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

Guifeng Deng [1,2], Mengfan Niu[1], Shuying Rao[1,2], Yuxi Luo[3], Jianjia Zhang[3], Junyi Xie[1], Zhenghe Yu[1], Wenjuan Liu[1], Junhang Zhang[1], Sha Zhao[4], Gang Pan[4], Xiaojing Li[1,4], Wei Deng[1,4], Wanjun Guo[1,4], Yaoyun Zhang[5], Tao Li [1,4*], Haiteng Jiang [1,4,6*]

[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] School of Biomedical Informatics, University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA.
[6] Department of Psychiatry and Mental Health, Wenzhou Medical University, Wenzhou 325035, Zhejiang Province, China.

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

Abstract

Sleep disorders affect billions worldwide, yet clinical polysomnography (PSG) analysis remains hindered by labor-intensive manual scoring and limited generalizability of automated sleep staging tools across heterogeneous protocols. We present LPSGM, a large-scale PSG model designed to address two critical challenges in sleep medicine: cross-center generalization and adaptable diagnosis of neuropsychiatric disorders. Trained on 220,500 hours of multi-center PSG data (24,000 full-night recordings from 16 public datasets), LPSGM integrates domain-adaptive pre-training, flexible channel configurations, and a unified architecture to mitigate variability in equipment, montages, and populations during sleep staging while enabling downstream fine-tuning for mental disorder detection. In prospective validation, LPSGM achieves expert-level consensus in sleep staging ($\kappa = 0.845 \pm 0.066$ vs. inter-expert $\kappa = 0.850 \pm 0.102$) and matches the performance of fully supervised models on two independent private cohorts. When fine-tuned, it attains 88.01% accuracy in narcolepsy detection and 100% accuracy in identifying major depressive disorder (MDD), highlighting shared physiological biomarkers between sleep architecture and neuropsychiatric symptoms. By bridging automated sleep staging with real-world clinical deployment, LPSGM establishes a scalable, data-efficient framework for integrated sleep and mental health diagnostics. The code and pre-trained model are publicly available at https://github.com/Deng-GuiFeng/LPSGM to advance reproducibility and translational research in sleep medicine.

Keywords: Polysomnography (PSG); Sleep staging; Narcolepsy; Major depressive disorder (MDD); Domain adaptation

Introduction

Sleep occupies nearly one-third of human life and serves as a critical determinant of overall health[1,2]. The escalating global burden of sleep disorders—including insomnia, obstructive sleep apnea (OSA), and narcolepsy—represents a major public health challenge. Epidemiological studies indicate that OSA affects approximately one billion adults worldwide[3], while insomnia disorder impacts 10% of the adult population, with an additional 20% experiencing transient symptoms[4].

Polysomnography (PSG), the gold-standard tool for sleep assessment, captures multimodal physiological signals—including electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) to analyze sleep architecture according to American Academy of Sleep Medicine (AASM) guidelines[5], which classify sleep into wakefulness (W), non-rapid eye movement (NREM: N1–N3), and rapid eye movement (REM) stages. By mapping these stages, PSG reveals how deviations in sleep architecture—such as shortened REM latency in narcolepsy[6,7] or fragmented sleep, reduced slow-wave activity and disinhibition REM sleep in MDD[8–11]—serve as biomarkers for neuropsychiatric conditions, positioning it as a dual-purpose tool for both sleep staging and mental health diagnostics.

Conventional manual PSG scoring requires clinicians to annotate 30-second epochs, a process requiring approximately two hours per full-night recording (7–9 hours of PSG data). This labor-intensive method suffers from significant inter-rater variability due to subjective interpretation of ambiguous patterns, undermining diagnostic consistency across clinical setting[12–16]. Machine learning advances have partially addressed these inefficiencies through automated sleep staging[17–34]. Early approaches relied on hand-engineered features and traditional machine learning models (e.g., support vector machines, random forests), but struggled to capture complex patterns of sleep physiology[17–23]. Subsequent deep learning frameworks—notably convolutional and recurrent neural networks (CNNs/RNNs)—enabled end-to-end learning of hierarchical representations directly from raw physiological data, achieving accuracy comparable to human experts[24–34]. However, current studies prioritize intra-dataset performance using fixed PSG channel configurations, exhibiting marked performance degradation in cross-domain settings due to variability in PSG protocols and population demographics. Transfer learning strategies[28–32]—including domain adaptation[28,29,32] and adversarial training[31]—improve out-of-distribution (OOD) generalization but persistently underperform models trained on target-domain data. These limitations underscore the urgent need for robust, generalizable frameworks capable of adapting to heterogeneous clinical environments.

Despite recent progress, clinical adoption of automated sleep staging systems remains constrained by two interrelated challenges: domain shift and data scarcity. PSG heterogeneity—driven by variability in recording protocols, equipment, and populations—contravenes the independent and identically distributed (IID) assumption critical to conventional deep learning, while the labor-intensive nature of manual annotation impedes scalable dataset curation. Although numerous public sleep datasets exist[6,35–52], their collective utility is diminished by inconsistent signal montages and channel configurations (Supplementary Table 2). Drawing inspiration from recent advances in large language models (LLMs) [53–55] and EEG foundation models[56–58], we propose that a unified framework integrating multi-center PSG data through domain-adaptive pre-training could mitigate domain shifts while unlocking downstream clinical utility.

Here, we present LPSGM, a large PSG model designed to overcome two critical barriers in sleep medicine: (1) cross-center generalization in automated sleep staging and (2) adaptable PSG-based diagnosis of neuropsychiatric disorders. Our framework integrates three core innovations: (i) Unified training protocol for harmonizing heterogeneous datasets with variable montages and channel configurations, (ii) Flexible inference architecture that dynamically adapts to diverse clinical PSG setups, (iii) Hybrid pre-training on 220,500 hours of PSG data (24,000 recordings) aggregated from 16 public datasets. LPSGM achieves performance parity with fully supervised models trained on two independent private datasets (HANG7 and SYSU; see supplementary materials for additional details). In prospective validation, it matches human expert consensus (model vs. experts: $\kappa = 0.845 \pm 0.066$; inter-expert agreement: $\kappa = 0.850 \pm 0.102$), demonstrating clinical-grade reliability. When

fine-tuned for disease detection, LPSGM attains 88.01% accuracy in narcolepsy diagnosis and 100% accuracy in identifying MDD, highlighting its versatility for downstream applications. By unifying robust cross-domain generalization with clinical adaptability, LPSGM bridges the gap between automated sleep staging and real-world deployment, enabling scalable, data-efficient analysis of sleep physiology and its neuropsychiatric correlates.

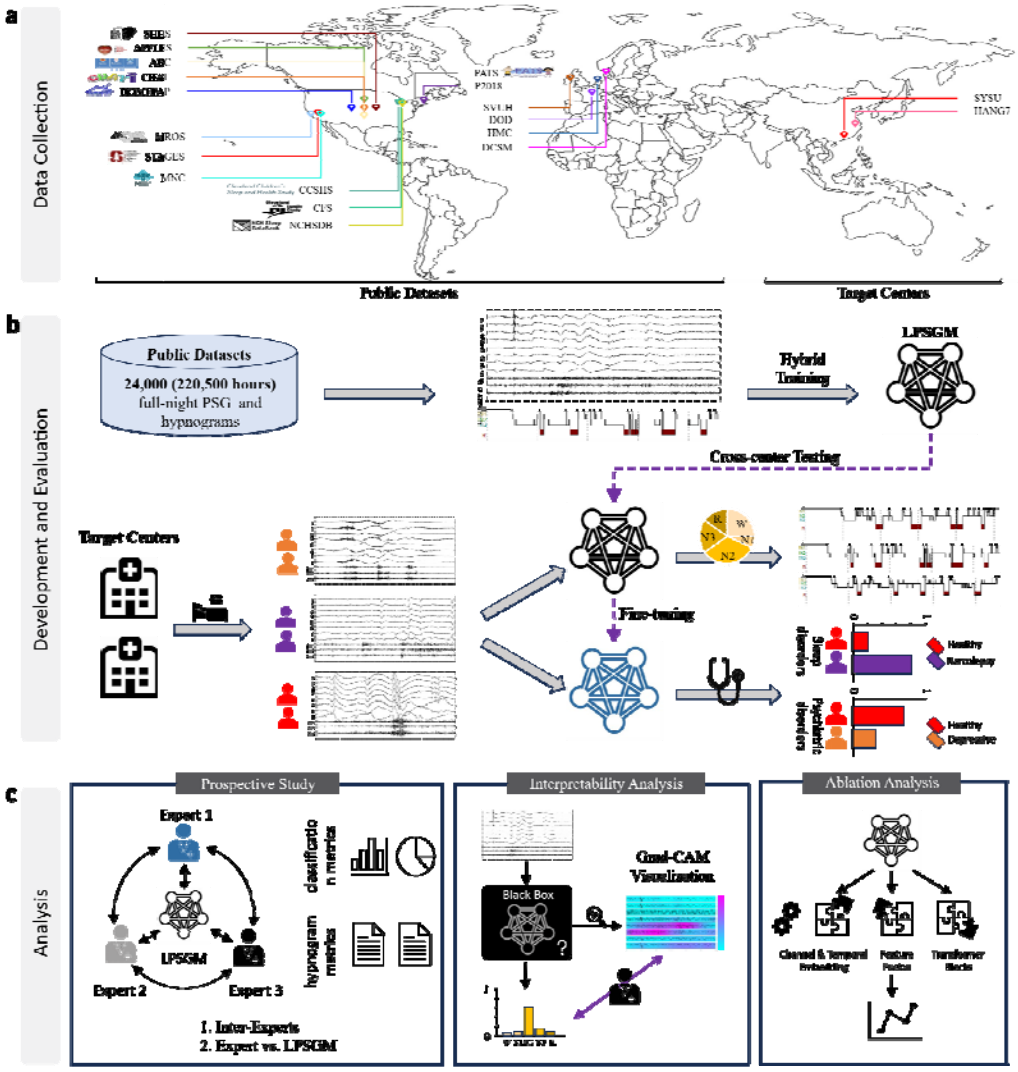

Fig. 1: Overview of the LPSGM framework. Panel (a): Data harmonization schematizes the aggregation of 220,500 hours of PSG data from 16 public datasets and 2 independent clinical cohorts, spanning diverse geographic populations and recording protocols. Panel (b): Cross-center generalization outlines the training-evaluation pipeline: LPSGM is pre-trained on multi-center public datasets, validated for cross-domain sleep staging on two unseen private datasets, and fine-tuned for downstream tasks including narcolepsy diagnosis and MDD detection. Panel (c): Analytical validation details the study's three-pronged evaluation: (1) a prospective clinical trial benchmarking LPSGM against expert consensus, (2) interpretability analysis to decode decision-making patterns, and (3) ablation studies

quantifying the contribution of key components.

Results

Cross-Center Classification Performance of Sleep Staging

We evaluated LPSGM's cross-center sleep staging performance on the HANG7 and SYSU datasets against two baselines (Table 1). Baseline 1 (lower bound) involved direct application of a model trained on one center's data to another without adaptation, while Baseline 2 (upper bound) used models fully trained on target-center data via five-fold cross-validation. LPSGM-Small outperformed LPSGM in baseline comparisons and was selected as the reference baseline. LPSGM achieved near-upper-bound performance on both datasets. For HANG7, LPSGM reached 85.68% accuracy, 82.88% macro-F1, and a kappa of 0.8138 (99.6%, 99.9%, and 99.5% of Baseline 2, respectively). On SYSU, it attained 84.13% accuracy, 77.88% macro-F1, and 0.7789 kappa (97.1%, 96.7%, and 95.7% of Baseline 2), demonstrating robust generalization across clinical centers.

Table 1: Comparison of the sleep staging results between LPSGM and baselines on two target private 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
	Base. 2	LPSGM	0.8568	0.8288	0.8138	0.9356	0.6348	0.8455	0.8636	0.8644
		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
	Base. 2	LPSGM	0.8413	0.7788	0.7789	0.7888	0.4688	0.8676	0.8703	0.8986
		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.

Notably, sleep staging performance exhibited variation across subject subgroups (Fig. 2b), suggesting potential demographic- or pathology-dependent effects on model generalizability. LPSGM achieved highest accuracy in healthy cohorts (88.99% on HANG7; 86.15% on SYSU) but declined slightly in populations with sleep or neurological disorders—notably, to 87.11% (hypersomnia) and 82.45% (narcolepsy) on HANG7, and 78.36% (MDD) on SYSU. This may reflect atypical sleep architecture in clinical populations.

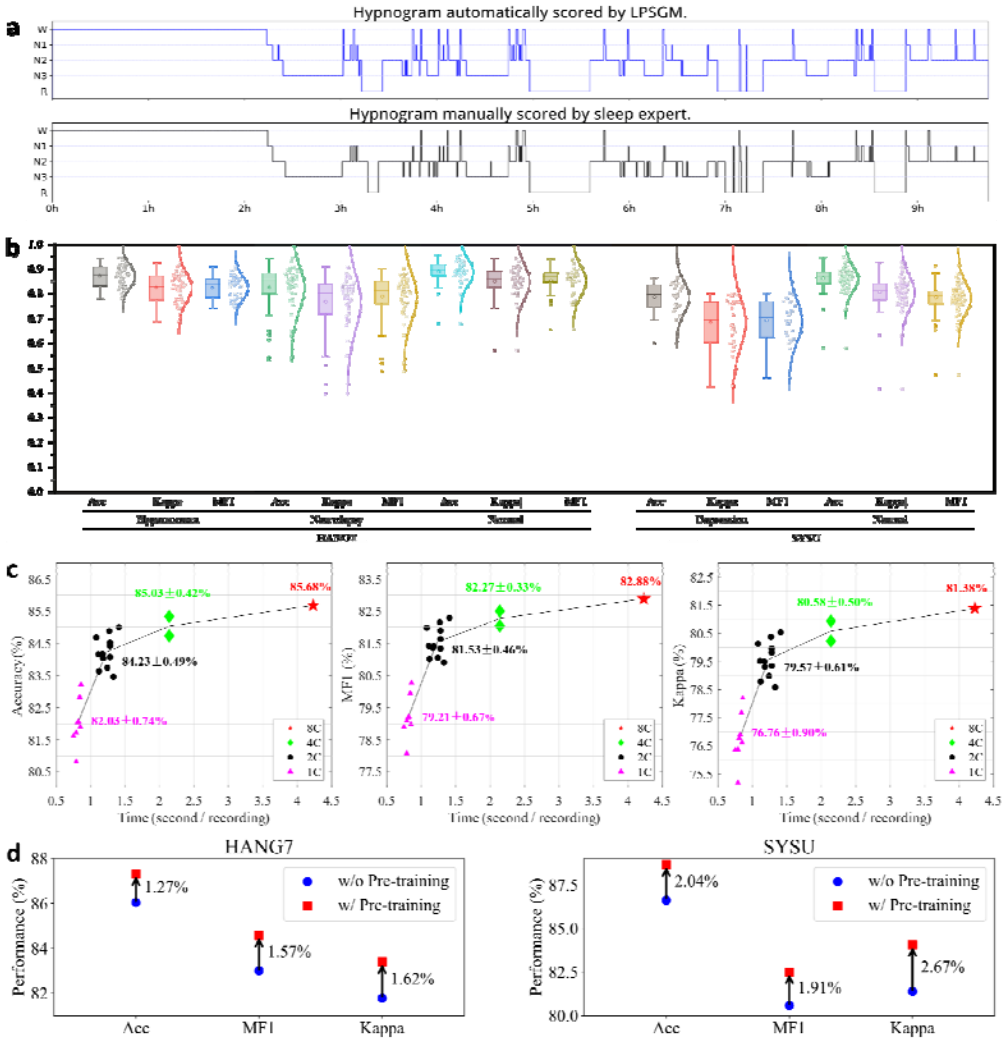

Fig. 2: Performance of LPSGM on Sleep Staging. (a) Hypnogram comparison of a full-night PSG recording from the HANG7 dataset. The top panel illustrates sleep stages predicted by LPSGM, while the bottom panel shows the expert-annotated reference hypnogram. (b) Boxplots summarizing distributions of accuracy (Acc), macro-F1 (MF1), and Cohen's κ across subject groups in the HANG7 and SYSU datasets. (c) Trade-off between model performance and inference speed under varying channel configurations (8C, 4C, 2C, and 1C; C = channels) on the HANG7 dataset. The x-axis denotes inference time per full-night recording, and the y-axis reports performance metrics (Acc, MF1, κ). (d) Five-fold cross-validation results on the

HANG7 and SYSU datasets, comparing models trained with and without large-scale hybrid pre-training. Blue circles denote models trained exclusively on target-center data (without pre-training), while red squares represent models pre-trained on public datasets before target-center fine-tuning (with pre-training).

During the inference stage, LPSGM dynamically balances accuracy and computational efficiency by adjusting 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. When using eight channels (8C), LPSGM achieves the highest accuracy of 85.68%, macro-F1 of 82.88%, and kappa of 81.38% , with an inference time of about 4 seconds per recording. As the number of channels decreases, the inference time reduces while there is a corresponding drop in performance metrics. When further reducing to two channels (2C), the accuracy is 84.23±0.49%, macro-F1 81.53±0.46%, kappa 79.57±0.61%, with the inference time dropping to about 1 second per recording. Using a single channel (1C), the model's accuracy drops more significantly to 82.03±0.74%, macro-F1 to 79.21±0.67%, and kappa to 76.76±0.90%, but the inference time is the fastest at approximately 0.5 seconds per recording.

In addition to cross-center generalization, we investigate the role of large-scale hybrid pre-training by comparing LPSGM models with and without pre-training through five-fold cross-validation (Fig.2d). The pre-trained model—initially developed on 220,500 hours of public, multi-center polysomnography (PSG) data—demonstrated superior performance following fine-tuning on target-center datasets. Quantitative analysis revealed consistent improvements across all evaluation metrics: pre-training enhanced accuracy, macro-F1 score, and Cohen's kappa by 2.04%, 1.91%, and 2.67%, respectively, for the SYSU dataset, while comparable improvements of 1.27%, 1.57%, and 1.62% were observed for the HANG7 dataset. These systematic enhancements not only validate the efficacy of the pre-trained LPSGM framework for domain adaptation but also demonstrate its practical advantages, including faster convergence and significantly better performance compared to models trained from scratch.

Comparison with State-of-the-Art Sleep Staging Methods

To our knowledge, LPSGM constitutes the largest PSG training framework to date among existing sleep staging models, a scale we hypothesize underlies its superior generalization performance. We systematically evaluated LPSGM against contemporary methods using standardized dataset sizes, categorizing comparison approaches as either non-transfer-based or transfer-based according to their original published methodologies. Non-transfer-based methods included: DeepSleepNet[24] (a classical CNN-BiLSTM network for extracting local features and learning transition rules), TinySleepNet[25] (a classical model based on CNN and RNN with fewer model

parameters), U-Time[33] (a fully-CNN encoder-decoder architecture for time series segmentation applied to sleep staging), and AttnSleep[34] (a hybrid architecture composed of a multi-resolution CNN and multi-head self-attention with causal convolutions). Transfer-based methods comprised SleepDG[32] (a combination of CNN and Transformer architectures that incorporates a proposed multi-level feature alignment technique to extract domain-invariant features) and RobustSleepNet[30] (channel-adaptive architecture trained on heterogeneous configurations). Notably, only RobustSleepNet and LPSGM supported flexible channel configurations. To ensure fair comparison, we standardized input channels to the overlapping C3-M2 and E1-M2 montages across all methods except RobustSleepNet and LPSGM, which were evaluated under both 2-channel (2C) and 8-channel (8C) configurations.

We implemented these methods based on their public code and default hyper-parameters under our cross-center protocol. LPSGM (8C) achieved state-of-the-art performance across all metrics (Table 2), outperforming both LPSGM (2C) and comparator methods. This demonstrates the value of multi-channel EEG integration through complementary signal information. Remarkably, LPSGM (2C) maintained superior performance over alternatives, highlighting the efficacy of Transformer-based channel encoding and attention mechanisms compared to conventional channel-stacking convolutional approaches. In non-transfer methods, we observed a model size-performance correlation, implying capacity limitations in large-scale training scenarios. Among transfer-based methods, SleepDG approached LPSGM (2C) performance, while RobustSleepNet's compact architecture (0.2M parameters) exhibited dataset-specific learning patterns, evidenced by pronounced HANG7-SYSU performance discrepancies, suggesting constrained 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. Model Size indicates the number of trainable parameters in each model, reported in millions (M).
b. 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.
c. 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

Beyond sleep stage classification, we extended LPSGM to two clinical diagnostic tasks: narcolepsy identification (HANG7 dataset, n=84) and MDD detection (SYSU dataset, n=44). Three fine-tuning strategies—partial (updating only the classifier while freezing other parameters), full (optimizing the entire model for diagnosis), and joint (simultaneously optimizing sleep staging and diagnosis)—were systematically compared against models trained from scratch. Diagnostic outcomes were determined via subject-level majority voting of sample-wise predictions, with five-fold cross-validation ensuring strict training-test separation. As shown in Fig. 3, fine-tuned LPSGM achieved 88.01% accuracy (90% sensitivity, 84.29% specificity) for narcolepsy detection and 100% accuracy for MDD identification. Joint fine-tuning yielded superior performance, with considerable improvement over training from scratch. These results underscore the clinical viability of unified PSG frameworks for multimodal diagnostics, bridging sleep physiology with neuropsychiatric disorder detection.

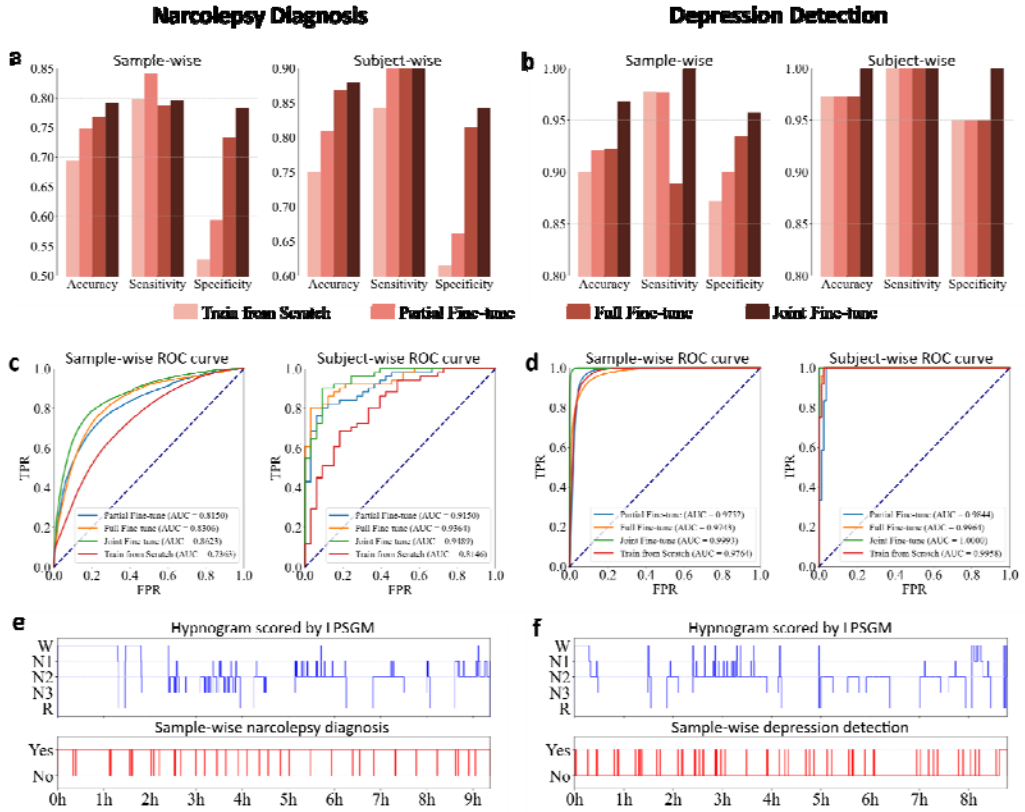

Fig. 3: Performance of fine-tuned LPSGM on narcolepsy diagnosis (left panels: a, c, e) and MDD detection (right panels: b, d, f). The model was fine-tuned using three different strategies and compared against models trained from scratch. Partial fine-tune, where only the classifier is fine-tuned for 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 Validation of LPSGM for Clinical Sleep Staging

We conducted a prospective validation study at the HANG7 center (protocol: Fig. 1c) to assess the clinical applicability of LPSGM. Three board-certified sleep experts independently scored twenty full-night PSG recordings, randomly selected over a one-month period. LPSGM predictions were generated for the same recordings, with comparative analyses evaluating: 1) Inter-expert variability: Each expert's annotations vs. consensus derived from the other two; 2) Model-expert agreement: LPSGM predictions vs. full expert consensus.

Classification metrics (Table 3) employed probabilistically adjusted accuracy and Cohen's κ (Supplementary Methods). Consensus labels were determined via epoch-wise majority voting, with tied probabilities evenly distributed. LPSGM achieved 89.2% accuracy ($\kappa = 0.845$), closely matching expert consensus (88.7% accuracy, $\kappa = 0.850$). Two-tailed *t*-tests revealed no significant differences between LPSGM and experts (accuracy: $p = 0.74$; κ: $p = 0.60$). Notably, LPSGM exhibited lower variability (SD: 4.6% accuracy, 6.6% κ) than individual experts (SD: 9.7% accuracy, 10.2% κ), highlighting superior consistency.

Hypnogram metrics (Table 4) quantified sleep architecture through stage transitions, categorized into: 1) Sleep latency: Time to enter specific stages. 2) Sleep duration: Total time per stage. 3) Stage distributions: Proportion relative to total sleep time. Mean absolute error (MAE) was calculated for each scorer against consensus contributors. For model-expert comparisons, MAE reflected LPSGM's deviation from individual experts, while inter-expert MAE captured pairwise differences. Statistical analysis identified significant discrepancies in N2 latency ($p < 0.05$) and N1 duration/distribution ($p < 0.01$), suggesting that while LPSGM robustly captures global sleep architecture, refining transient N1/N2 transition detection could further enhance clinical utility.

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. Two-tailed *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 (Supplementary Methods).
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 (SPT)	Episodes (#)	—	5.0±4.4	4.7±3.2	5.4±4.2	4.0±3.1	4.9±5.3
	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 of LPSGM

The black-box nature of deep learning models remains a critical barrier to clinical adoption, where transparency in decision-making is essential. To decode LPSGM's inference mechanisms, we applied gradient-weighted class activation mapping (Grad-CAM)[59], a post hoc interpretability method that generates spatial heatmaps highlighting input regions most salient to model predictions. Analyses focused on the final convolutional layer of the Epoch Encoder's dual-branch CNN (Fig. 5b), with activation maps resampled to EEG resolution and aggregated across branches to visualize attention patterns during sleep staging.

Figure 4 illustrates Grad-CAM results for six representative 30-second EEG epochs. Panels a–d showcase correct classifications, with attention peaks aligning with clinical biomarkers per AASM guidelines: (a) Posterior dominant rhythm (PDR) in occipital channels (>50% epoch coverage), pathognomonic of wakefulness. (b) Frontal slow-wave activity diagnostic of N3 sleep. (c) Low-amplitude mixed-frequency (LAMF) EEG with slow eye movements (SEM), indicative of N1 sleep. (d) Rapid eye movements (REMs) confirming REM sleep. In contrast, panels e–f depict misclassifications where activation patterns diverge from established biomarkers—either misallocating attention (e.g., transient artifacts) or lacking discriminative features for ambiguous epochs.

These visualizations provide two key insights into LPSGM's interpretability: (1) the model successfully attends to clinically relevant EEG features, confirming its biological plausibility, yet (2) reveals limitations in processing ambiguous or complex patterns, suggesting areas for future improvement.

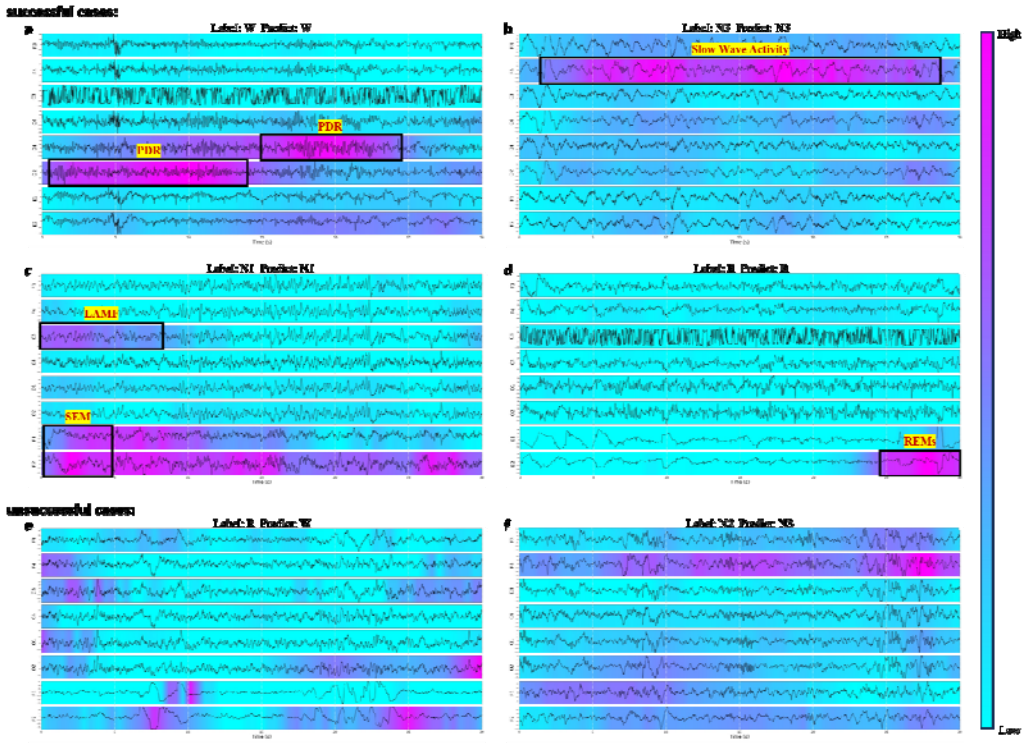

Fig. 4: Grad-CAM visualizations of LPSGM predictions for sleep staging. The figure presents six 30-second EEG epochs analyzed using Grad-CAM to highlight model-attended regions across eight EEG channels (F3, F4, C3, C4, O1, O2, E1, E2). Panels (a)-(d) depict correctly classified epochs, while panels (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

Ablation Study

To validate key architectural decisions in LPSGM, we performed systematic ablation analyses focusing on two core components: channel & temporal encoding approaches, and feature fusion strategies. These innovations, described in detail in the Methods section, are designed to enhance the model's ability to process heterogeneous PSG datasets. We evaluated three encoding variants (no encoding, addition-based encoding, and concatenation-based encoding) alongside two fusion approaches (channel averaging versus CLS token-based fusion). As shown in Table 5, concatenation-based encoding paired with CLS token fusion achieved optimal performance. The baseline (no encoding) exhibited the weakest results, while

addition-based encoding showed partial improvement. This hierarchy underscores the necessity of preserving spatiotemporal information in Transformer architectures. We posit that addition-based encoding underperforms due to dimensional conflation during feature integration, which degrades discriminative signal retention. In contrast, concatenation prevents dimensional overlap, enabling robust representation of heterogeneous inputs. CLS token fusion outperformed channel averaging by replacing static linear aggregation with dynamic, attention-driven weighting. While naive averaging indiscriminately combines features, CLS tokens leverage self-attention to emphasize diagnostically salient patterns, thereby preserving feature distinctiveness while adaptively prioritizing clinically relevant biomarkers.

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

To determine the optimal architecture for clinical deployment, we evaluated LPSGM's performance across varying numbers of Transformer blocks (N = 1–6). As shown in Table 6, increasing N yielded divergent outcomes: the HANG7 cohort exhibited consistent performance gains (accuracy: +2.32% from N = 1 to N = 6), while SYSU showed non-linear improvement, peaking at N = 5 (+3.90% vs. N = 1). This dataset-specific response suggests differing sensitivities to model complexity, likely arising from variations in data heterogeneity between clinical centers.

Balancing computational efficiency and performance, we selected N = 4 as the optimal configuration. This choice retained >99.5% of maximum achievable accuracy while reducing memory demands by 30.8% (9.9M parameters vs. 14.3M for N = 6). The diminishing returns beyond N = 5 highlight the necessity of architectural parsimony for real-world clinical applications, where resource constraints and deployment scalability are paramount.

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 represents a transformative advance in automated sleep staging and mental disorder diagnosis through PSG. Trained on 220,500 hours of PSG data (24,000 full-night recordings aggregated across 16 public datasets)[6,35–52], LPSGM achieves robust cross-center generalization, matching the performance of models trained exclusively on target-center data. This capability stems from its ability to approximate real-world clinical diversity: multi-center training aggregates heterogeneous sampling distributions arising from variations in population demographics, recording protocols, and annotation criteria. While individual datasets offer limited snapshots of clinical reality, their integration yields a comprehensive approximation of true data distributions, enabling near-optimal performance (99.6% and 97.1% of center-specific accuracy on independent HANG7 and SYSU datasets, respectively). These results highlight coordinated multi-institutional data sharing as a critical pathway to enhance robustness and accelerate clinical adoption of automated sleep staging technologies.

LPSGM's dynamic channel adaptability addresses the inherent heterogeneity of PSG protocols across clinical environments, enabling scalable sleep assessment tailored to diverse diagnostic needs. In specialized sleep laboratories, the model maximizes precision by leveraging multi-channel EEG configurations to detect subtle neurological or psychiatric abnormalities. Conversely, in resource-constrained settings—such as primary care clinics or home-based monitoring—LPSGM maintains reliability even with reduced channel setups, facilitating accessible preliminary assessments and longitudinal compliance. This flexibility bridges the gap between high-resolution diagnostics and real-world practicality while establishing a foundation for future integration of multimodal physiological signals (e.g., ECG, respiratory metrics), which could refine diagnostic accuracy and enhance pathophysiological insights into sleep disorders.

Interpretability remains foundational to deploying deep learning models in clinical workflows[60–62]. Grad-CAM[59] analyses reveal that LPSGM prioritizes EEG features aligned with clinical biomarkers during sleep staging—for example, posterior dominant rhythms (PDR) for wakefulness and slow-wave activity for N3 sleep. This congruence with expert-defined criteria (AASM guidelines[5]) fosters trust in LPSGM's decision-making, as visualized in attention maps from the final convolutional layer of the Epoch Encoder (Fig. 5b). However, misclassifications correlate with divergent attention patterns, such as transient artifact focus or ambiguous EEG signatures (e.g., N1/N2 transitions). These cases highlight opportunities to refine temporal context modeling and artifact resilience. Future iterations could integrate clinician feedback loops during training and embed real-time saliency overlays into user interfaces, enabling practitioners to validate model attention against clinical intuition[63,64]. Such transparency enhancements are critical for adoption in routine workflows, bridging

the gap between algorithmic outputs and actionable clinical insights.

Sleep and mental health are bidirectionally linked[65–68] through neurochemical and circadian pathways—disrupted sleep exacerbates psychiatric symptoms (e.g., depression, anxiety) by impairing emotional regulation, while mental disorders often dysregulate sleep architecture, perpetuating a harmful cycle. Notably, our LPSGM demonstrates promising potential for identifying neuropsychiatric disorders through PSG, achieving 88.01% accuracy in narcolepsy detection and 100% accuracy in MDD identification when fine-tuned for diagnostic tasks. These results suggest that sleep-related physiological features may encode biomarkers for psychiatric conditions, with joint optimization of sleep staging and diagnosis implying shared latent representations between sleep architecture and symptom profiles. However, these findings require cautious interpretation. Limited sample sizes (n = 84 for narcolepsy; n = 44 for MDD) and absence of direct comparisons to gold-standard clinical diagnostics necessitate validation in larger, diverse cohorts. Although the perfect MDD classification is provocative, replication across populations and clinical workflows is critical to confirm generalizability.

While LPSGM demonstrates promising performance, several limitations merit consideration. First, hypnogram analysis revealed opportunities for refinement: transient N1/N2 transitions—critical for assessing sleep fragmentation—were less reliably detected by LPSGM compared to expert consensus. Future iterations could prioritize temporal context modeling (e.g., attention to stage progression patterns) to improve resolution of these clinically salient dynamics. Second, while LPSGM achieves strong performance with multi-channel configurations, accuracy drops in single-channel setups—a critical barrier for home or resource-limited applications. Future work should prioritize optimizing feature extraction algorithms for minimal-channel configurations to enhance accessibility. Furthermore, to advance clinical utility, subsequent studies could extend LPSGM's scope to other disorders such as sleep apnea and anxiety disorders, while integrating multimodal signals (e.g., chin electromyography, heart rate variability) to disentangle comorbidities and refine diagnostic specificity. Finally, adaptation for wearable devices could facilitate broader access to PSG-based monitoring, enabling scalable, real-world at-home sleep assessments that bridge diagnostic gaps in underserved populations.

In conclusion, this work presents LPSGM as a unified, scalable framework for automated PSG analysis, combining robust cross-domain sleep staging with a novel extension to neuropsychiatric disorder diagnosis. By harmonizing large-scale multi-center PSG data and enabling adaptive channel configurations, the model bridges critical gaps between technical innovation and clinical applicability. To facilitate further research and clinical translation, we release both the source code and pre-trained models, offering the scientific community a foundation for advancing generalizable, clinically adaptable PSG technologies.

Methods

Overview

LPSGM operates as a sequence-to-sequence model for automatic sleep staging using multichannel PSG signals (Fig. 5a). The architecture comprises three core components: an Epoch Encoder that extracts local intra-epoch features, a Sequence Encoder that models global inter-epoch dependencies via self-attention mechanisms, and a Classifier that assigns sleep stages to sequential 30-second PSG epochs.

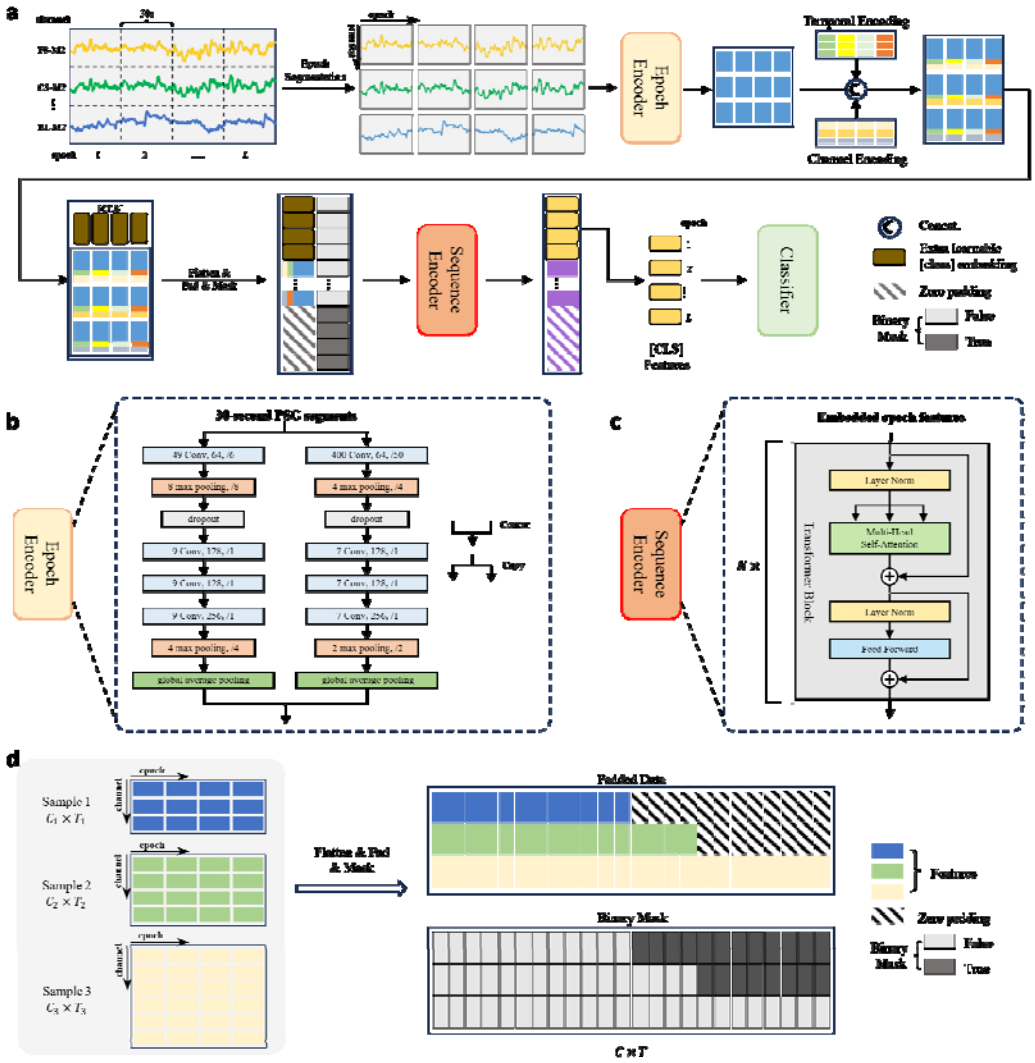

Fig. 5: Overall architecture of LPSGM. (a) LPSGM consists of an Epoch Encoder, Sequence Encoder, and Classifier, designed for both sleep staging and disorder diagnosis. (b) The Epoch Encoder employs a dual-branch CNN to extract local intra-epoch features from each 30-second PSG segment, using small and 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 address heterogeneous channel configurations, PSG signals are segmented into 30-second epochs. For a multi-channel PSG sequence $x_i = \{x_i^1, x_i^2, \dots, x_i^L\}$ containing L epochs, each epoch $x_i^l \in \mathbb{R}^{C_i \times n}$ comprises C_i channels and $n = 30 \times f_s$ samples. Segmentation is performed across temporal and channel dimensions, yielding $L \times C_i$ single-channel segments $x_i^{l,c} \in R^n$. This standardization enables consistent input formatting for downstream encoding and classification.

Epoch Encoder

Each single-channel segment $x_i^{l,c} \in \mathbb{R}^n$ is processed by a dual-branch convolutional neural network (CNN) inspired by DeepSleepNet[24]. The two branches employ small and large filters to capture high- and low-frequency features, respectively. Each branch consists of four convolutional layers (performing 1D convolution, batch normalization and ReLU activation sequentially), two max pooling layers for downsampling, and a global average pooling layer. Outputs from both branches are concatenated along the feature dimension, 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, Transformer architectures lack inherent positional encoding mechanisms for capturing temporal or channel-specific information. To address this, we introduce channel and temporal embeddings inspired by LaBraM[58], but replace additive fusion with concatenation to preserve dimensional distinctiveness (Fig. 5c).

Specifically, we maintain two embedding sets: a channel embedding set $CE = \{ce_1, ce_2, \dots, ce_{|C|}\}$ and a temporal embedding set $TE = \{te_1, te_2, \dots, te_L\}$. Let d_{ce} and d_{te} denote the dimension of the embedding vectors in the CE and TE, respectively. For each feature vector $e_i^{l,c}$ from epoch encoder, we concatenate its 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, where input sequences maintain a fixed temporal length (L epochs) but exhibit variability in channel counts (C_i), we address hardware limitations in batch processing through a padding and masking strategy adapted from natural language processing (NLP). Following segmentation and encoding, feature sequences of variable length $L \times C_i$ are standardized by unfolding each batch of B sequences into $C_i \times T_i$ matrices ($T_i = L$, fixed to 10-minute epochs). These matrices are padded with zero vectors to a uniform dimension of $C_{max} \times T_{max}$, where C_{max} and T_{max} are the maximum number of channels and temporal length within the batch.

A binary mask $M \in \{0,1\}^{B \times (C_{max} \times T_{max})}$ is generated to distinguish valid data from padding:

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

where $M_{ij} = 1$ indicates that the j-th element of the i-th sequence is padding while $M_{ij} = 0$ indicates valid data. The mask serves dual roles: suppressing padded positions during Transformer self-attention computations to prevent spurious feature interactions, and isolating loss calculations exclusively to valid data points during backpropagation. By enforcing $T_i = L$, we ensure temporal consistency across sequences while enabling efficient batch training.

Insertion of CLS token

To generate unified representations for variable-length sequences, we introduce the classification (CLS) token, a technique commonly used in Transformer-based architectures. Specifically, L learnable CLS tokens are prepended to each feature sequence $\widetilde{e_i}$. These tokens share the dimensionality of the encoded features ($d + d_{ce} + d_{te}$), thereby extending the sequence to $\widetilde{e_i}'$ with a length of $L \times (C_i + 1)$. The CLS tokens are initialized randomly and optimized during training to capture global sequence characteristics.

Sequence Encoder

The sequence encoder processes the padded and masked feature sequences (including CLS tokens) to extract global temporal dependencies from multi-channel, multi-epoch sleep data. Built on the Transformer architecture, it comprises N identical Transformer blocks, each containing a multi-head self-attention (MSA) layer

and a feed-forward network (FFN), interleaved with layer normalization (LN).

Let E_0 denote the input feature sequence and E_{out} the output sequence. The encoding process for each Transformer block (indexed $\ell = 1, \ldots, N$) is defined as:

$$E_\ell^{'} = \mathrm{MSA}\big(\mathrm{LN}(E_{\ell-1})\big) + E_{\ell-1}, \quad \ell = 1, \ldots, N \#(3)$$

$$E_\ell = \mathrm{FFN}\left(\mathrm{LN}\left(E_\ell^{'}\right)\right) + E_\ell^{'}, \quad \ell = 1, \ldots, N \#(4)$$

The final output is computed as:

$$E_{out} = \mathrm{LN}(E_N) \#(5)$$

The sequence retains its original dimensions through this process. To generate classification features E_{cls}, the first L elements of E_{out} (corresponding to CLS tokens) are extracted:

$$E_{cls} = E_{out}[0{:}\,L] \#(6)$$

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 conducted 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:

$$\mathcal{L}_{classify} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_k \cdot y_i^k \log \hat{y}_i^k \#(7)$$

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 35, 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 mental 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[35,36], STAGES[35], ABC[35,37], HOMEPAP[35,52], SHHS[35,38], PATS[35,39,40], CHAT[35,41], CCSHS[35,42], CFS[35,43], MNC[6,35], NCHSDB[35,44] and MROS[35,45] 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[46], HMC[47], P2018[48] and CAP[49] 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[50] dataset can be accessed from https://sleeptight.isr.uc.pt/. DOD-H and DOD-O datasets[51] can be downloaded at https://github.com/Dreem-Organization/dreem-learning-open. Access to the HANG7 and SYSU dataset[69] is governed by data-use agreements, and it is therefore not

publicly available.

Code Availability

The source code and pre-trained models 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. R. Why sleep is important for health: a psychoneuroimmunology perspective. *Annu. Rev. Psychol.* **66**, 143–172 (2015).
3. Benjafield, A. V. *et al.* Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir. Med.* **7**, 687–698 (2019).
4. Morin, C. M. & Jarrin, D. C. Epidemiology of insomnia: prevalence, course, risk factors, and public health burden. *Sleep Med. Clin.* **17**, 173–191 (2022).
5. American Academy of Sleep Medicine. *AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* 3rd edn (American Academy of Sleep Medicine, Darien, IL, 2023).
6. Stephansen, J. B. *et al.* Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **9**, 5229 (2018).
7. Quinnell, T. G. & Smith, I. E. Narcolepsy, idiopathic hypersomnolence and related conditions. *Clin. Med.* **11**, 282 (2011).
8. Lopez, J. *et al.* Reduced sleep spindle activity in early-onset and elevated risk for depression. *J. Am. Acad. Child Adolesc. Psychiatry* **49**, 934–943 (2010).
9. Riemann, D., Krone, L. B., Wulff, K. & Nissen, C. Sleep, insomnia, and depression. *Neuropsychopharmacol.* **45**, 74–89 (2020).
10. Kupfer, D. J. REM latency: a psychobiologic marker for primary depressive disease. *Biol. Psychiatry* **11**, 159–174 (1976).
11. Kupfer, D. & Foster, F. G. Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression. *Lancet* **300**, 684–686 (1972).
12. Rosenberg, R. S. & Van Hout, S. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* **09**, 81–87 (2013).
13. Zhang, X. *et al.* Process and outcome for international reliability in sleep scoring. *Sleep Breath* **19**, 191–195 (2015).
14. Danker-Hopfe, H. *et al.* Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J. Sleep Res.* **18**, 74–84 (2009).
15. MacLean, A. W., Lue, F. & Moldofsky, H. The reliability of visual scoring of alpha EEG activity during sleep. *Sleep* **18**, 565–569 (1995).
16. Kim, Y. D. *et al.* 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 Clin. Neurosci.* **47**, 91–97 (1993).
17. Hassan, A. R. & Bhuiyan, M. I. H. Automatic sleep stage classification. *In 2015 2nd Int. Conf. on Electr. Inf. Commun. Tech. (EICT)* **211–216** (IEEE, 2015).

18. Rahman, M. A. *et al.* Optimization of sleep stage classification using single-channel EEG signals. *In 2019 4th Int. Conf. on Electr. Inf. Commun. Tech. (EICT)* **1–6** (IEEE, 2019).
19. Hassan, A. R. & Bhuiyan, M. I. H. A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. *J. Neurosci. Methods* **271**, 107–118 (2016).
20. 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. *Comput. Meth. Programs Biomed.* **108**, 10–19 (2012).
21. Şen, B., Peker, M., Çavuşoğlu, A. & Çelebi, F. V. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. *J. Med. Syst.* **38**, 1–21 (2014).
22. Zhao, S. *et al.* Evaluation of a single-channel EEG-based sleep staging algorithm. *Int. J. Environ. Res. Public Health* **19**, 2845 (2022).
23. Touil, M., Bahatti, L. & El Magri, A. Sleep's depth detection using electroencephalogram signal processing and neural network classification. *J. Med. Artif. Intell.* **5**, 9 (2022).
24. Supratak, A., Dong, H., Wu, C. & Guo, Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 1998–2008 (2017).
25. Supratak, A. & Guo, Y. TinySleepNet: an efficient deep learning model for sleep stage scoring based on raw single-channel EEG. *In 2020 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)* **641–644** (IEEE, 2020).
26. Biswal, S. *et al.* Expert-level sleep scoring with deep neural networks. *J. Am. Med. Inform. Assoc.* **25**, 1643–1650 (2018).
27. Phan, H. *et al.* XSleepNet: multi-view sequential model for automatic sleep staging. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 5903–5915 (2022).
28. Yoo, C., Lee, H. W. & Kang, J.-W. Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network. *IEEE J. Biomed. Health Inform.* **26**, 1273–1284 (2022).
29. Fan, J. *et al.* Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 205–216 (2022).
30. Guillot, A. & Thorey, V. RobustSleepNet: transfer learning for automated sleep staging at scale. *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 1441–1451 (2021).
31. Deng, Z. *et al.* Adversarial training helps transfer learning via better representations. *Adv. Neural Inf. Process. Syst.* **34**, 25179–25191 (2021).
32. Wang, J. *et al.* Generalizable sleep staging via multi-level domain alignment. *Proc. AAAI Conf. Artif. Intell.* **38**, 265–273 (2024).
33. Perslev, M. *et al.* U-Time: a fully convolutional network for time series segmentation applied to sleep staging. *Adv. Neural Inf. Process. Syst.* **32**, – (2019).
34. Eldele, E. *et al.* An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 809–818 (2021).
35. Zhang, G.-Q. *et al.* The National Sleep Research Resource: towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **25**, 1351–1358 (2018).
36. Quan, S. F. *et al.* The association between obstructive sleep apnea and neurocognitive performance—The Apnea Positive Pressure Long-term Efficacy Study (APPLES). *Sleep* **34**,

303–314 (2011).

37. Bakker, J. P. *et al.* Gastric banding surgery versus continuous positive airway pressure for obstructive sleep apnea: a randomized controlled trial. *Am. J. Respir. Crit. Care Med.* **197**, 1080–1083 (2018).

38. Quan, S. F. *et al.* The sleep heart health study: design, rationale, and methods. *Sleep* **20**, 1077–1085 (1997).

39. Wang, R. *et al.* 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).

40. Redline, S. *et al.* Adenotonsillectomy for snoring and mild sleep apnea in children: a randomized clinical trial. *JAMA* **330**, 2084–2095 (2023).

41. Marcus, C. L. *et al.* A randomized trial of adenotonsillectomy for childhood sleep apnea. *N. Engl. J. Med.* **368**, 2366–2376 (2013).

42. Rosen, C. L. *et al.* Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. *J. Pediatr.* **142**, 383–389 (2003).

43. Redline, S. *et al.* The familial aggregation of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.* **151**, 682–687 (1995)..

44. Lee, H. *et al.* A large collection of real-world pediatric sleep studies. *Sci. Data* **9**, 421 (2022).

45. Blackwell, T. *et al.* 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.* **59**, 2217–2225 (2011).

46. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000).

47. Alvarez-Estevez, D. & Rijsman, R. M. Haaglanden Medisch Centrum sleep staging database (version 1.0.1). *PhysioNet* (2021).

48. Ghassemi, M. *et al.* You snooze, you win: the PhysioNet/Computing in Cardiology Challenge 2018. *In 2018 Comput. Cardiol. Conf. (CinC)* **45**, 1–4 (2018).

49. Terzano, M. G. *et al.* Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep Med.* **2**, 537–554 (2001).

50. Khalighi, S. *et al.* ISRUC-Sleep: a comprehensive public dataset for sleep researchers. *Comput. Meth. Programs Biomed.* **124**, 180–192 (2016).

51. Guillot, A. *et al.* Dreem open datasets: multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**, 1955–1965 (2020).

52. Rosen, C. L. *et al.* 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).

53. Achiam, J. *et al.* GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).

54. Lewis, M. *et al.* BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *In Proc. 58th Annu. Meet. Assoc. Comput. Linguist.* **7871–7880** (ACL, 2020).

55. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).

56. 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. In *Proceedings of the 13th International Conference on Learning Representations* (2025).
57. Yi, K. *et al.* Learning topology-agnostic EEG representations with geometry-aware modeling. *Adv. Neural Inf. Process. Syst.* **36** (2024).
58. Jiang, W.-B., Zhao, L.-M. & Lu, B.-L. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations* (2024).
59. Selvaraju, R. R. *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. *In 2017 IEEE Int. Conf. Computer Vision (ICCV)* **618–626** (IEEE, 2017).
60. Esser-Skala, W. & Fortelny, N. Reliable interpretability of biology-inspired deep neural networks. *NPJ Syst. Biol. Appl.* **9**, 50 (2023).
61. Sapoval, N. et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **13**, 1728 (2022).
62. Mamede, S. & Schmidt, H. G. Making large language models into reliable physician assistants. *Nat. Med.* **31**, 1071–1072 (2025).
63. Nasarian, E., Alizadehsani, R., Acharya, U. R. & Tsui, K.-L. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician–AI–collaboration framework. *Info. Fusion* **108**, 102412 (2024).
64. Phan, H. *et al.* SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans. Biomed. Eng.* **69**, 2456–2467 (2022).
65. Alvaro, P. K., Roberts, R. M. & Harris, J. K. A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. *Sleep* **36**, 1059–1068 (2013)
66. Scott, A. J., Webb, T. L., Martyn-St James, M., Rowse, G. & Weich, S. Improving sleep quality leads to better mental health: a meta-analysis of randomised controlled trials. *Sleep Med. Rev.* **60**, 101556 (2021).
67. Tahmasian, M. *et al.* The interrelation of sleep and mental and physical health is anchored in grey-matter neuroanatomy and under genetic control. *Commun. Biol.* **3**, 171 (2020).
68. Meyer, N. *et al.* The sleep-circadian interface: a window into mental disorders. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2214756121 (2024).
69. Wei, Y., Zhu, Y., Zhou, Y., Yu, X. & Luo, Y. Automatic sleep staging based on contextual scalograms and attention convolution neural network using single-channel EEG. *IEEE J. Biomed. Health Inform.* **28**, 801–811 (2024).

Acknowledgements

This work was supported by STI2030-Major Projects (2022ZD0212400), National Natural Science Foundation of China (82371453), Key R&D Program of Zhejiang (2024C03006，2024C04024, 2024SSYS0017), Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2023-PT310-01) and Fundamental

Research Funds for the Central Universities（2025ZFJH01-01）

Author Contributions

G.D. and H.J. designed the methodology. J.Z. and Y.Z provided critical insights into the methodological design. G.D. conducted the experiments and performed the primary analysis. G.D. and M.N. carried out the interpretability analysis. S.R., J.X., S.Z., and G.P. contributed to the analysis and interpretation of results. M.N., 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