BCI. arXiv preprint arXiv:2405.18765 , (2024).
28. Dosovitskiy, A.. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 , (2020).
29. Perslev, M., Jensen, M., Darkner, S., Jennum, P.. U-time: A fully convolutional network for time series segmentation applied to sleep staging. Advances in Neural Information Processing Systems 32, (2019).
30. Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C., Li, X., Guan, C.. An attention-based deep learning approach for sleep stage classification with single-channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering 29, 809--818 (2021).
31. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.. Grad-CAM: visual explanations from deep networks via gradient-based localization. International journal of computer vision 128, 336--359 (2020).
32. Zhang, G., 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. Journal of the American Medical Informatics Association 25, 1351--1358 (2018).
33. Quan, S., Chan, C., Dement, W., Gevins, A., Goodwin, J., Gottlieb, D., Green, S., Guilleminault, C., Hirshkowitz, M., Hyde, P., others. The association between obstructive sleep apnea and neurocognitive performance—the Apnea Positive Pressure Long-term Efficacy Study (APPLES). Sleep 34, 303--314 (2011).
34. Bakker, J., Tavakkoli, A., Rueschman, M., Wang, W., Andrews, R., Malhotra, A., Owens, R., Anand, A., Dudley, K., Patel, S.. Gastric banding surgery versus continuous positive airway pressure for obstructive sleep apnea: a randomized controlled trial. American journal of respiratory and critical care medicine 197, 1080--1083 (2018).
35. Rosen, C., 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 35, 757--767 (2012).
36. Quan, S., Howard, B., Iber, C., Kiley, J., Nieto, F., O'Connor, G., Rapoport, D.,

Redline, S., Robbins, J., Samet, J., others. The sleep heart health study: design, rationale, and methods. Sleep 20, 1077--1085 (1997).

37. Wang, R., Bakker, J., Chervin, R., Garetz, S., Hassan, F., Ishman, S., Mitchell, R., Morrical, M., Naqvi, S., Radcliffe, J., others. 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 10, e033889 (2020).
38. Redline, S., Cook, K., Chervin, R., Ishman, S., Baldassari, C., Mitchell, R., Tapia, I., Amin, R., Hassan, F., Ibrahim, S., others. Adenotonsillectomy for snoring and mild sleep apnea in children: a randomized clinical trial. JAMA 330, 2084--2095 (2023).
39. Marcus, C., Moore, R.. A randomized trial of adenotonsillectomy for childhood sleep apnea. New England Journal of Medicine 368, 2366--2376 (2013).
40. Rosen, C., Larkin, E., Kirchner, H., Emancipator, J., Bivins, S., Surovec, S., Martin, R., Redline, S.. Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. The Journal of pediatrics 142, 383--389 (2003).
41. Redline, S., Tishler, P., Tosteson, T., Williamson, J., Kump, K., Browner, I., Ferrette, V., Krejci, P.. The familial aggregation of obstructive sleep apnea.. American journal of respiratory and critical care medicine 151, 682--687 (1995).
42. Stephansen, J., Olesen, A., Olsen, M., Ambati, A., Leary, E., Moore, H., Carrillo, O., Lin, L., Han, F., Yan, H., others. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nature communications 9, 5229 (2018).
43. Lee, H., Li, B., DeForte, S., Splaingard, M., Huang, Y., Chi, Y., Linwood, S.. A large collection of real-world pediatric sleep studies. Scientific Data 9, 421 (2022).
44. Blackwell, T., Yaffe, K., Ancoli-Israel, S., Redline, S., Ensrud, K., Stefanick, M., Laffan, A., Stone, K., 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. Journal of the American Geriatrics Society 59, 2217--2225 (2011).
45. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., Stanley, H.. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation 101, e215--e220 (2000).
46. Alvarez-Estevez, D., Rijsman, R.. Haaglanden medisch centrum sleep staging database (version 1.0. 1). PhysioNet , (2021).
47. Ghassemi, M., Moody, B., Lehman, L., Song, C., Li, Q., Sun, H., Mark, R., Westover, M., Clifford, G.. You snooze, you win: the physionet/computing in cardiology challenge 2018. 2018 Computing in Cardiology Conference (CinC) 1--4 (2018).
48. Terzano, M., Parrino, L., Sherieri, A., Chervin, R., Chokroverty, S., Guilleminault, C., Hirshkowitz, M., Mahowald, M., Moldofsky, H., Rosa, A., others. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in

human sleep. Sleep medicine 2, 537--554 (2001).
49. Khalighi, S., Sousa, T., Santos, J.. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. Computer methods and programs in biomedicine 124, 180--192 (2016).
50. Guillot, A., Sauvet, F., During, E., Thorey, V.. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. IEEE transactions on neural systems and rehabilitation engineering 28, 1955--1965 (2020).
51. Rosenberg, R., Van Hout, S.. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. Journal of clinical sleep medicine 9, 81--87 (2013).
52. Zhang, X., Dong, X., Kantelhardt, J., Li, J., Zhao, L., Garcia, C., Glos, M., Penzel, T., Han, F.. Process and outcome for international reliability in sleep scoring. Sleep and Breathing 19, 191--195 (2015).
53. Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., others. Interrater reliability for sleep scoring according to the Rechtschaffen \& Kales and the new AASM standard. Journal of sleep research 18, 74--84 (2009).
54. MacLean, A., Lue, F., Moldofksy, H.. The reliability of visual scoring of alpha EEG activity during sleep. Sleep 18, 565--569 (1995).
55. do Kim, Y., Kurachi, M., Horita, M., Matsuura, K., Kamikawa, Y.. Agreement of visual scoring of sleep stages among many laboratories in Japan: effect of a supplementary definition of slow wave on scoring of slow wave sleep. Psychiatry and Clinical Neurosciences 47, 91--97 (1993).
56. Xu, S., Faust, O., Seoni, S., Chakraborty, S., Barua, P., Loh, H., Elphick, H., Molinari, F., Acharya, U.. A review of automated sleep disorder detection. Computers in Biology and Medicine 150, 106100 (2022).
57. Quinnell, T., Smith, I.. Narcolepsy, idiopathic hypersomnolence and related conditions. Clinical medicine 11, 282 (2011).
58. Lopez, J., Hoffmann, R., Armitage, R.. Reduced sleep spindle activity in early-onset and elevated risk for depression. Journal of the American Academy of Child \& Adolescent Psychiatry 49, 934--943 (2010).
59. Riemann, D., Krone, L., Wulff, K., Nissen, C.. Sleep, insomnia, and depression. Neuropsychopharmacology 45, 74--89 (2020).
60. Kupfer, D.. REM latency: a psychobiologic marker for primary depressive disease.. Biological psychiatry 11, 159--174 (1976).
61. Kupfer, D., Foster, F.. Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression. The Lancet 300, 684--686 (1972).

Acknowledgements

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）

Author Contributions

G.D. and H.J. designed the methodology. G.D. conducted the experiments and performed the primary analysis. G.D. and M.N. carried out the interpretability analysis. S.R., J.X. and G.P. contributed to the analysis and interpretation of results. M.N., J.S., and Z.Y. collected, organized, and preprocessed the HANG7 dataset. Y.L. collected, organized, and preprocessed the SYSU dataset. W.L., X.L., W.D., W.G., and T.L. led the prospective study.

Competing Interests

The authors declare no competing interests.

Additional Information