A Unified Flexible Large Polysomnography Model for Sleep Staging and Brain 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 brain 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 for sleep disorder diagnosis, LPSGM achieved 80.47% accuracy on the large-scale MNC dataset (773 subjects) for a three-class classification (Healthy Control vs. T1 Narcolepsy vs. Other Hypersomnia). The model also demonstrated strong cross-institutional generalizability, with an AUC of 0.8791 on independent cohorts for a binary (Normal vs. Abnormal) classification. While depression screening on smaller datasets showed perfect accuracy in controlled settings, larger-scale validation is necessary. By bridging automated sleep staging with real-world clinical deployment, LPSGM establishes a scalable framework for integrated sleep and brain disorder 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; Sleep disorder diagnosis; Depression screening; 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 brain disorder 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 brain 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 Methods 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 brain disorder diagnosis, LPSGM demonstrates robust cross-dataset generalizability, achieving strong performance in sleep disorder classification on a large-scale multi-center dataset (773 subjects) and maintaining diagnostic reliability when transferred to independent clinical cohorts without additional training. By unifying robust cross-domain generalization with clinical adaptability, LPSGM bridges the gap between automated sleep staging and real-world deployment, enabling scalable analysis of sleep physiology and its neuropsychiatric correlates.

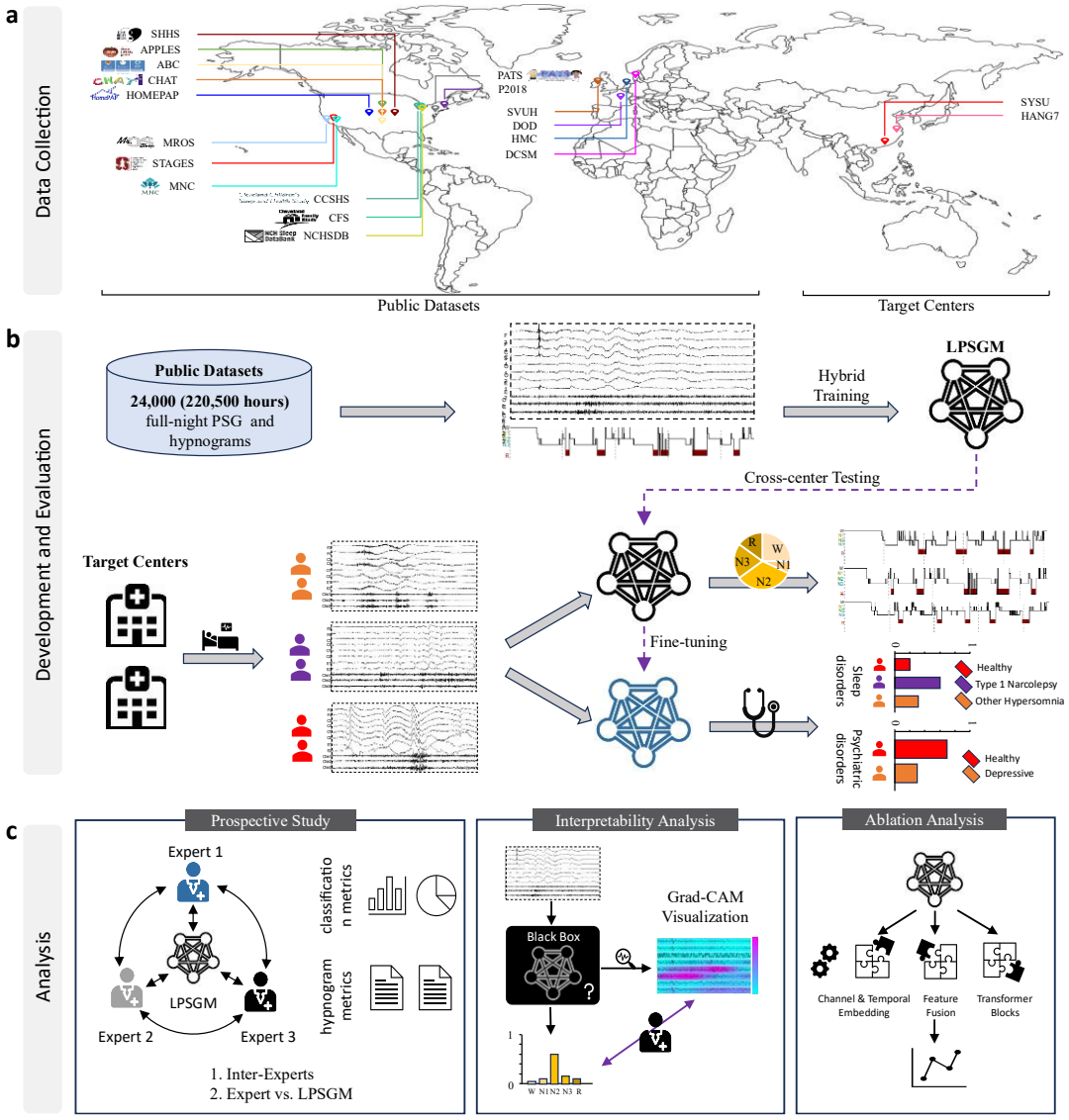


Fig. 1: Overview of the LPSGM framework. Panel (a): Data harmonization schematizes the aggregation of 220,500 hours of PSG data from 16 public datasets and 2 independent clinical cohorts, spanning diverse geographic populations and recording protocols. Panel (b): Cross-center generalization outlines 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 sleep disorder diagnosis and MDD screening. 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.

Additional validation on three public datasets—the MESA[71] and two subsets from the MASS (MASS-SS1 and MASS-SS3)[70] —further confirmed LPSGM's robust generalization capabilities, achieving accuracy of 83.74%, 79.98%, and 84.97% respectively, demonstrating consistent performance across diverse populations, age groups, and recording protocols (Supplementary Fig. 2).

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.

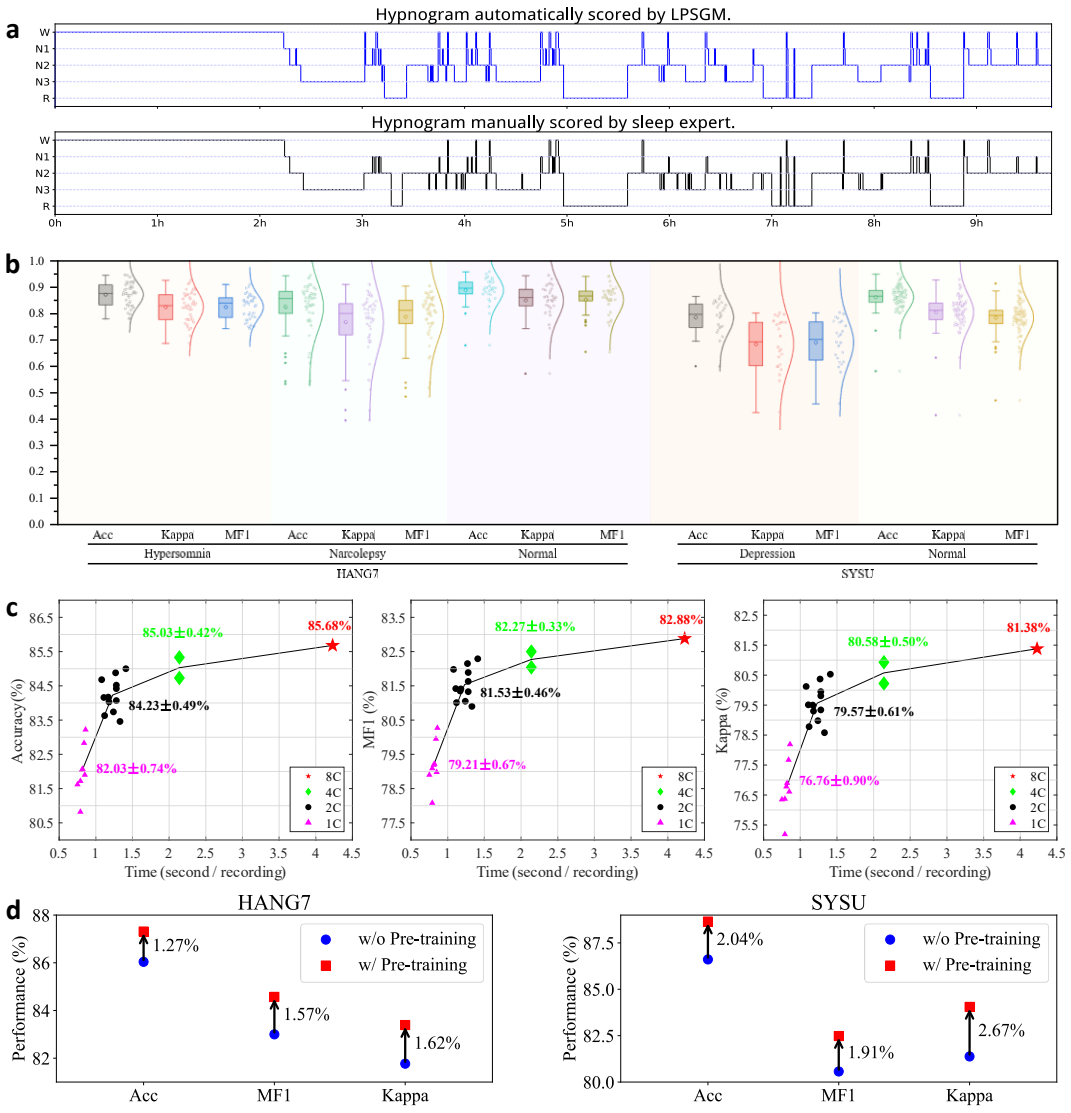


Fig. 2: Performance of LPSGM on Sleep Staging. (a) Hypnogram comparison of a full-night PSG recording from the HANG7 dataset. The top panel illustrates sleep stages predicted by LPSGM, while the bottom panel shows the expert-annotated reference hypnogram. (b) Boxplots summarizing distributions of accuracy (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. See Supplementary Table 5 for comprehensive performance metrics across all tested channel configurations.

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 Brain Disorder Diagnosis

To assess LPSGM's diagnostic capabilities beyond sleep staging, we conducted fine-tuning experiments for brain disorder classification across two independent clinical datasets with distinct patient populations and recording protocols. The MNC dataset comprised 773 PSG recordings from six international cohorts spanning three diagnostic categories: non-narcolepsy controls (n=310), Type 1 narcolepsy patients (n=254), and patients with other hypersomnia conditions (n=209) (Supplementary Methods, Supplementary Table 3). The HANG7 dataset included 127 subjects from a single clinical center: 33 healthy controls, 51 narcolepsy patients (13 Type 1, 38 Type 2), and 43 patients with hypersomnia symptoms associated with anxiety and depression (Supplementary Methods). We evaluated three fine-tuning strategies—Full Fine-tune, Partial Fine-tune, and Joint Fine-tune—against Train from Scratch baselines across multiple classification tasks, with particular emphasis on cross-dataset generalization performance.

On the MNC dataset, three-class classification among the diagnostic categories demonstrated clear advantages for fine-tuning approaches. At the subject level, Full Fine-tune achieved 80.47% accuracy ($\kappa = 0.7043$), outperforming Partial Fine-tune at 77.75% accuracy ($\kappa = 0.6643$) and Train from Scratch at 72.70% accuracy ($\kappa = 0.5880$) (Fig. 3a). Binary classification tasks provided additional validation of LPSGM's discriminative capacity. For Normal (Non-narcolepsy Healthy Control) versus Abnormal (T1 Narcolepsy and Other Hypersomnia) discrimination, Full Fine-tune achieved subject-wise AUC of 0.9663, compared to Partial Fine-tune (0.9442) and Train from Scratch (0.9394) (Fig. 3b). Type 1 narcolepsy identification against Non-T1 Narcolepsy Control demonstrated Full Fine-tune reaching subject-wise AUC of 0.9115, with Partial Fine-tune achieving 0.9065 and Train from Scratch attaining 0.8403 (Fig. 3c).

Cross-dataset evaluation provided direct assessment of generalization capabilities critical for clinical deployment. Models fine-tuned on MNC for Normal versus Abnormal classification were directly applied to HANG7 without additional training, achieving subject-wise AUC of 0.8791 (Fig. 3d). This cross-institutional performance validates LPSGM's capacity to maintain diagnostic accuracy across different patient populations, recording equipment, and clinical protocols. Within-dataset evaluation on HANG7 for Normal (Healthy Control) versus Abnormal (Narcolepsy and Hypersomnia) classification confirmed superior performance across fine-tuning approaches: Joint Fine-tune achieved subject-wise AUC of 0.9146, Full Fine-tune reached 0.9055, and Partial Fine-tune attained 0.8607, all substantially exceeding Train from Scratch performance at 0.8253 (Fig. 3d). The availability of complete sleep stage annotations in HANG7 enabled Joint Fine-tune evaluation, which maintained simultaneous sleep

staging and disease classification capabilities. Subject-wise evaluation consistently yielded superior performance compared to sample-wise metrics across all tasks, reflecting the clinical practice of diagnostic assessment based on comprehensive overnight recordings rather than individual epoch classifications (Fig. 3e).

In addition to the sleep disorder diagnostic tasks, we evaluated LPSGM's capacity for depression screening on two additional datasets. On the SYSU dataset comprising healthy controls (n=20) and patients with major depressive disorder (n=24), LPSGM demonstrated exceptional performance with Joint Fine-tune achieving perfect subject-wise accuracy (100%) and AUC of 1. However, performance declined on the larger APPLES dataset (928 healthy, 163 depressed subjects), where the classification target was depression history. With Full Fine-tune, subject-wise accuracy was 72.0% and the AUC was 0.7236. Detailed results for depression screening tasks are provided in Supplementary Fig. 8.

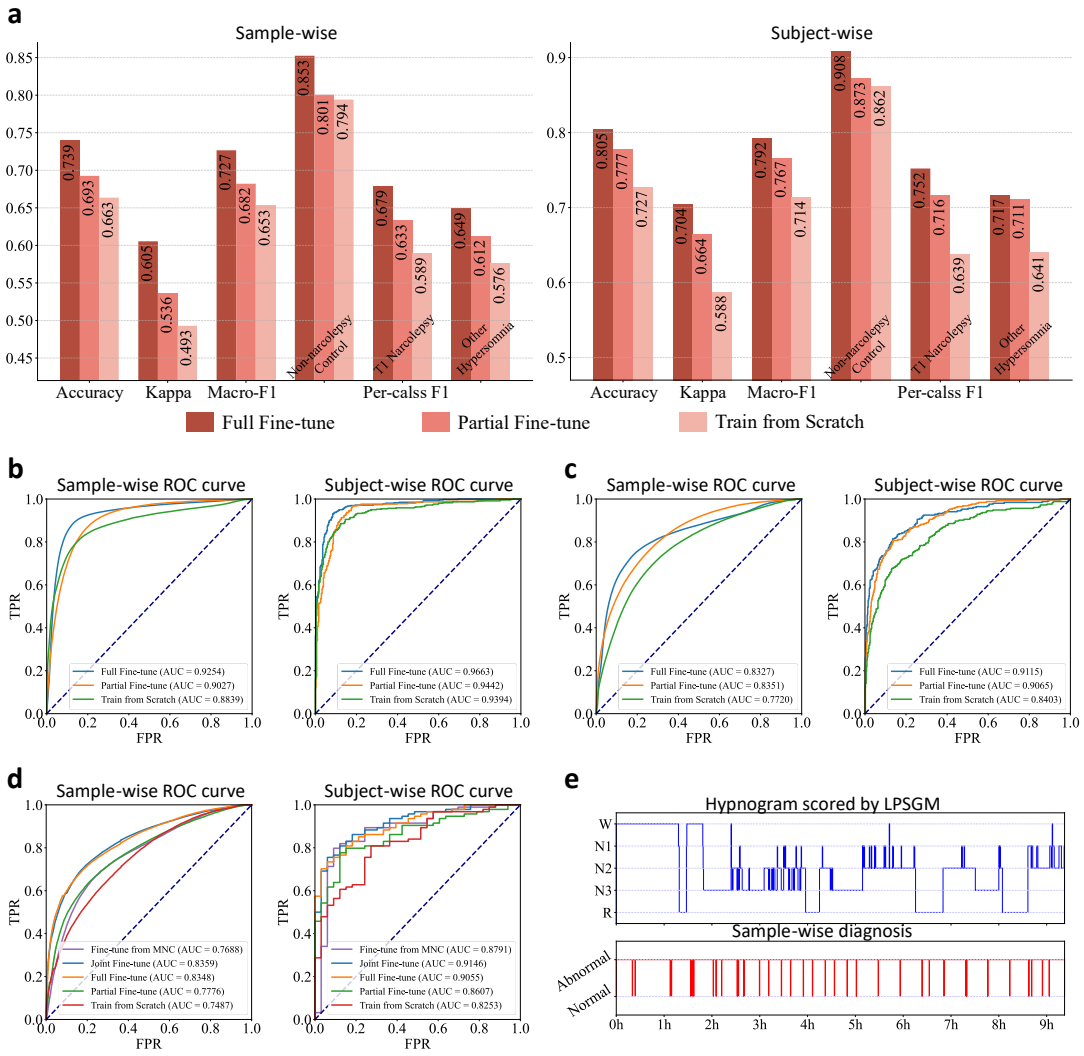


Fig. 3: Performance evaluation of LPSGM after fine-tuning for sleep disorder diagnosis on MNC and HANG7 datasets. Three fine-tuning strategies were compared: Full Fine-tune (all parameters updated), Partial Fine-tune (only disease classifier updated), and Joint Fine-tune (simultaneous optimization for sleep staging and disease diagnosis),

alongside Train from Scratch baseline. All evaluations used five-fold cross-validation unless specified otherwise. (a) Three-class classification on MNC dataset distinguishing Non-narcolepsy Healthy Controls, T1 Narcolepsy, and Other Hypersomnia, with performance metrics shown for both sample-wise (left) and subject-wise (right) evaluations. (b) Binary classification ROC analysis on MNC dataset for Normal (Non-narcolepsy) versus Abnormal (T1 Narcolepsy and Other Hypersomnia) discrimination, displaying sample-wise (left) and subject-wise (right) receiver operating characteristic curves with corresponding area under curve (AUC) values. (c) Binary classification ROC analysis on MNC dataset for T1 Narcolepsy identification, comparing Non-T1 Narcolepsy Control (Non-narcolepsy and Other Hypersomnia) versus T1 Narcolepsy, with sample-wise (left) and subject-wise (right) performance. (d) Binary classification ROC analysis on HANG7 dataset for Normal (Healthy Control) versus Abnormal (Narcolepsy and Hypersomnia) discrimination, including both within-dataset five-fold cross-validation and cross-dataset transfer evaluation (Fine-tune from MNC curve), demonstrating model generalizability across independent clinical cohorts. (e) Representative example of LPSGM diagnostic workflow showing overnight hypnogram with sleep stages automatically scored by LPSGM (top), and corresponding epoch-wise diagnostic predictions alternating between Normal and Abnormal classifications (bottom). Subject-level diagnosis is determined through majority voting of epoch-wise predictions across the entire recording. Note that joint fine-tuning was not applicable to MNC dataset due to incomplete sleep stage annotations. See Supplementary Fig. 3-7 for detailed confusion matrices for all classification tasks and fine-tuning approaches.

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	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 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.

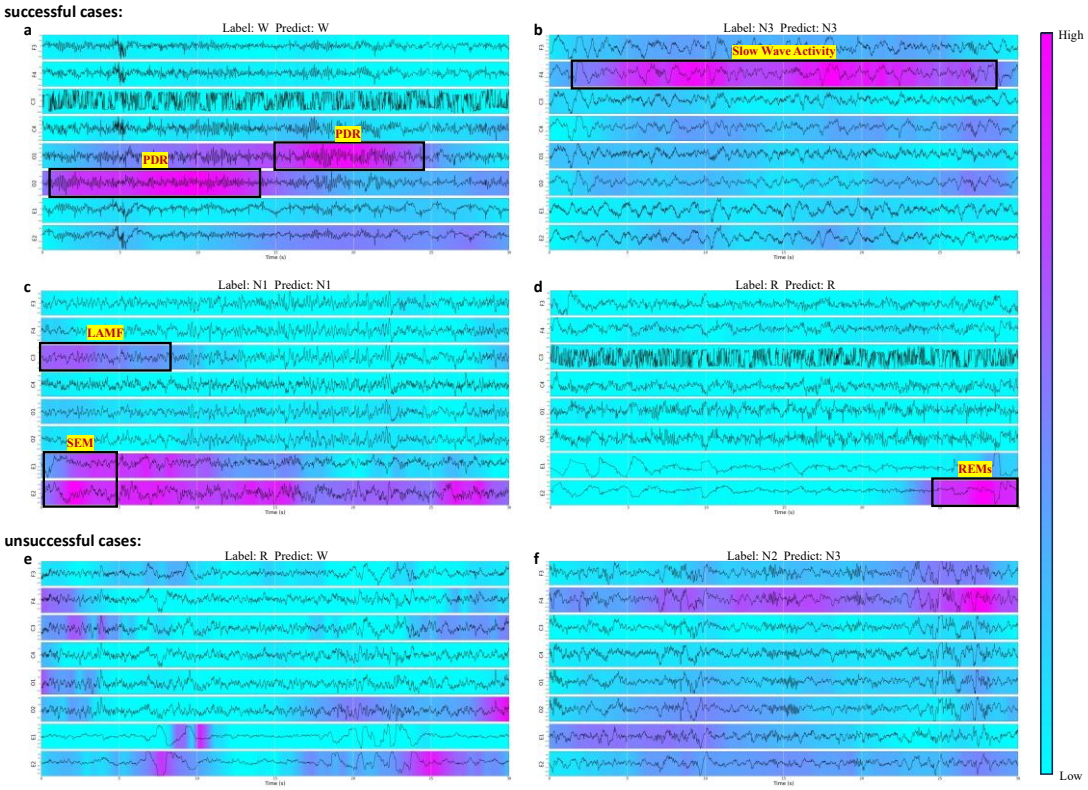


Fig. 4: Grad-CAM visualizations of LPSGM predictions for sleep staging. The figure presents six 30-second EEG epochs analyzed using Grad-CAM to highlight model-attended regions across eight EEG channels (F3, F4, C3, C4, O1, O2, E1, E2). Panels (a)-(d) depict correctly classified epochs, while panels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

mental disorders often dysregulate sleep architecture. Our evaluation of LPSGM for brain disorder diagnosis yielded encouraging yet nuanced results across different clinical contexts. For sleep-related disorders, LPSGM demonstrated robust performance in narcolepsy and hypersomnia classification on the large-scale MNC dataset (773 subjects), achieving subject-wise accuracy of 80.47% for three-class classification and AUC of 0.9663 for binary Normal versus Abnormal discrimination. Critically, the model maintained diagnostic reliability when transferred from MNC to the independent HANG7 dataset without additional training (AUC=0.8791), validating cross-institutional generalizability essential for clinical deployment across diverse healthcare settings.

Depression screening presented more complex findings requiring careful interpretation. On the SYSU dataset, LPSGM achieved exceptional performance with perfect subject-wise accuracy (100%) for Healthy versus Depressive classification. However, performance declined substantially on the larger APPLES dataset (subject-wise accuracy = 72.0%, AUC = 0.7236). Several factors may explain this discrepancy. First, the SYSU dataset comprised distinctly contrasting populations—healthy controls and patients with major depressive disorder—representing extreme phenotypic differences that facilitate classification. Second, APPLES employed a surrogate depression label derived from medical history records rather than formal clinical diagnosis, introducing potential labeling inaccuracies. Third, APPLES exclusion criteria specifically excluded patients with Hamilton Depression Rating Scale (HAMD) scores of 24 or greater, suggesting that even depression-positive subjects exhibited relatively mild symptomatology. Finally, the APPLES cohort consisted primarily of obstructive sleep apnea patients, whose sleep architecture is inherently disrupted by frequent arousals and respiratory events, potentially masking depression-specific sleep pattern alterations and complicating PSG-based mood disorder detection. These depression screening results require cautious interpretation and warrant validation through larger-scale studies with standardized clinical assessments and diverse patient populations. Nevertheless, our findings demonstrate LPSGM's potential as a foundation for PSG-based brain disorder screening, establishing an important initial framework for integrating sleep physiology with psychiatric diagnosis.

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. Additionally, current disease classification operates at the sequence level, processing samples comprising L consecutive epochs as described in Methods. While this temporal context proves sufficient for sleep staging tasks, brain disorder diagnosis may require analysis of

longer contextual windows to detect subtle pathological sleep patterns that manifest across extended time scales. 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 brain 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.

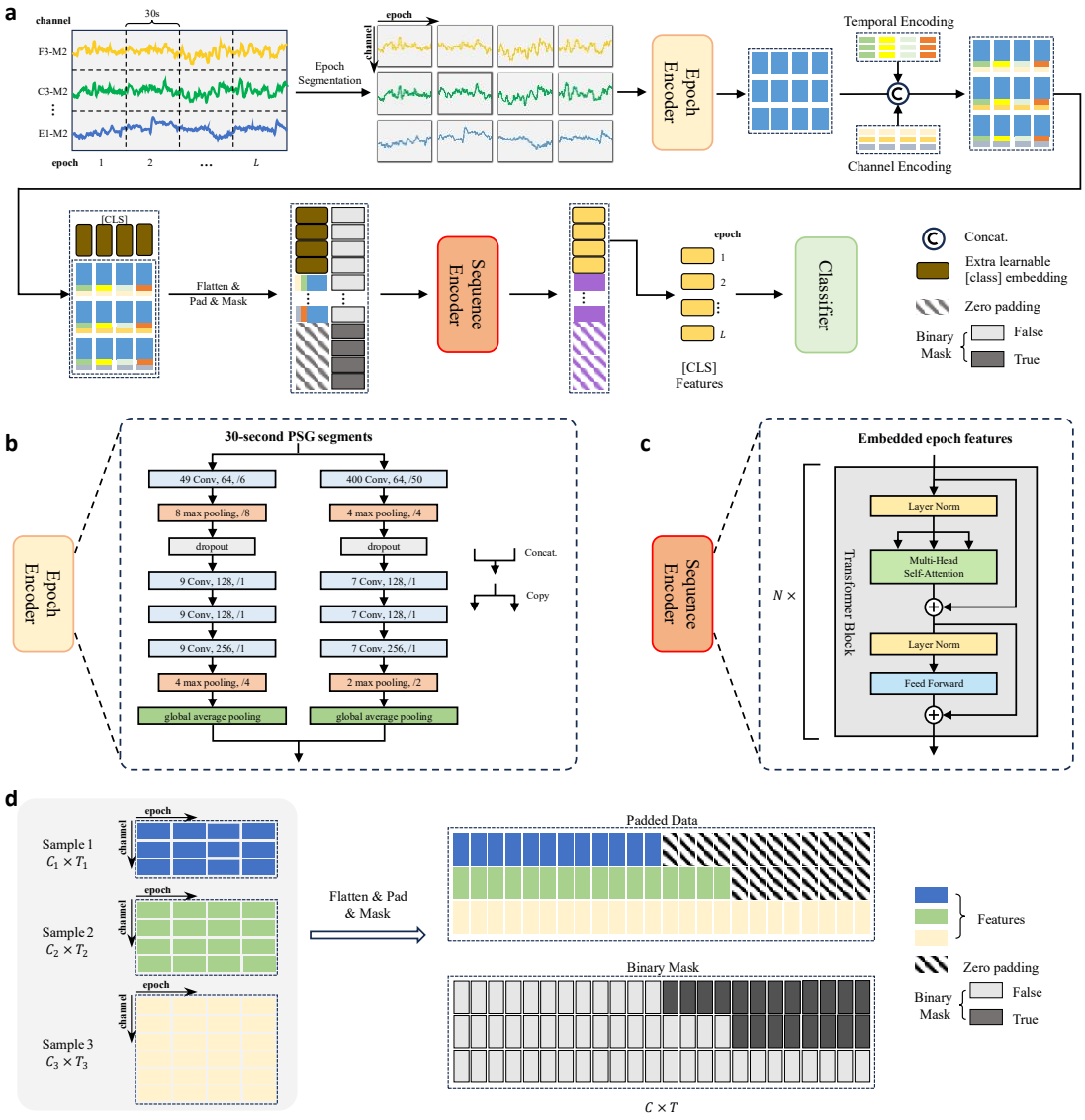


Fig. 5: Overall architecture of LPSGM. (a) LPSGM consists of an Epoch Encoder, Sequence Encoder, and Classifier, designed for both sleep staging and disorder diagnosis. (b) The Epoch Encoder employs a dual-branch CNN to extract local intra-epoch features from each 30-second PSG segment, using small and large convolutional filters to capture high- and low-frequency EEG features, respectively. (c) The Sequence Encoder consists of a series of N 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 $x_i = \{x_i^1, x_i^2, \ldots, x_i^L\}$ 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:

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

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

Padding & Masking

In cross-dataset hybrid training, 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:

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

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 $\tilde{e}_i$. These tokens share the dimensionality of the encoded features ($d + d_{ce} + d_{te}$), thereby extending the sequence to $\tilde{e}_i^{'}$ 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 $\ell = 1, \dots, N$) is defined as:

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

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

The final output is computed as:

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

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:

$$E_{cls} = E_{out}[0:L] \tag{6}$$

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:

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

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 brain 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], MROS[35,45], MESA[71] and MNC[6,35] 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. MASS[70] dataset can be accessed from https://borealisdata.ca/dataverse/MASS. 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).
70. O'Reilly, C., Gosselin, N., Carrier, J. & Nielsen, T. Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. *J. Sleep Res.* **23**, 628–635 (2014).
71. Chen, X. et al. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* **38**, 877–888 (2015).

Acknowledgements

This work was supported by STI2030-Major Projects (2022ZD0212400, 2021ZD0200404) National Natural Science Foundation of China (82371453), Key R&D Program of Zhejiang（2024C03006，2024C04024, 2024ZY01010）, Fundamental Research Funds for the Central Universities （ 2025ZFJH01-01 ） and the Construction Fund of Key Medical Disciplines of Hangzhou (2025HZGF10)

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

Supplementary Information for

A Unified Flexible Large Polysomnography Model for Sleep Staging and Brain 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

Supplementary Methods

Supplementary Figures

Supplementary Tables

Supplementary Notes

Supplementary Methods

Datasets

To the best of our knowledge, this work represents the most extensive use of data for training a large-scale sleep staging model to date. We aggregated approximately 24,000 full-night PSG recordings, encompassing roughly 220,500 hours of sleep data, from 16 publicly available datasets as source domains for training. A summary of the public datasets is provided below.

(1) The Apnea Positive Pressure Long-term Efficacy Study (APPLES) is a multi-center dataset that includes overnight PSG recordings from 1,104 patients with obstructive sleep apnea syndrome (OSAS). The dataset includes four EEG and two EOG channels, and was scored using Rechtschaffen and Kales (R&K) criteria. For depression screening evaluation, we utilized the depression medical history variable (depressionmedhxhp) from the physician's history and physical form, which categorizes subjects as: None (n=841), Resolved (n=87), Ongoing (n=163), and Accidental skip (n=7). Subjects with ongoing depression were classified as the Depression group, while those with None or Resolved status were classified as Healthy controls, excluding subjects with accidental skip entries.

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

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

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

(5) The Institute of Systems and Robotics, University of Coimbra (ISRUC) dataset consists of 126 PSG recordings from the Sleep Medicine Center of the Hospital of the University of Coimbra, Portugal. The dataset comprises three groups of data. Data in group one concerning 100 subjects, with one recording session per subject. Data in group two is gathered from 8 subjects and two recording sessions were performed per subject. Data in group three is collected from one recording session related to 10 healthy subjects. The dataset includes six EEG and two EOG channels and was scored according to the AASM criteria.

(6) The St. Vincent's University Hospital (SVUH) dataset contains 25 full overnight PSG with suspected sleep-disordered breathing. The dataset contains two EEG and two EOG channels and was scored according to R&K criteria.
(7) P2018 (You Snooze You Win: The PhysioNet/Computing in Cardiology Challenge 2018) dataset was contributed by the Massachusetts General Hospital's (MGH) Computational Clinical Neurophysiology Laboratory (CCNL), and the Clinical Data Animation Laboratory (CDAC). The dataset consists of 994 training examples and 989 test examples, with only the training data having labels publicly available. We only use the training data, which includes 6 EEG and 1 EOG channels, and uses the AASM criteria for sleep staging.
(8) The Stanford Technology Analytics and Genomics in Sleep (STAGES) dataset was collected on 1500 patients evaluated for sleep disorders from six centers. The dataset contains six EEG and two EOG channels and is annotated based on AASM criteria.
(9) The Apnea, Bariatric surgery, and CPAP (ABC) dataset includes 80 patients with severe OSAS, with six EEG and two EOG channels, being scored based on AASM criteria.
(10) The Nationwide Children's Hospital Sleep DataBank (NCHSDB) has 3,984 pediatric sleep studies on 3,673 unique patients conducted at NCH in Columbus, Ohio, USA between 2017 and 2019. The dataset includes six EEG and two EOG channels and using AASM criteria.
(11) The Home Positive Airway Pressure (HOMEPAP) study was a multi-center dataset that enrolled 373 patients with suspected moderate and severe OSAS. Subjects were randomized to lab-based and home-based management. We only use the lab-based subset as it includes the channels we need. The lab-based subset includes six EEG and two EOG channels with AASM criteria annotation.
(12) The Childhood Adenotonsillectomy Trial (CHAT) is a multi-center dataset that enrolled 1447 children with mild to moderate OSAS. The dataset includes six EEG and two EOG channels with AASM criteria annotation.
(13) The Cleveland Children's Sleep and Health Study (CCSHS) dataset consists of 515 PSG recordings from 907 children aged between 8 and 11 years old. The dataset includes two EEG and two EOG channels, and was scored according to AASM criteria.
(14) The Cleveland Family Study (CFS) is a family-based study of sleep apnea worldwide, compring 730 overnight PSG from 2284 indivisuals. The dataset includes two EEG and two EOG channels with AASM criteria.
(15) The MROS dataset enrolled 5994 men 65 years or older at six clinical centers. The dataset consists of 3929 PSG recordings with two EEG and two EOG channals, and was scored according to AASM criteria.
(16) The Sleep Heart Health Study (SHHS) is a multi-center cohort study implemented by the National Heart Lung & Blood Institute, consisting of two subset. SHHS-1 and SHHS-2 contain 5793 and 2651 overnight PSG, respectively. The dataset contains two EEG and EOG channels and was scored according to R&K criteria.

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

(1) The HANG7 dataset was acquired from the Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine. Data were randomly selected from the sleep department and collected between October 2018 and July 2022 using the Australian Compumedics Grael polysomnography system. Data collection was conducted at Zhejiang University with Institutional Review Board approval, and written consent was obtained from all subjects or their caregivers. The dataset comprises PSG recordings from 127 subjects (73 females and 54 males, aged 11-57 years with mean age 23.1 ± 9.9 years), including 33 healthy controls, 51 patients diagnosed with narcolepsy (13 with type 1 and 38 with type 2), and 43 patients exhibiting hypersomnia symptoms associated with anxiety and depression but not meeting the diagnostic criteria for narcolepsy. The dataset includes six EEG and two EOG channels sampled at a frequency of 512 Hz. PSG recordings were scored by experienced clinicians according to AASM criteria.

(2) The SYSU dataset includes two groups: 20 healthy controls and 24 patients with major depressive disorder (MDD). Data were collected using the Compumedics Profusion EEG recording system with Neuvo amplifier. All subjects were equipped with the same device to collect overnight PSG signals. This study received approval from the Ethics Committee of Guangdong 999 Brain Hospital (approval number: 2020-010-059). The experiments involving healthy individuals were conducted at the sleep laboratory of Sun Yat-sen University, encompassing 80 PSG recordings from 20 healthy undergraduate and graduate students (10 males and 10 females, mean age 21.9 ± 1.2 years) sampled at 500 Hz. Data collection involved 4 consecutive nights per participant, and all participants were normal sleepers without sleep disorders (Pittsburgh Sleep Quality Index: 5.2 ± 3.0, sleep efficiency: 91.05% ± 4.35%). The MDD dataset comprised 24 PSG recordings from 24 MDD patients (13 males and 11 females, mean age 21.4 ± 7.1 years), who were diagnosed by two experienced psychiatrists based on the criteria of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The scores of 24-item Hamilton Depression Scale (HAMD-24) and the self-rating Depression Scale (SDS) were 22.6±6.20 and 65.57±9.53, respectively. All patients were without drug abuse, suicide risk, pregnancy, present or history of head injuries, seizures, or epilepsy. The EEG signals were sampled at 256 Hz. Both groups included six EEG and two EOG channels and were scored according to the AASM scoring manual by two well-trained sleep technologists.

To evaluate the performance of our fine-tuned model for disease diagnosis, we utilized the Mignot Nature Communications (MNC) dataset, a multi-center collection of raw polysomnography data from 10 different cohorts recorded at 12 sleep centers

across 3 continents. In this study, we utilized six cohorts from the MNC dataset, encompassing a total of 773 PSG recordings. These cohorts include the Chinese Narcolepsy Cohort (CNC), Danish Hypersomnia Cohort (DHC), French Hypersomnia Cohort (FHC), Italian Hypersomnia Cohort (IHC), Korean Hypersomnia Cohort (KHC), and the Stanford Sleep Cohort (SSC). For our analysis, we used only the raw PSG signals and diagnostic labels from this dataset, without utilizing hypnogram annotations. The dataset comprises three diagnostic categories: non-narcolepsy controls (*n*=310), Type 1 narcolepsy patients (*n*=254), and patients with other hypersomnia conditions (*n*=209). Type 1 narcolepsy diagnosis was confirmed through clinical criteria including cataplexy and/or cerebrospinal fluid hypocretin-1 levels $\leqslant$110 pg/ml. The other hypersomnia group included patients with Type 2 narcolepsy, idiopathic hypersomnia, and other forms of excessive daytime sleepiness. The detailed diagnostic distribution across cohorts is provided in Supplementary Table 4.

To further evaluate sleep staging performance across diverse populations and recording protocols, we additionally tested LPSGM on three public datasets: the Multi-Ethnic Study of Atherosclerosis (MESA) and two subsets from the Montreal Archive of Sleep Studies (MASS). The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep dataset comprises 2,033 overnight unattended polysomnography recordings collected between 2010-2012 from a multi-ethnic population including African American, White/Caucasian, Hispanic, and Chinese-American participants. The dataset contains one EEG channel and two EOG channels, with sleep staging performed according to AASM criteria. The Montreal Archive of Sleep Studies (MASS) is an open-access database of laboratory-based polysomnography recordings. MASS comprises five subsets (SS1-SS5), from which we selected SS1 and SS3 as they employ 30-second epoch annotations consistent with clinical practice, while other subsets use 20-second epochs. MASS-SS1 comprises 53 overnight PSG recordings, while MASS-SS3 includes 62 overnight PSG recordings. Both subsets contain six EEG and two EOG channels, and were scored according to AASM criteria.

Preprocessing

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

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

$$x[c] = \frac{x[c] - \mathrm{mean}(x[c])}{\mathrm{std}(x[c])}, c \in C \quad (1)$$

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

We adhered to the current AASM sleep staging standards. For datasets originally

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

Channel Standardization and Indexing

A key objective of LPSGM is to flexibly process PSG recordings from heterogeneous sources (summarized in Supplementary Table 2), which vary in both the number of available channels and their specific montages. The model achieves this flexibility not by enforcing a strict channel order, but by mapping available channels to a fixed set of channel identities. This process ensures that channel-specific information is correctly interpreted by the model, regardless of the input configuration.

The model internally maintains eight learnable channel embedding vectors $\{ce_1, ce_2, \ldots, ce_8\}$, as described in the Methods. These embeddings were randomly initialized and optimized during training to represent the unique characteristics of the eight AASM-recommended channels selected for this study: F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1, E1-M2 and E2-M1.

During preprocessing for any given dataset, we implemented a protocol to map the dataset's provided channels to these eight model-specific identities:

(1) **EEG Channel Mapping (F3, F4, C3, C4, O1, O2):** For each of the six target EEG locations, we first identified the corresponding channel in the dataset's metadata.
AASM-Referenced: If the channel was already referenced according to AASM guidelines (e.g., F3-M2, F3-A2, F4-M1, F4-A1), it was directly mapped to its corresponding model identity (e.g., F3-A2 was mapped to the F3-M2 identity).
Globally-Referenced: If the channel used a global reference (e.g., F3, F3-REF), we checked for the presence of AASM reference signals (M1, M2, A1, A2) in the recording file. If available, we computationally re-referenced the channel by calculating the differential (e.g., F3 signal minus M2 signal) to create the target channel. If AASM reference signals were not available, we directly adopted the globally-referenced channel (e.g., F3) and mapped it to its corresponding model identity (e.g., F3-M2).

(2) **EOG Channel Mapping (E1, E2):** The same principles were applied to the EOG channels. An additional rule was implemented for datasets providing a differential EOG channel (e.g., E1-E2, LOC-ROC); in such cases, the signal was mapped to the E1-M2 model identity.

This identity-based mapping, rather than a fixed-order input, is foundational to the model's design. Before the Epoch Encoder's features are fed into the Sequence Encoder (Transformer), each channel's feature vector is concatenated with its corresponding learnable embedding (e.g., ce_3 for C3-M2). The Transformer architecture, being inherently permutation-invariant, processes these tagged feature vectors in parallel.

This mechanism, combined with the Padding & Masking strategy (see Methods), allows the model to inherently manage variable numbers of input channels, seamlessly handling both multi-channel and minimal-channel configurations without structural modification or retraining.

For a comprehensive implementation of the preprocessing scripts used for each dataset, please refer to our publicly available code repository.

Classification Metrics with Probabilistic Extensions

Let n be the total number of 30-s epochs retained, C be the number of sleep-stage classes. For each epoch $i = 1, \dots, n$

$$y_i^{\text{true}} = \left(y_{i,1}^{\text{true}}, \dots, y_{i,C}^{\text{true}}\right) \text{ and } y_i^{\text{pred}} = \left(y_{i,1}^{\text{pred}}, \dots, y_{i,C}^{\text{pred}}\right)$$

are the true-label and predicted-probability vectors (each in $[0,1]^C$ and summing to 1).

Probabilistic Accuracy

The probabilistic accuracy is an extended metric that calculates the average match between the true and predicted probabilistic distributions across all samples. It quantifies how well the predicted probabilities align with the true probabilities on a per-sample basis.

$$\text{Acc} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C} y_{i,k}^{\text{true}} y_{i,k}^{\text{pred}} \times 100\% \tag{2}$$

This metric essentially computes the weighted sum of the element-wise product of the true and predicted probability distributions for each sample, then averages these values across all samples.

Probabilistic Cohen's Kappa

The probabilistic Cohen's Kappa extends the traditional Cohen's Kappa to accommodate probabilistic predictions. It measures the agreement between the true and predicted distributions, accounting for the possibility of multiple correct classes per sample.

First construct the soft confusion matrix

$$M_{ab} = \sum_{i=1}^{n} y_{i,a}^{\text{true}}\, y_{i,b}^{\text{pred}} \quad (a, b = 1, \dots, C) \tag{3}$$

and denote its total weight $N = \sum_{a,b} M_{ab}$. Then the observed agreement is

$$P_o = \frac{1}{N}\sum_{k=1}^{C} M_{kk} \tag{4}$$

and the chance agreement

$$P_e = \frac{1}{N^2}\sum_{k=1}^{C}\left(\sum_{b=1}^{C} M_{kb}\right)\left(\sum_{a=1}^{C} Mak\right) \tag{5}$$

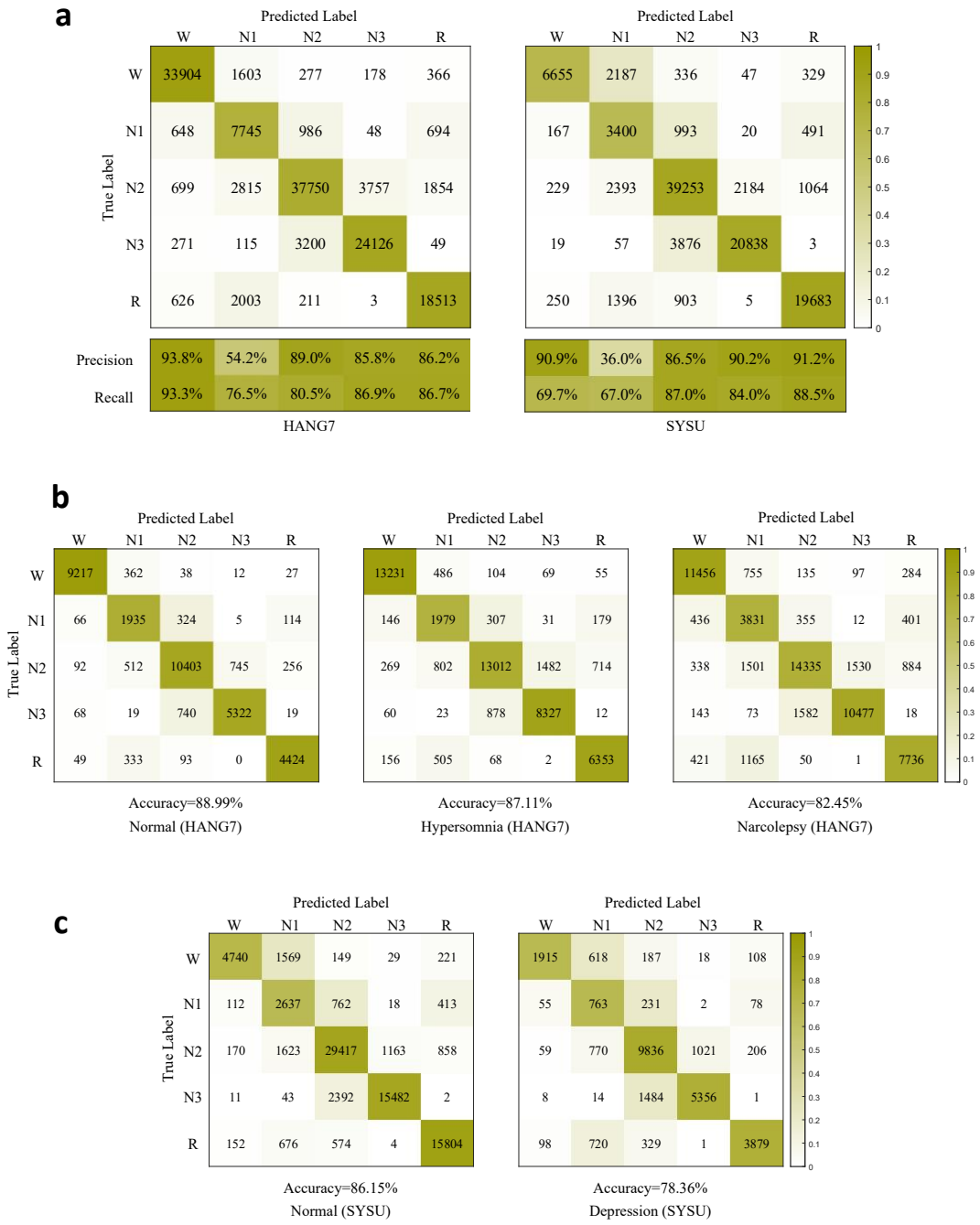
Finally,

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

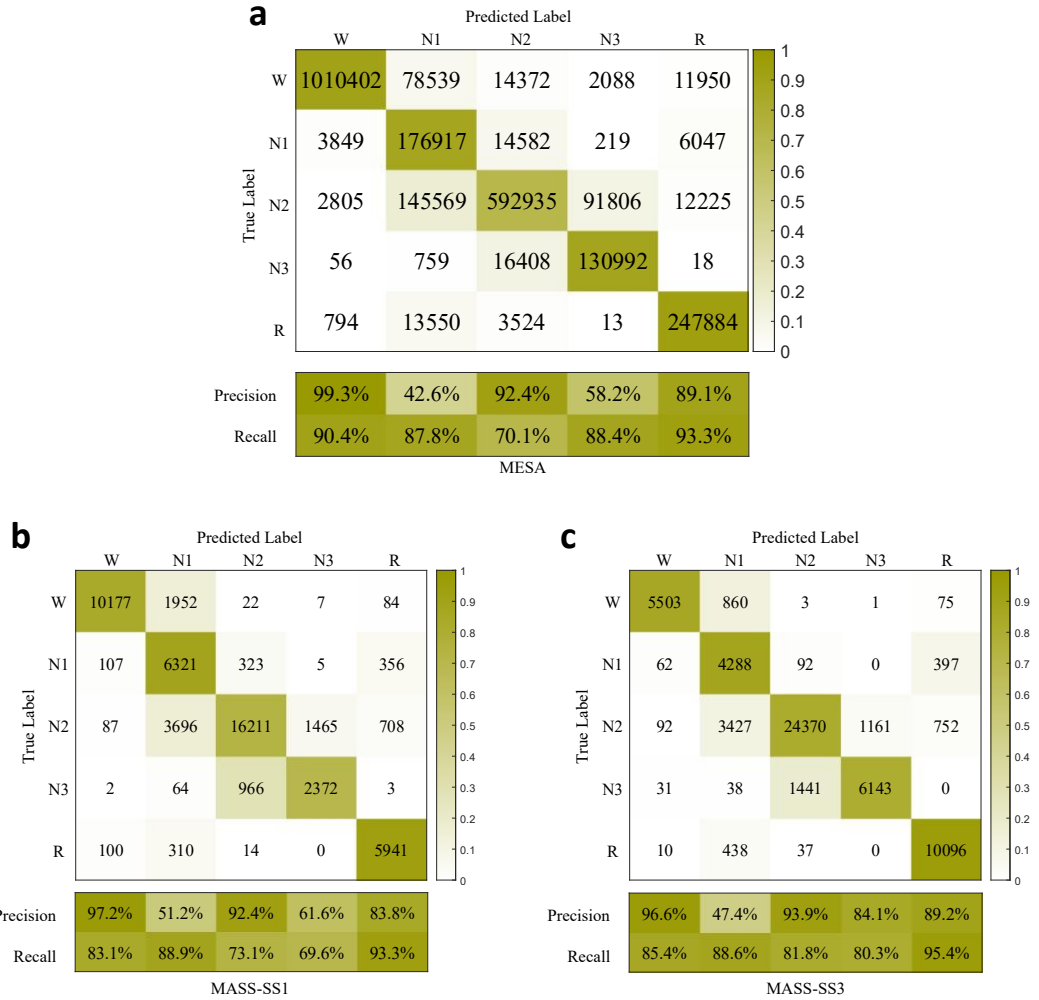
This extension allows for a more nuanced assessment of model performance when dealing with probabilistic predictions and scenarios where multiple classes may be partially correct for a given sample.

It should be noted that when the probability distributions are deterministic (i.e., the probability is concentrated in a single class for each sample, such as [1,0,0], [0,1,0], or [0,0,1]), the probabilistic accuracy and Cohen's Kappa reduce to their traditional counterparts. In such cases, the probabilistic metrics yield identical results to the conventional accuracy and Cohen's Kappa, ensuring consistency with standard evaluation practices when dealing with hard classifications.

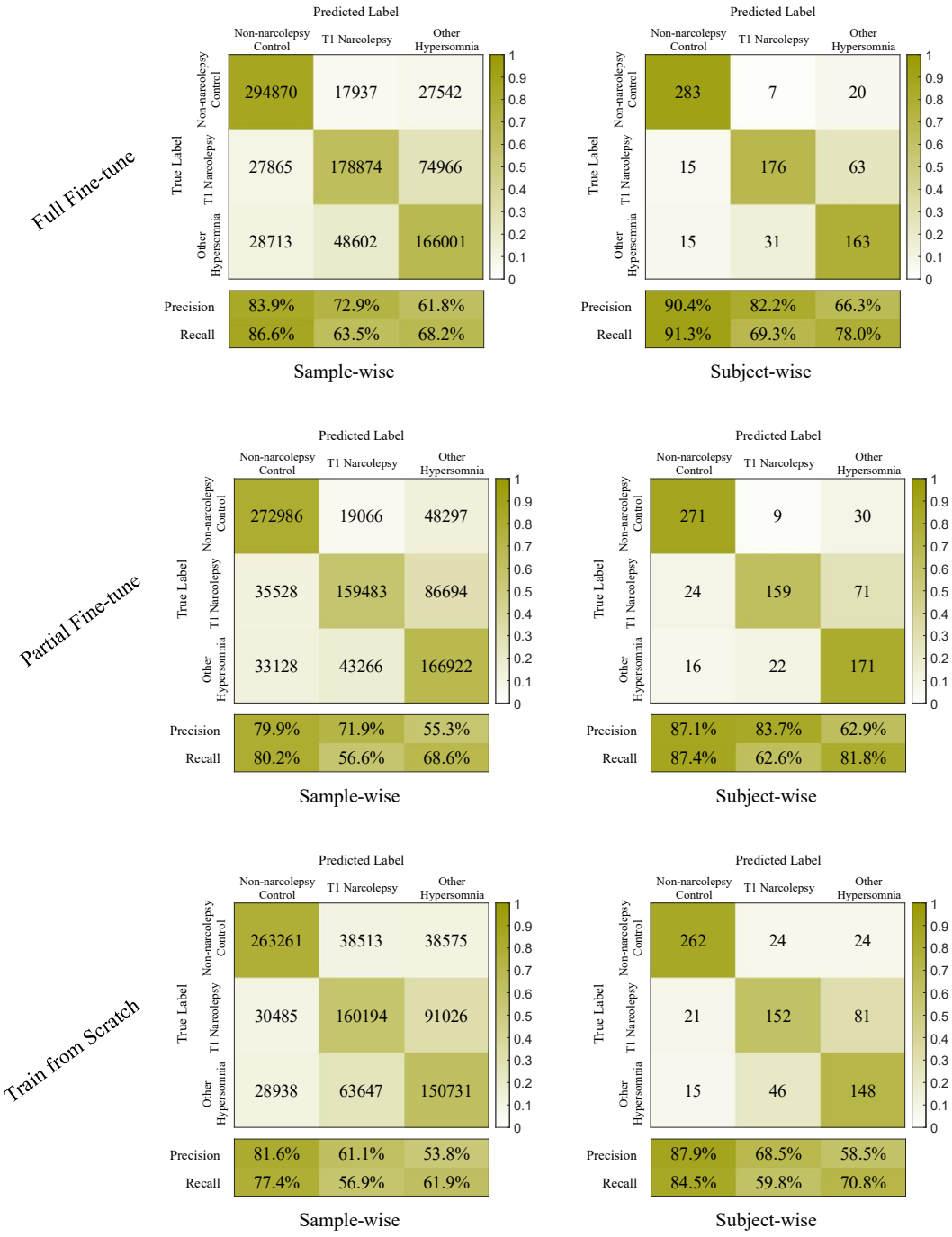
Supplementary Figures



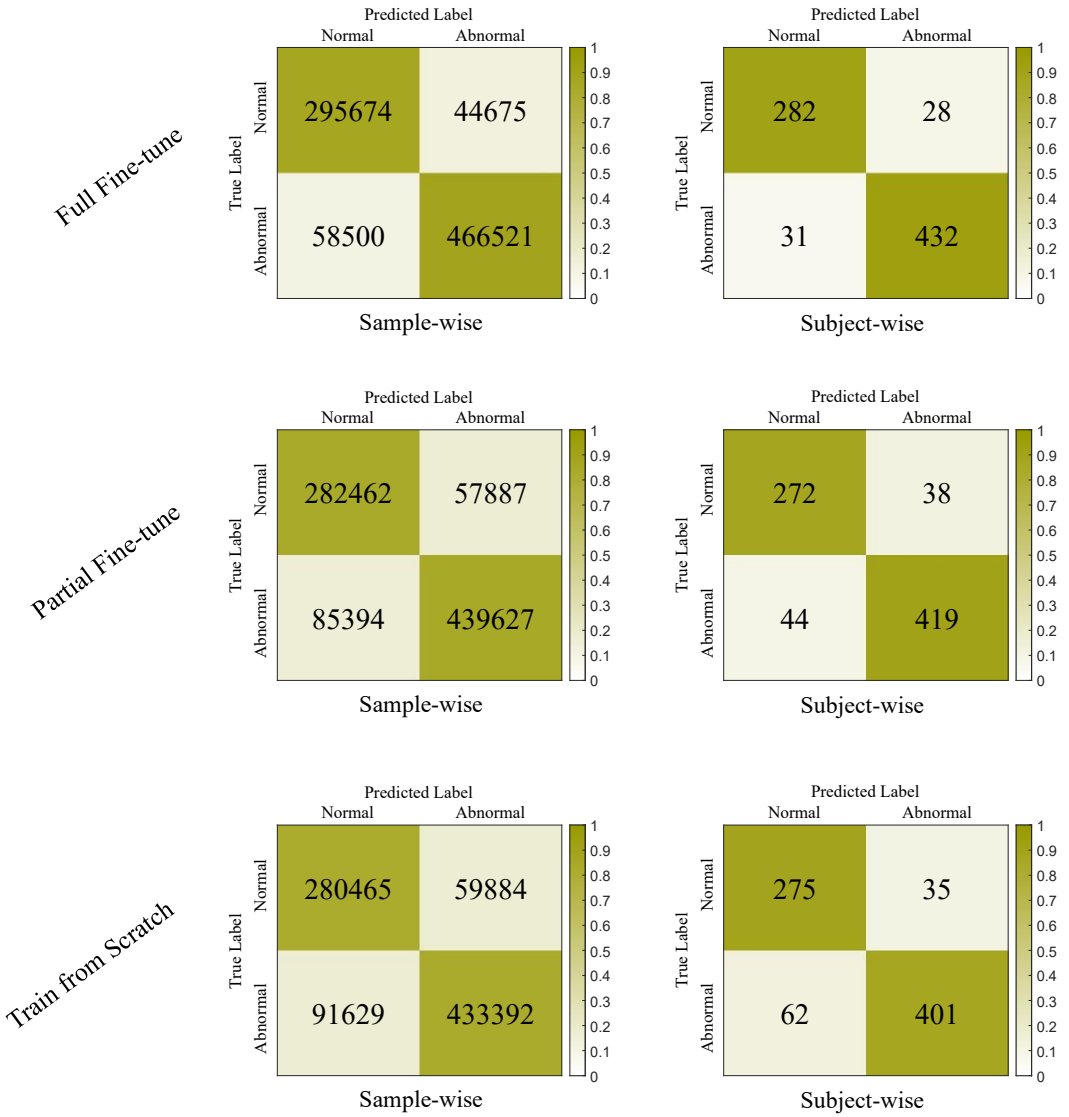
Supplementary Figure 1: Cross-center sleep staging confusion matrices of LPSGM. Panel (a) presents the confusion matrices of LPSGM on the HANG7 and SYSU datasets. The values in each cell represent the number of 30-second epochs classified into each sleep stage, with darker shades indicating higher counts. Precision and recall scores for each sleep stage are reported below. Panel (b) shows confusion matrices for different subject groups within the HANG7 dataset. Panel (c) displays confusion matrices for different subject groups within the SYSU dataset.



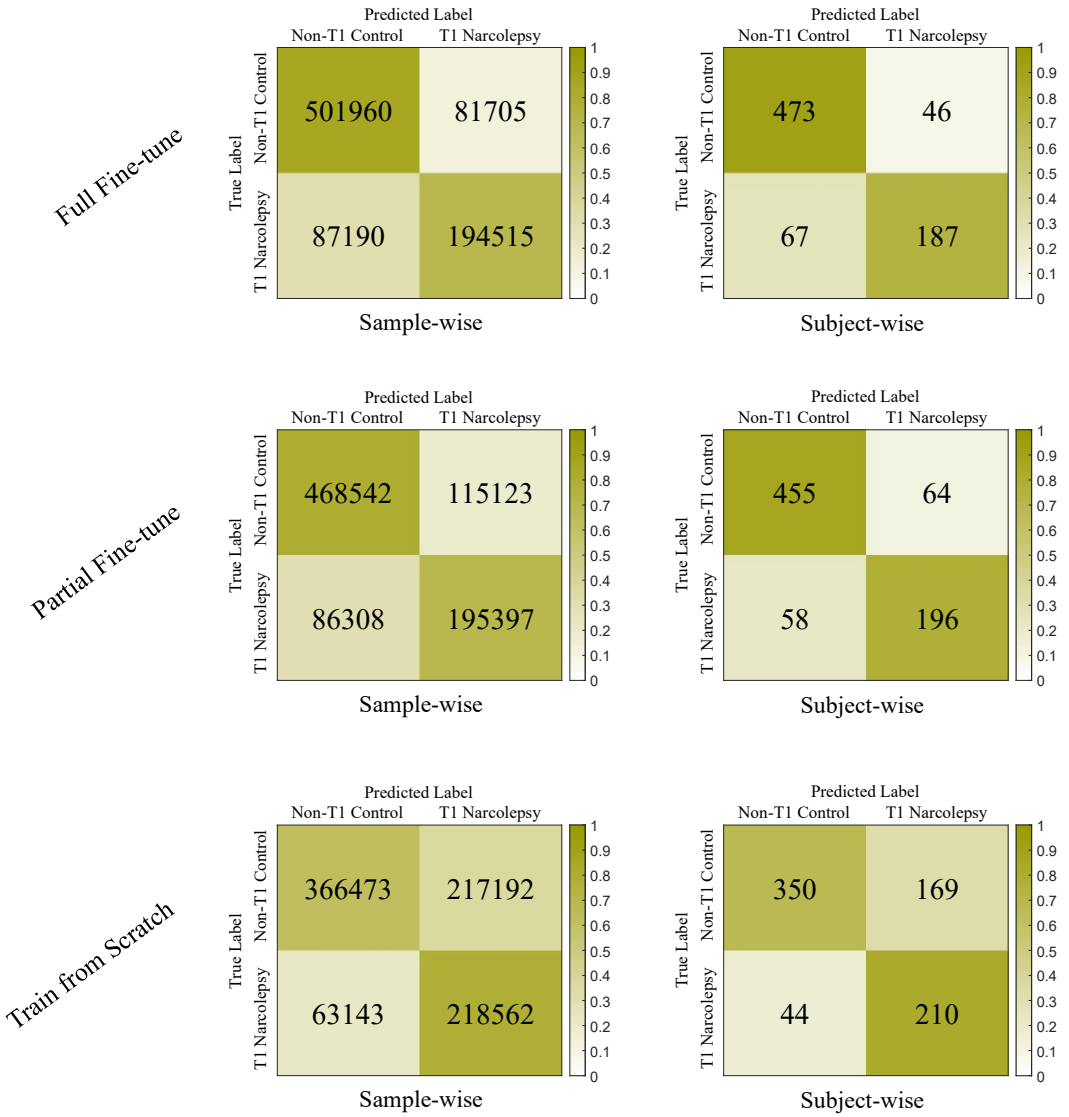
Supplementary Figure 2: Cross-center sleep staging confusion matrices of LPSGM on the (a) MESA, (b) MASS-SS1 and (c) MASS-SS3 datasets. The values in each cell represent the number of 30-second epochs classified into each sleep stage, with darker shades indicating higher counts. Precision and recall scores for each sleep stage are reported below. MESA achieved overall accuracy of 83.74%, macro-F1 of 78.61%, and Cohen's kappa of 77.37%. MASS-SS1 achieved overall accuracy of 79.98%, macro-F1 of 77.98%, and Cohen's kappa of 73.23%. MASS-SS3 achieved overall accuracy of 84.97%, macro-F1 of 82.83%, and Cohen's kappa of 78.75%. Per-class F1 scores for MESA were: W (94.64%), N1 (57.35%), N2 (79.74%), N3 (70.17%), R (91.15%). Per-class F1 scores for MASS-SS1 were: W (89.61%), N1 (64.98%), N2 (81.66%), N3 (65.38%), R (88.30%). Per-class F1 scores for MASS-SS3 were: W (90.66%), N1 (61.74%), N2 (87.43%), N3 (82.14%), R (92.20%).



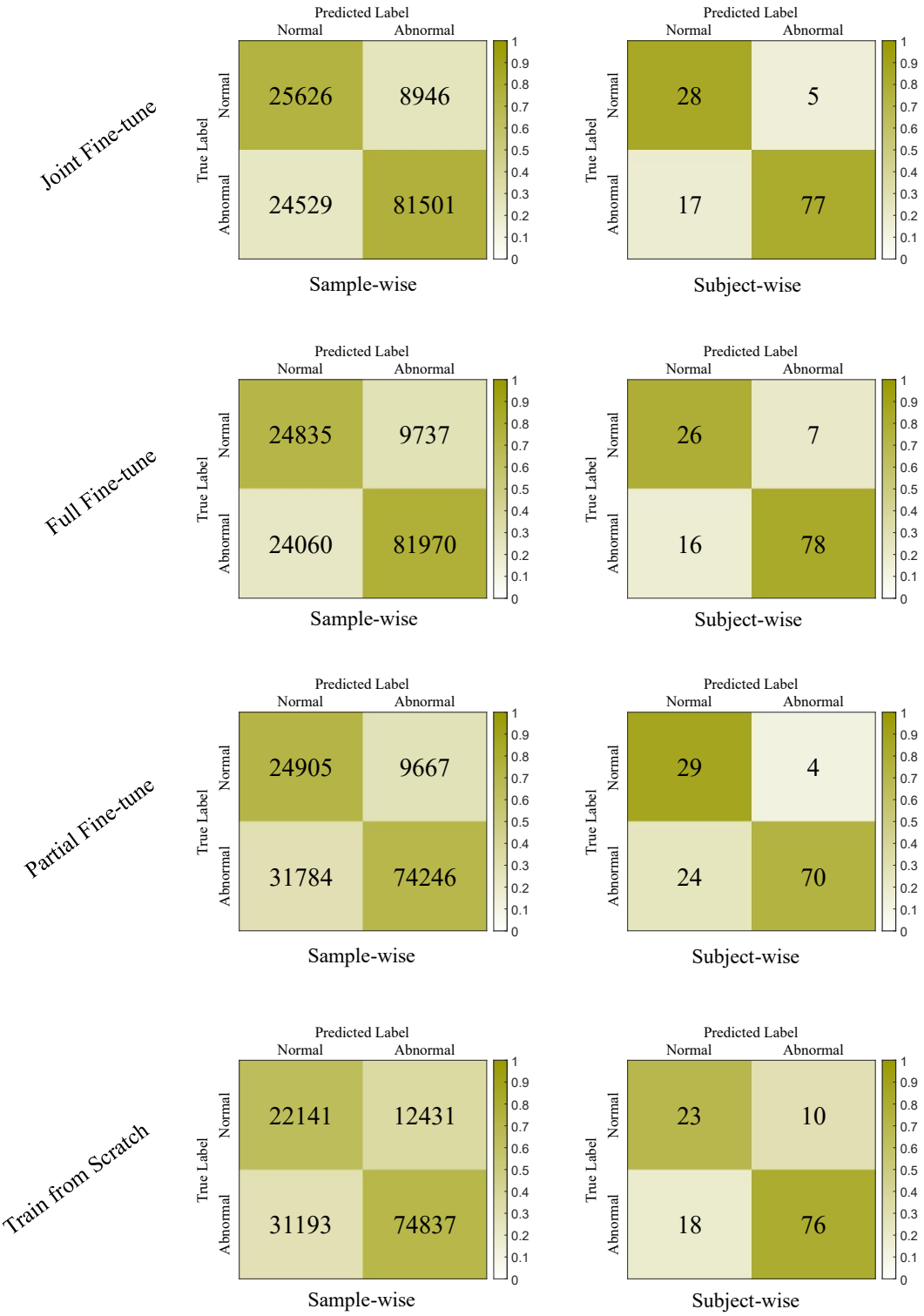
Supplementary Figure 3: Confusion matrices for three-class classification on MNC dataset distinguishing Non-narcolepsy Control, T1 Narcolepsy, and Other Hypersomnia. Results are shown for Full Fine-tune, Partial Fine-tune, and Train from Scratch approaches, evaluated at both sample-wise (left panels) and subject-wise (right panels) levels. Each matrix displays absolute prediction counts with precision and recall values for individual classes presented below the confusion matrices.



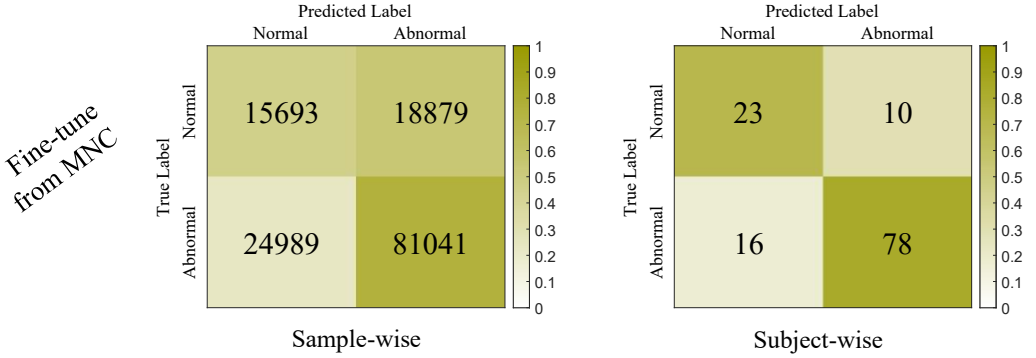
Supplementary Figure 4: Confusion matrices for binary classification on MNC dataset discriminating Normal (Non-narcolepsy) versus Abnormal (T1 Narcolepsy and Other Hypersomnia). Results are presented for Full Fine-tune, Partial Fine-tune, and Train from Scratch approaches, with evaluation at both sample-wise (left panels) and subject-wise (right panels) levels. Matrices show absolute prediction counts for each true-predicted label combination.



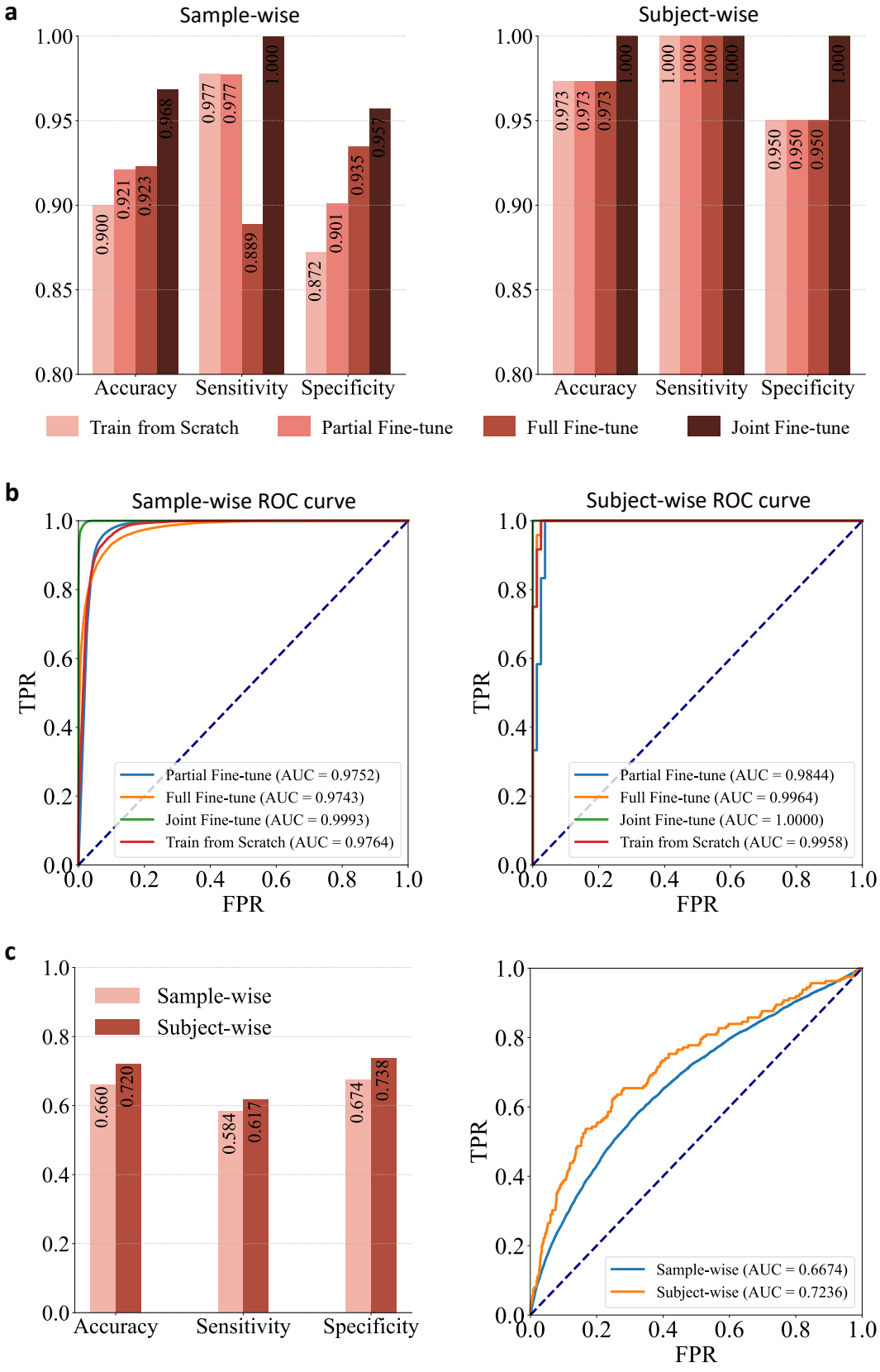
Supplementary Figure 5: Confusion matrices for binary classification on MNC dataset identifying T1 Narcolepsy, comparing Non-T1 Control (Non-narcolepsy and Other Hypersomnia) versus T1 Narcolepsy. Results are displayed for Full Fine-tune, Partial Fine-tune, and Train from Scratch approaches, evaluated at both sample-wise (left panels) and subject-wise (right panels) levels. Matrices present absolute prediction counts for true versus predicted classifications.



Supplementary Figure 6: Confusion matrices for binary classification on HANG7 dataset discriminating Normal (Healthy Control) versus Abnormal (Narcolepsy and Hypersomnia). Results are shown for Joint Fine-tune, Full Fine-tune, Partial Fine-tune, and Train from Scratch approaches, with evaluation at both sample-wise (left panels) and subject-wise (right panels) levels. Matrices display absolute prediction counts demonstrating within-dataset cross-validation performance.

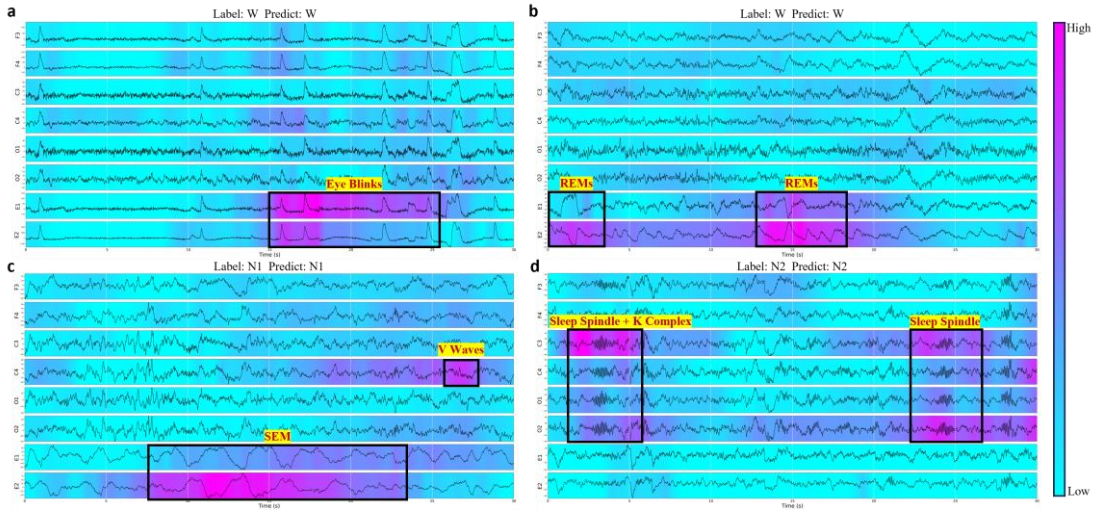


Supplementary Figure 7: Confusion matrices for cross-dataset transfer evaluation from MNC to HANG7, showing binary classification performance for Normal versus Abnormal discrimination without additional fine-tuning on the target dataset. Results demonstrate the generalizability of models fine-tuned on MNC dataset when directly applied to HANG7 cohort, evaluated at both sample-wise (left panel) and subject-wise (right panel) levels.



Supplementary Figure 8: LPSGM performance for depression screening across datasets. Performance evaluation for Healthy versus Depressive classification on the SYSU dataset, comparing Train from Scratch, Partial Fine-tune, Full Fine-tune, and

Joint Fine-tune approaches through both accuracy metrics (a) and receiver operating characteristic (ROC) curve analysis (b), with results displayed at sample-wise (left panels) and subject-wise (right panels) levels and corresponding area under the curve (AUC) values for all fine-tuning strategies. (c) Performance evaluation for depression screening on the APPLES dataset using Full Fine-tune approach, showing accuracy, sensitivity, and specificity metrics (left) and ROC curve analysis (right) at both sample-wise and subject-wise levels. Depression classification was based on ongoing depression status from medical history records in the APPLES dataset.



Supplementary Figure 9: Grad-CAM visualizations of LPSGM predictions for subjects lacking a posterior dominant rhythm (PDR). According to AASM guidelines, a subset of individuals (approximately 10%) do not generate an alpha rhythm upon eye closure, necessitating reliance on alternative biomarkers for scoring Wakefulness and Stage N1. This figure displays representative 30-second epochs from such subjects, where heatmaps (magenta) indicate regions with high positive contribution to the model's decision. (a) Stage W: The model directs its attention to the EOG channels (E1/E2), highlighting high-amplitude waveforms characteristic of Eye Blinks. (b) Stage W: The model focuses on Rapid Eye Movements (REMs) in the EOG channels (E1/E2), which serve as an alternative criterion for scoring Wakefulness in the absence of alpha rhythm. (c) Stage N1: Lacking alpha attenuation as a marker, the model identifies N1 by focusing on Slow Eye Movements (SEM) in the EOG channels and Vertex Sharp Waves (V Waves) in the central EEG (C4). (d) Stage N2: The model consistently attends to characteristic K-complexes and Sleep Spindles, demonstrating that the detection of these phasic events remains robust independent of the background rhythm.

Supplementary Tables

Supplementary Table 1: Definitions of hypnogram metrics.

Hypnogram Metrics		Definition
Sleep Latency (min)		
N1		The time from the onset of the first non-wake sleep stage to the onset of the first Stage N1 sleep.
N2		The time from the onset of the first non-wake sleep stage to the onset of the first Stage N2 sleep.
N3		The time from the onset of the first non-wake sleep stage to the onset of the first Stage N3 sleep.
REM		The time from the onset of the first non-wake sleep stage to the onset of the first REM sleep episode.
Sleep Duration (min)		
TST (Total Sleep Time)		The total duration of sleep during the sleep period, calculated as the sum of all NREM and REM sleep stages.
REM		The cumulative duration of all REM sleep episodes during the sleep period.
NREM		The cumulative duration of all non-REM sleep stages (N1, N2, and N3) during the sleep period.
SWS		The cumulative duration of Stage N3 sleep (slow-wave sleep) during the sleep period.
Sleep Stages		
W (SPT)	Episodes (#)	The number of wake episodes (Wake after Sleep Onset, WASO) during the sleep period time (SPT).
	Duration (min)	The total time spent awake during the sleep period time (SPT).
R	Duration (min)	Identical to NREM.
	TST (%)	The proportion of REM sleep relative to the total sleep time (TST).
N1	Duration (min)	The cumulative duration of all Stage N1 sleep episodes during the sleep period.
	TST (%)	The proportion of Stage N1 sleep relative to the total sleep time (TST).
N2	Duration (min)	The cumulative duration of all Stage N2 sleep episodes during the sleep period.
	TST (%)	The proportion of Stage N2 sleep relative to the total sleep time (TST).
N3	Duration (min)	The cumulative duration of all Stage N3 sleep episodes during the sleep period.
	TST (%)	The proportion of Stage N3 sleep (SWS) relative to the total sleep time (TST).

Supplementary Table 2: Overview of sleep staging datasets used in our experiments.

Datasets		Recordings	Annotation	EEG						EOG		Channels	Sample Rate	
				F3-M2	F4-M1	C3-M2	C4-M1	O1-M2	O2-M1	E1-M2	E2-M1		EEG	EOG
APPLES		1067	R&K			√	√	√	√	√	√	6	100	100
DCSM		255	AASM	√	√	√	√	√	√	√	√	8	256	256
DOD	DOD-H	25	AASM	√	√	√				√	√	5	250	250
	DOD-O	56	AASM	√		√	√	√	√	√	√	7	250	250
HMC		151	AASM		√	√	√		√	√	√	6	256	256
ISRUC		126	AASM	√	√	√	√	√	√	√	√	8	200	200
SVUH		25	R&K			√	√			√	√	4	128	64
P2018		994	AASM	√	√	√	√	√	√	√		7	200	200
STAGES	BOGN	85	AASM	√	√	√	√	√	√	√	√	8	200	200
	STNF	525		√	√	√	√	√	√	√	√	8	256	256
	GSDV	288		√	√	√	√	√	√	√	√	8	200	200
	MSTR	286		√	√	√	√	√	√	√	√	8	256	256
	GSBB	38		√	√	√	√	√	√	√	√	8	200	200
	GSLH	51		√	√	√	√	√	√	√	√	8	200	200
	GSSA	34		√	√	√	√	√	√	√	√	8	200	200
	GSSW	131		√	√	√	√	√	√	√	√	8	200	200
	MSMI	61		√	√	√	√	√	√	√	√	8	200	200
	MSNF	35		√	√	√	√	√	√	√	√	8	200	200
	MSQW	145		√	√	√	√	√	√	√	√	8	200	200
	MSTH	31		√	√	√	√	√	√	√	√	8	256	256
	STLK	156		√	√	√	√	√	√	√	√	8	500	500
ABC		132	AASM	√	√	√	√	√	√	√	√	8	256	256

NCHSDB		3947	AASM	√	√	√	√	√	√	√	√	8	256/400/512	256/400/512
HOMEPAP		245	AASM	√	√	√	√	√	√	√	√	8	200/256	200/256
CHAT		1638	AASM	√	√	√	√	√	√	√	√	8	200	200
CCSHS		515	AASM			√	√			√	√	4	128	128
CFS		730	AASM			√	√			√	√	4	128	128
MROS		3929	AASM			√	√			√	√	4	256	256
SHHS	SHHS-1	5793	R&K			√	√			√	√	4	125	50
	SHHS-2	2651				√	√			√	√	4	128	32
HANG7		127	AASM	√	√	√	√	√	√	√	√	8	512	512
SYSU		104	AASM	√	√	√	√	√	√	√	√	8	500/256	500/256
MASS	SS1	53	AASM	√	√	√	√	√	√	√	√	8	256	256
	SS3	62	AASM	√	√	√	√	√	√	√	√	8	256	256
MESA		2033	AASM			√				√	√	3	256	256

Supplementary Table 3: Number of sleep stages of the datasets after preprocessing.

Datasets		Total	Epochs					Ratio %				
			W	N1	N2	N3	R	W	N1	N2	N3	R
APPLES		1049110	256128	147217	481003	24295	140467	24.4	14	45.8	2.3	13.4
DCSM		304266	79636	21140	113027	43637	46826	26.2	6.9	37.1	14.3	15.4
DOD	DOD-H	24662	3037	1505	11879	3514	4727	12.3	6.1	48.2	14.2	19.2
	DOD-O	54197	10660	2898	26650	5683	8306	19.7	5.3	49.2	10.5	15.3
HMC		137243	23686	15548	50083	26671	21255	17.3	11.3	36.5	19.4	15.5
ISRUC		107784	23198	14254	34661	21489	14182	21.5	13.2	32.2	19.9	13.2
SVUH		20774	4707	3016	3403	7658	1990	22.7	14.5	16.4	36.9	9.5
P2018		892262	157945	136978	377870	102592	116877	17.7	15.4	42.3	11.5	13.1
STAGES	BOGN	75930	23185	4623	29819	8236	10067	30.5	6.1	39.3	10.8	13.3
	STNF	592027	202352	57652	171620	52147	108256	34.2	9.7	29	8.8	18.3
	GSDV	218608	54699	13363	114794	10194	25558	25	6.1	52.5	4.7	11.7
	MSTR	221466	42522	23789	103118	24201	27836	19.2	10.7	46.6	10.9	12.6
	GSBB	29843	7954	2469	14423	1668	3329	26.7	8.3	48.3	5.6	11.2
	GSLH	32118	8540	2684	16485	1597	2812	26.6	8.4	51.3	5	8.8
	GSSA	27751	6998	920	15233	881	3719	25.2	3.3	54.9	3.2	13.4
	GSSW	90341	25790	5944	44213	3549	10845	28.5	6.6	48.9	3.9	12
	MSMI	45892	8035	4227	22959	4847	5824	17.5	9.2	50	10.6	12.7
	MSNF	27061	4905	1385	13454	4012	3305	18.1	5.1	49.7	14.8	12.2
	MSQW	113085	25039	13622	53319	8125	12980	22.1	12	47.1	7.2	11.5
	MSTH	23682	5036	2146	12033	1755	2712	21.3	9.1	50.8	7.4	11.5
	STLK	154691	29429	11937	76878	11426	25021	19	7.7	49.7	7.4	16.2

ABC		133000	30938	19296	52334	11761	18671	23.3	14.5	39.3	8.8	14
NCHSDB		3661376	665063	128183	1382551	874762	610817	18.2	3.5	37.8	23.9	16.7
HOMEPAP		229604	63395	24759	86718	22645	32087	27.6	10.8	37.8	9.9	14
CHAT		1957293	469804	119436	628932	464993	274128	24	6.1	32.1	23.8	14
CCSHS		691401	212027	19221	249698	110191	100264	30.7	2.8	36.1	15.9	14.5
CFS		866204	321333	26394	306264	111937	100276	37.1	3	35.4	12.9	11.6
MROS		5373725	2609224	218233	1735010	278620	532638	48.6	4.1	32.3	5.2	9.9
SHHS	SHHS-1	5863207	1691288	217583	2397460	739403	817473	28.8	3.7	40.9	12.6	13.9
	SHHS-2	3192507	1208326	111456	1147780	313790	411155	13.9	3.5	36	9.8	12.9
HANG7		142450	36337	10121	46875	27761	21356	25.5	7.1	32.9	19.5	15
SYSU		106778	9554	5071	45123	24793	22237	8.9	4.7	42.3	23.2	20.8
MASS	SS1	51293	12242	7112	22167	3407	6365	23.9	13.9	43.2	6.6	12.4
	SS3	59317	6442	4839	29802	7653	10581	10.9	8.2	50.2	12.9	17.8
MESA		2578303	1117351	201614	845340	148233	265765	43.3	7.8	32.8	5.7	10.3

Supplementary Table 4: Subject distribution across diagnostic categories in the MNC dataset cohorts.

Cohort	Non-narcolepsy Control	T1 Narcolepsy	Other Hypersomnia	Total
CNC	24	55	0	79
DHC	20	21	38	79
FHC	13	35	10	58
IHC	0	70	78	148
KHC	14	66	78	158
SSC	239	7	5	251
Total	310	254	209	773

Supplementary Tabel 5: LPSGM sleep staging performance across different channel configurations on the HANG7 dataset.

Channel Count	Channel Selection								Metrics		
	F3-M2	F4-M1	C3-M2	C4-M1	O1-M2	O2-M1	E1-M2	E2-M2	Accuracy	MF1	Kappa
8	√	√	√	√	√	√	√	√	0.8568	0.8288	0.8138
4	√		√		√		√		0.8533	0.825	0.8093
4		√		√		√		√	0.8473	0.8204	0.8022
2	√						√		0.8442	0.8163	0.7981
2			√				√		0.8407	0.8133	0.7934
2					√		√		0.8488	0.8215	0.8037
2		√						√	0.8374	0.8105	0.7898
2				√				√	0.8346	0.809	0.7858
2						√		√	0.8452	0.8189	0.7995
2	√							√	0.8417	0.8135	0.795
2			√					√	0.8363	0.8101	0.7878
2					√			√	0.8468	0.8198	0.8012
2		√					√		0.8416	0.8142	0.7951
2				√			√		0.8403	0.8141	0.793
2						√	√		0.85	0.8229	0.8053
1	√								0.8207	0.7922	0.769
1		√							0.8162	0.789	0.7635
1			√						0.8082	0.7808	0.7519
1				√					0.8172	0.7909	0.7637
1					√				0.8204	0.7919	0.7678
1						√			0.819	0.7898	0.7661
1							√		0.8322	0.8027	0.7819
1								√	0.8283	0.7995	0.7767

Supplementary Notes

Acknowledgements

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

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

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

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

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

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

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

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

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

The Mignot Nature Communications research was mostly supported by a grant from Jazz Pharmaceuticals to E.M. Additional funding came from: NIH grant R01HL62252 (to P.E.P.); Ministry of Science and Technology 2015CB856405 and National Foundation of Science of China 81420108002,81670087 (to F.H.); H. Lundbeck A/S, Lundbeck Foundation, Technical University of Denmark and Center for Healthy Aging, University of Copenhagen (to P.J. and H.B.D.S). Additional support was provided by the Klarman Family, Otto Mønsted, Stibo, Vera & Carl Johan Michaelsens, Knud Højgaards, Reinholdt W. Jorck and Hustrus and Augustinus Foundations (to A.N.O.).

NCH Sleep DataBank was supported by the National Institute of Biomedical

Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB025018.

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

The Mignot Nature Communications research was mostly supported by a grant from Jazz Pharmaceuticals to E.M. Additional funding came from: NIH grant R01HL62252 (to P.E.P.); Ministry of Science and Technology 2015CB856405 and National Foundation of Science of China 81420108002,81670087 (to F.H.); H. Lundbeck A/S, Lundbeck Foundation, Technical University of Denmark and Center for Healthy Aging, University of Copenhagen (to P.J. and H.B.D.S). Additional support was provided by the Klarman Family, Otto Mønsted, Stibo, Vera & Carl Johan Michaelsens, Knud Højgaards, Reinholdt W. Jorck and Hustrus and Augustinus Foundations (to A.N.O.). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

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