A Language Model Built on Sleep Stage Sequences Enables Efficient Sleep Assessment

Tianyou Yu[1,3†], Zhenghui Gu[1,3†], Rui Huang[1], Fei Wang[9,3], Man Li[1,3], Jingang Yu[1,3], Zhuliang Yu[1,3], Jun Zhang[10,3], Yan Xu[11], Haiteng Jiang[4,5], Wenjuan Liu[4,6], Guifeng Deng[4,6], Zhengrun Gao[7], Yiwen Wu[8], Jun Liu[8], Yu Zhang[12], Matt W Jones[13], Yuanqing Li[1,3*], Jun Xiao[2,3*], Wei Wu[7*]

[1]School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510641, China.
[2]School of Electric Power Engineering, South China University of Technology, Guangzhou, 510641, China.
[3]Pazhou Lab, Guangzhou, 510330, China.
[4]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital and Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, 310058, China.
[5]School of Brain Science and Brain Medicine, MOE Frontier Science Center for Brain Science and Brain-machine Integration, State Key Laboratory of Brain-machine Intelligence, Zhejiang University, Hangzhou, 311121, China.
[6]College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, 310058, China.
[7]Songjiang Hospital & Songjiang Research Institute, Shanghai Key Laboratory of Emotions and Affective Disorders, Shanghai Jiao Tong University School of Medicine, Shanghai, 201600, China.
[8]Department of Neurology & Institute of Neurology, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China.
[9]School of Artificial Intelligence, South China Normal University, Guangzhou, 510631, China.
[10]School of Information Engineering, Guangdong University of Technology, Guangzhou, 510320, China.
[11]Sleep Medicine Center, Department of Psychiatry, Nanfang Hospital Affiliated to Southern Medical University, Guangzhou, 510515, China.

[12]Department of Neurology, Songjiang Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, 201699, China.
[13]School of Physiology, Pharmacology & Neuroscience, University of Bristol, Bristol, BS8 1QU, UK.

*Corresponding author(s). E-mail(s): auyqli@scut.edu.cn; junxiao@scut.edu.cn; weiwuneuro@sjtu.edu.cn;
†These authors contributed equally to this work.

Abstract

Sleep assessment is fundamental to understanding sleep architecture, identifying sleep disorders, and advancing personalized sleep medicine. However, current clinical sleep assessment methods rely on time-consuming and often costly procedures, limiting their accessibility and scalability. This study introduces SleepGPT, the first GPT-based language model for efficient sleep assessment encompassing both sleep staging and disorder identification. SleepGPT leverages the sequential structure of sleep hypnograms, recognizing strong correlations between successive sleep stages to extract relevant patterns and transitions. Following self-supervised pretraining on manually annotated large-scale whole-night hypnograms, SleepGPT yielded consistent performance gains in sleep staging and disorder diagnosis across five publicly available datasets, with successful blinded replications on three independent datasets. Notably, experiments on established sleep staging benchmarks validate SleepGPT as a robust add-on module that reliably enhances the performance of existing methods. SleepGPT-powered models furthermore achieved comparable sleep staging accuracy using wearable EEG and polysomnography (PSG) in a dataset recorded simultaneously with both modalities. Moreover, a SleepGPT-powered transformer model substantially surpassed state-of-the-art performance in classifying abnormal sleep stage sequences and diagnosing Type-1 narcolepsy. These findings underscore the potential of SleepGPT-powered models as clinically translatable and scalable artificial intelligence (AI) tools for sleep assessment, opening new avenues to advancing precision medicine for sleep disorders.

Keywords: Sleep language model, automated sleep staging, sleep disorder diagnosis, hypnogram, generative pre-trained transformers

Introduction

Sleep accounts for nearly one-third of the human lifespan and is central to overall health and well-being. Disrupted sleep patterns are linked to a range of prevalent disorders, including insomnia, sleep apnea, and narcolepsy, which collectively affect hundreds of millions worldwide and contribute to serious health issues including

hypertension, diabetes, cardiovascular disease, psychiatric disorders, and neurodegenerative diseases [1, 2]. Comprehensive sleep assessment, which involves evaluating sleep stages, duration, and regularity, is foundational to understanding sleep architecture and diagnosing these disorders. The primary method for such assessment, sleep staging, segments a night's sleep into specific stages. These stages encompass wakefulness (W), rapid eye movement (REM) sleep, and non-rapid eye movement (NREM) sleep, which are further segmented into N1, N2, and N3 according to the American Academy of Sleep Medicine (AASM) standard [3]. Typically, this procedure employs nocturnal polysomnography (PSG), a composite recording featuring multiple digital signals encompassing electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) of the chin and legs, and electrocardiography (ECG), as well as measures of breathing effort, oxygen saturation, and airflow. Manual sleep stage labeling from PSG signals may involve visually inspecting each 30-second segment (epoch) of an entire night's recording. For an 8-hour sleep study, this translates to over 900 epochs, each requiring meticulous examination of multiple signals in PSG. This frame-by-frame process is time-consuming and subjective, with inter-scorer reliability reaching only 82.6% [4–6]. Consequently, there is a strong demand for more efficient sleep staging.

Recent advancements in machine learning (ML), particularly deep learning (DL), have driven substantial progress in automated sleep staging. Traditional ML approaches for sleep staging rely on manually engineered features to classify sleep stages [7, 8]. Capitalizing on the rapid advancements in DL and the increasing accessibility of large sleep datasets, deep neural networks, such as convolutional neural networks (CNNs)[9–12], recurrent neural networks (RNNs)[13–15], transformers[16, 17], and hybrid networks[18–21], have been proposed for the automated extraction of features from raw PSG signals, subsequently facilitating sleep stage classification. These deep neural networks can process either the raw multimodal time series (EEG, EOG, ECG, EMG, etc.) or time-frequency representations derived from the PSG signals [22, 23]. Moreover, sleep staging involves a classification of discrete time series, in which adjacent segments are highly correlated. Common intuition suggests that models that stage multiple consecutive epochs generally have better performance than those that stage a single epoch. Hence, RNN models equipped with memory, such as long short-term memory (LSTM) and gated recurrent unit (GRU) are commonly employed to exploit contextual information from adjacent epochs [13, 14, 24, 25].

Beyond sleep staging, an effective sleep assessment framework must also support the detection and diagnosis of sleep disorders. Sleep patterns, including the regularity and transitions between stages, are linked to numerous chronic conditions such as obesity, hypertension, and mental health disorders [2]. Traditional diagnostic processes require highly trained sleep specialists to manually analyze and interpret results, leading to the problem of inter- and intra-operator variability and resource-intensive procedures. For instance, diagnosing narcolepsy often requires a multiple sleep latency test (MSLT), a 10-hour test where patients take 4-5 naps spaced 2 hours apart. During each nap, sleep latency and REM sleep onset are measured. Narcolepsy is diagnosed if the mean sleep latency is under 8 minutes and at least two naps show

REM onset (SOREMPs), or one SOREMP is detected with a short REM latency during overnight PSG [26, 27]. Despite these criteria, patients often face a delay of 7-10 years from symptom onset to diagnosis due to symptom misinterpretation and limited testing access [28]. Automated methods for detecting sleep disorders have been proposed to increase cost-effectiveness and mitigate inter- and intra-operator variability [29]. In particular, traditional ML approaches have been employed for automated sleep disorder detection using PSG data. However, challenges such as inter-subject variability, the high dimensionality of PSG data, and limited availability of labeled instances complicate the training of ML models with robust generalization capabilities. To address these challenges, various features, such as time and frequency representations of single-lead or multichannel EEG or ECG signals [30], disorder-specific events or waveforms [31], and sleep macrostructure statistics [26], have been used to train classifiers for sleep disorder diagnosis tasks, though with limited success. The lack of generalization capabilities and the need for extensive feature engineering have hindered the development of robust and accurate sleep disorder diagnostic models. To date, no study has employed raw sleep stage annotation sequences to diagnose sleep disorders in an end-to-end manner.

A whole-night sleep stage sequence, or hypnogram, obtained through manual annotation or automated models, is crucial for quantifying sleep macrostructure, including cycles, stage-specific durations, transitions, latency, and efficiency. A typical sleep cycle progresses from wakefulness through light and deep sleep, culminating in REM, and repeats approximately 4 to 6 times per night, each lasting around 90 minutes. These cycles reflect fundamental neurophysiological mechanisms, suggesting strong correlations between consecutive stages and revealing inherent sequential and transition patterns that can improve the accuracy of sleep staging and disorder diagnosis [32, 33]. For instance, sleep stage transitions have been leveraged for automated sleep staging using rule-based corrections or data-driven methods like Markov models [8, 32, 34, 35]. Clinically, disorders such as obstructive sleep apnea (OSA) show disruptions in REM-to-NREM transitions [36], while insomnia patients exhibit increased light sleep and reduced deep and REM stages. Narcolepsy, in contrast, is marked by short-latency REM and rapid transitions between wakefulness and REM sleep [26, 37]. These patterns highlight the importance of capturing sequential attributes within sleep stage sequences for accurate diagnosis.

Although sequential attributes have been used to enhance sleep staging, current methods often depend heavily on specific dataset characteristics and may struggle to capture long-term dependencies. Notably, the contextual dependencies in sleep stages resemble those in natural language. Recent advances in language models for biological data analysis underscore their potential. For example, scBERT aids in single-cell RNA-seq cell type annotation [38], Born's regression transformer supports molecular modeling [39], and protein language models predict viral evolution, protein structures, and secondary features [40–42].

Inspired by these applications, we introduce SleepGPT, a sleep language model designed for efficient sleep assessment (Fig. 1). SleepGPT undergoes self-supervised training on millions of sleep stage annotations, learning to predict the next stage from preceding ones, similar to GPT models[43, 44]. To our knowledge, this is the first study

using a language model to capture the sequential dynamics and transition patterns within sleep stage sequences. The pretrained SleepGPT model enhances both sleep staging and disorder diagnosis. Specifically, we approach PSG-based sleep staging as a speech recognition task, with SleepGPT acting as a language model to refine sleep stage predictions. Likewise, we frame sleep disorder diagnosis as a text classification task, using sleep stage sequences as the input. Extensive experiments demonstrate that SleepGPT serves as an effective plug-and-play module, consistently improving sleep staging performance and supporting sleep disorder diagnosis. These results suggest SleepGPT's potential for sleep monitoring and biomarker discovery in sleep disorders and other CNS-related diseases.

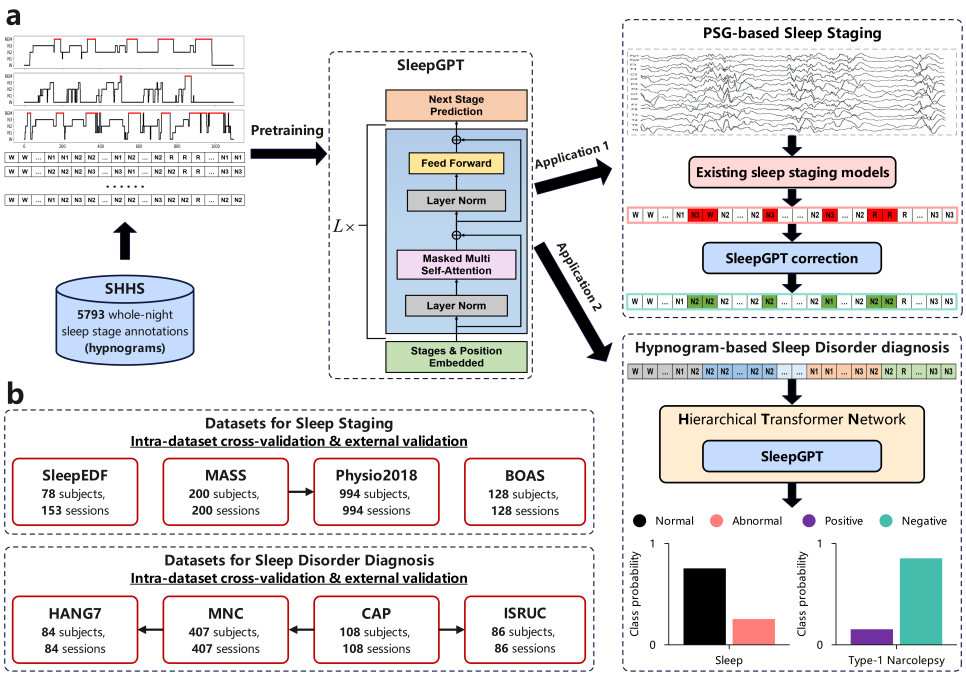

Fig. 1: **Overview of the proposed SleepGPT model and applications to sleep staging enhancement and sleep disorder diagnosis**. (**a**) The SleepGPT model is pretrained on a large sleep stage annotation dataset SHHS [45] and is used to correct the sleep stage predictions of existing sleep staging models. Moreover, a hierarchical transformer network (HTN) is employed for sleep disorder diagnosis, with SleepGPT acting as a local feature extractor. (**b**) Datasets for evaluating the proposed artificial intelligence (AI) models. For sleep staging, cross-validation of the AI models is performed on the SleepEDF [46] and MASS [47] datasets. The models trained from the MASS datasets are then externally validated on the Physio2018 [48] dataset for generalizability assessment. Furthermore, the translatability of the SleepGPT model on wearable EEG data is validated on the BOAS [49] dataset with simultaneously collected PSG and headband EEG data. For sleep disorder diagnosis, cross-validation of the AI models is performed on the CAP [50] and MNC [26] datasets. For generalizability assessment, models trained on the CAP dataset for distinguishing normal from abnormal sleep stage sequences are validated externally using the ISRUC [51], MNC, and HANG7 [52] datasets, while those trained on the MNC dataset for distinguishing Type-1 narcolepsy from other hypersomnia and healthy controls are externally validated on the HANG7 dataset.

Results

Development of SleepGPT

The SleepGPT model, built on the GPT-2 architecture [44], is trained using the SHHS-1 [45] dataset ($N = 5793$) with a next-stage prediction objective (Fig. **S4**). Specifically, each subject's sleep stage annotations from SHHS-1 are organized into 30-second sequences ranging from 360 to 1199 stages in length, with an average of 1,012 stages. In total, the dataset comprises 5,863,207 sleep stage annotations.

The pretrained SleepGPT model is then evaluated on the sleep staging task (Fig. **S5**) using the SleepEDF [46] ($N = 153$) and MASS [47] ($N = 200$) datasets. Concretely, several state-of-the-art sleep staging models are employed to predict sleep stages from PSG signals, and the predictions with and without SleepGPT corrections are compared to demonstrate the enhancement enabled by SleepGPT. Both intra-cohort cross-validation and inter-cohort validation are conducted on the SleepEDF and MASS datasets to assess the capability of the SleepGPT model in enhancing the performance of existing sleep staging models. Additionally, the Physio2018 [48] dataset ($N = 994$), along with the YASA toolbox [53] containing pretrained sleep staging models, are included to blindly (i.e., after the models are finalized and locked) assess the out-of-sample prediction performance of the models. To assess the translational utility of SleepGPT, the BOAS [49] dataset ($N = 128$), which includes simultaneous PSG and wearable EEG recordings, is employed to compare sleep staging performance between these two modalities.

The SleepGPT model is further utilized as a local feature extractor within a hierarchical transformer network (HTN) for the diagnosis of sleep disorders (Fig. **S6**). The HTN model is cross-validated on the CAP [50] ($N = 108$) dataset to distinguish normal from abnormal sleep stage sequences and blindly validated on the ISRUC [51] dataset ($N = 86$), the MNC [26] dataset ($N = 407$) and the HANG7 [52] dataset ($N = 84$). To demonstrate its performance on a potentially more challenging task, the diagnosis of Type-1 narcolepsy (vs. other hypersomnia and healthy controls) is also performed on the MNC dataset ($N = 407$) via cross-validation and blindly validated on the HANG7 dataset ($N = 84$). Finally, a visualization example of the learned global attention weights of HTN is provided to offer insights into the salient patterns captured by the model.

Enhancing sleep staging with SleepGPT

The performance of the proposed SleepGPT model is assessed in terms of accuracy, macro-averaging F1-score (MF1), and Cohen's kappa coefficient (κ) for the sleep staging task on the SleepEDF [46] and MASS [47] datasets. Among these performance metrics, accuracy indicates the proportion of correctly classified sleep stages, while MF1 ensures that performance is balanced across all classes, addressing the impact of class imbalance [54]. Cohen's Kappa (κ) adjusts for chance agreement, providing a more robust assessment of model performance, especially in multi-class classification [55]. Table 1 presents the results of the proposed methods alongside those of contemporary state-of-the-art methodologies, with the referenced ones directly sourced

from the original publications. Despite its simple design (Fig. **S7**), TinySleepNet [12] attains competitive performance across both datasets. Moreover, by integrating information from preceding EEG epochs, TinySleepNet outperforms its non-sequential version (i.e., without an LSTM module), TinySleepNet-nonseq. This improvement stems from the ability of the LSTM layer to capture the sequential attributes of EEG data, a feature leveraged by most advanced sleep staging models. Improvement is also observed between the multi-view architecture-based XSleepNet models with and without an LSTM sequential module [23], highlighting the importance of capturing sleep stage sequences to enhance sleep staging performance. The integration of SleepGPT improves the accuracy of TinySleepNet-nonseq by 2.1% and 1.5%, and that of XSleepNet-nonseq (non-sequential version of XSleepNet) by 2.5% and 1.3%, on the SleepEDF and MASS datasets, respectively. Notably, the performance of the non-sequential models on the SleepEDF dataset approaches that of their sequential counterparts, i.e., TinySleepNet and XSleepNet. This demonstrates the ability of the SleepGPT model to capture the sequential characteristics of sleep stages, compensating for the limitations of the non-sequential models in encoding the contextual information of EEG signals. Importantly, the staging accuracy of TinySleepNet with SleepGPT exhibits improvements of 0.5% and 0.6% on the SleepEDF and MASS datasets, respectively. The incorporation of SleepGPT further enhances the accuracy of XSleepNet by 0.6% and 0.4% on the SleepEDF and MASS datasets, respectively. These results underscore the effectiveness of the SleepGPT model in enhancing the performance of state-of-the-art sleep staging models, even those already equipped with sequential modules.

A visualization of the enhancement in sleep staging performance by SleepGPT is provided in Fig. **S2**, which illustrates the hypnograms of a representative subject from the SleepEDF and MASS datasets. It demonstrates that the context-aware nature of SleepGPT corrects the predicted sleep stages produced by sleep staging models, particularly for non-sequential models, resulting in a hypnogram more consistent with the ground truth.

The improvements in MF1 and κ scores further validate the robustness of the SleepGPT model in enhancing the performance of existing sleep staging models. SleepGPT significantly enhances the XSleepNet model and achieves the highest MF1 score of 78.7% and 81.5%, and κ score of 0.781 and 0.797 on the SleepEDF and MASS datasets, respectively. Additionally, the sleep stage-specific results demonstrate that the SleepGPT-powered models perform favorably on the hard-to-classify sleep stages, such as N1 and REM, which existing models are more prone to misclassifying.

Testing generalization of the sleep staging models

To evaluate the out-of-sample generalizability of SleepGPT for sleep staging, we applied the sleep staging models trained on the MASS dataset to the Physio2018 dataset [48], which contains 994 PSG recordings. We aimed to assess whether SleepGPT could similarly enhance sleep staging performance in an independent dataset, as observed in the cross-validation experiments. The MASS dataset was chosen for training due to its identical EEG channel configurations (C3-A2/C4-A1) to those in the Physio2018 dataset, ensuring consistent input features in the PSG recordings.

Table 1: Performance of state-of-the-art sleep staging methods and the proposed SleepGPT-powered models on the SleepEDF and MASS datasets. The best results are highlighted in bold, and the results with the SleepGPT-powered models are gray-shaded. **SleepGPT**: with (w) or without (w/o) SleepGPT, **ACC**: accuracy (%), **MF1**: macro-averaging F1-score (%), **kappa**: Cohen's kappa coefficient. **W**, **N1**, **N2**, **N3** and **REM**: sleep stage-specific accuracies (%). **TinySleepNet-Nonseq**: TinySleepNet without an LSTM module. **XSleepNet-Nonseq**: XSleepNet without an LSTM module.

Dataset	Method	SleepGPT	ACC	MF1	kappa	W	N1	N2	N3	REM
SleepEDF	DeepSleepNet [9]	w/o	77.8	71.8	0.700	90.8	44.8	78.5	67.9	71.3
	SleepEEGNet [11]	w/o	80.0	73.6	0.730	91.7	44.1	82.5	73.5	76.1
	AttnSleepNet [20]	w/o	81.3	75.1	0.740	92.0	42.0	85.0	**82.1**	74.2
	SeqSleepNet [14]	w/o	82.6	76.4	0.760	-	-	-	-	-
	TinySleepNet-Nonseq	w/o	80.5	71.9	0.727	93.7	29.1	87.6	70.1	74.1
		w	82.6	74.3	0.755	**94.7**	30.2	**91.4**	73.0	74.8
	TinySleepNet [12]	w/o	83.1	76.9	0.764	93.3	41.6	87.9	76.8	80.5
		w	83.6	77.6	0.772	93.6	44.2	88.1	77.1	81.1
	XSleepNet-Nonseq	w/o	80.3	72.7	0.725	90.4	31.8	88.5	70.9	76.6
		w	82.8	76.3	0.761	93.3	40.7	88.5	72.0	80.4
	XSleepNet [23]	w/o	83.7	77.9	0.774	93.6	45.0	87.7	78.6	81.4
		w	**84.3**	**78.7**	**0.781**	93.7	**47.2**	88.3	79.3	**81.5**
MASS	DeepSleepNet [9]	w/o	83.9	78.9	0.769	86.3	53.9	88.2	79.6	86.8
	IITNet* [24]	w/o	86.3	80.5	0.790	85.4	54.1	91.3	86.8	84.8
	TinySleepNet-Nonseq	w/o	81.7	74.3	0.736	86.5	29.4	89.4	79.1	84.5
		w	83.2	76.0	0.756	87.0	30.2	**91.8**	78.7	86.2
	TinySleepNet [12]	w/o	84.2	79.7	0.774	87.5	51.5	89.2	82.4	85.2
		w	84.8	80.2	0.784	87.9	53.1	89.5	**84.5**	85.6
	XSleepNet-Nonseq	w/o	82.1	75.1	0.742	86.7	32.3	89.7	79.5	83.7
		w	83.4	77.0	0.761	88.3	36.6	90.6	80.9	84.5
	XSleepNet [23]	w/o	85.3	80.8	0.791	**88.9**	53.8	90.1	83.3	86.5
		w	**85.7**	**81.5**	**0.797**	88.7	**56.1**	90.2	83.7	**87.3**

* the results are not directly comparable since they were evaluated on the SS3 subset with 62 healthy subjects in the MASS dataset.

To confirm the performance improvements were bidirectional after the generalization assessment of the MASS models was completed, we also reversed the process by applying the sleep staging models trained on the Physio2018 dataset to the MASS dataset. As an additional part of out-of-sample validation, YASA (Yet Another Spindle Algorithm) [53], an open-source sleep analysis toolbox with pretrained sleep staging models, was also applied to stage the sleep EEG data from the SleepEDF and MASS datasets. SleepGPT was assessed to determine whether it could improve the performance of this established sleep staging tool.

The results, shown in Table 2 and Fig. **S1**, indicate that SleepGPT consistently enhances the performance of YASA, TinySleepNet, and XSleepNet across all three datasets. Specifically, it improves YASA's accuracy on the SleepEDF and MASS datasets by 4.2% and 1.6%, respectively. Furthermore, integrating SleepGPT enhances the accuracy of TinySleepNet by 2.9% and 1.6%, and that of XSleepNet by 2.3% and 1.9%, during cross-dataset validation between the Physio2018 and MASS datasets. Consistent improvements in MF1, κ, and stage-specific accuracies are also observed. Notably, models trained on the larger Physio2018 dataset ($N = 994$) and tested on the

Table 2: Results of three state-of-the-art staging methods with and without SleepGPT when performing cross-dataset sleep staging on the SleepEDF, MASS, and Physio2018 datasets. **Dataset**: **Source** → **Target** indicate that the staging model is trained from the **Source** dataset and evaluated on the **Target** dataset. **SleepGPT**: with or without SleepGPT, **ACC**: accuracy (%), **MF1**: macro-averaging F1-score (%), **kappa**: Cohen's kappa coefficient. **W**, **N1**, **N2**, **N3** and **REM**: sleep stage-specific accuracies (%).

Method	Dataset	SleepGPT	ACC	MF1	kappa	W	N1	N2	N3	REM
YASA	YASA → SleepEDF	w/o	72.7	63.3	0.611	92.6	13.2	73.9	73.2	64.9
		w	76.9	68.2	0.677	89.1	20.2	81.3	81.4	76.8
	YASA → MASS	w/o	78.6	71.2	0.700	92.8	21.7	82.0	91.6	84.3
		w	80.2	73.6	0.724	95.4	26.1	82.3	94.0	86.5
	YASA → Physio2018	w/o	71.4	66.2	0.613	92.5	20.5	77.3	78.1	77.4
		w	73.5	68.6	0.640	93.6	22.8	80.7	79.3	77.7
TinySleepNet	Physio2018 → MASS	w/o	74.1	67.1	0.617	66.1	27.4	89.2	74.1	69.6
		w	77.0	71.3	0.665	66.2	39.1	90.1	82.9	71.8
	MASS → Physio2018	w/o	70.5	66.3	0.600	94.7	35.0	75.9	48.6	81.6
		w	72.1	68.0	0.621	95.1	36.4	78.2	51.3	81.8
XSleepNet	Physio2018 → MASS	w/o	74.4	68.5	0.626	64.3	39.1	88.6	83.0	61.4
		w	76.7	71.3	0.661	68.8	43.0	90.0	86.5	62.0
	MASS → Physio2018	w/o	70.3	65.8	0.594	94.2	41.8	77.1	38.8	77.0
		w	72.2	67.4	0.617	95.0	42.9	80.6	39.0	77.8

smaller MASS dataset ($N = 200$) demonstrated superior staging performance than the reverse, suggesting that the size and composition of the training dataset, particularly the distribution of sleep stages, have a considerable impact on model generalizability. A visualization example of the correction made by SleepGPT during cross-dataset validation is provided in Fig. **S3**. These findings validate the generalizability and robustness of SleepGPT, as it reliably improves the out-of-sample prediction performance of existing sleep staging models across diverse datasets.

Assessing translatability of the SleepGPT-powered models on wearable EEG data

Advances in wearable devices offer new possibilities for at-home sleep assessment, including sleep staging. To assess the translational potential of the SleepGPT-powered models, we evaluated their sleep staging performance using the BOAS dataset [49], which includes simultaneously recorded PSG and headband EEG data of 128 subjects. The TinySleepNet and XSleepNet models, with and without SleepGPT, were cross-validated on the PSG and the headband EEG data, respectively. As detailed in Table 3, the results show that both the TinySleepNet and XSleepNet models exhibit comparable performance across the PSG and headband EEG data. Importantly, the SleepGPT-powered models consistently enhanced the performance of both models, leading to improvements in accuracy, MF1, κ scores, and stage-specific accuracies across both datasets. Notably, the sleepGPT-powered XSleepNet model achieved the highest accuracy of 84.3% and 83.9% for the PSG and headband EEG data, respectively. The results underscore the robustness of the SleepGPT models in enhancing

sleep staging performance across different EEG modalities and support their potential for integration into wearable sleep monitoring applications, enabling adoption of low-burden, low-cost, longitudinal sleep monitoring without significantly compromising accuracy.

Table 3: Performance of state-of-the-art sleep staging methods and the proposed SleepGPT-powered models on the BOAS wearable dataset. The best results are highlighted in bold, while the results with the SleepGPT-powered model are gray-shaded. **SleepGPT**: with (w) or without (w/o) SleepGPT, **ACC**: accuracy (%), **MF1**: macro-averaging F1-score (%), **kappa**: Cohen's kappa coefficient. **W**, **N1**, **N2**, **N3** and **REM**: sleep stage-specific accuracies (%).

Dataset	Method	SleepGPT	ACC	MF1	kappa	W	N1	N2	N3	REM
PSG	TinySleepNet	w/o	80.9	64.1	0.660	78.0	20.8	90.8	48.0	69.4
		w	81.6	65.6	0.673	78.4	23.1	**91.2**	51.1	70.5
	XSleepNet	w/o	83.4	68.8	0.715	88.7	27.0	88.2	53.9	81.1
		w	**84.3**	**70.4**	**0.728**	**89.0**	**27.9**	89.2	**56.3**	**81.7**
Headband	TinySleepNet	w/o	80.7	62.5	0.652	78.5	18.3	91.1	39.5	68.8
		w	81.3	64.5	0.667	79.5	20.8	91.0	**48.0**	69.4
	XSleepNet	w/o	83.0	65.7	0.694	84.7	23.9	91.8	36.8	73.4
		w	**83.9**	**67.8**	**0.712**	**87.0**	**26.9**	**91.9**	43.8	**74.5**

Sleep disorder diagnosis with SleepGPT

We next assessed the capability of SleepGPT as a feature extractor for sleep disorder diagnosis compared to existing state-of-the-art methods. Performance was evaluated using balanced classification accuracy (BACC), sensitivity, and specificity across the modified CAP [50] and MNC [26] datasets. Table 4 presents the results for two empirical feature-based XGBoost classifiers—Hypnogram [53] and Hypnodensity [26]—as well as a baseline end-to-end neural network (BaseNet; Fig. **S8**), the proposed hierarchical transformer network (HTN) trained from scratch, and the HTN model incorporating pretrained SleepGPT parameters. The receiver operating characteristic (ROC) curves of each method across the CAP and MNC datasets are also shown in Fig. 2 to provide a detailed breakdown of the results (Also see Fig. **S9** for the confusion matrix of each method).

On the CAP dataset, in which the task was to distinguish between normal sleep and abnormal sleep (including sleep disorders such as insomnia, disordered breathing, narcolepsy, etc.), the XGBoost classifier trained with hypnogram features achieved a balanced accuracy of 90.91%, with a sensitivity of 95.65% and a specificity of 86.17% [1]. While end-to-end models like BaseNet can learn useful features from hypnograms, they may struggle to capture the sequential attributes of sleep stages if the model architecture is not carefully designed, as evidenced by a balanced accuracy of 84.97%, with a sensitivity of 86.96% and a specificity of 82.98%. The SleepGPT-based HTN

[1] Hypnodensity-based XGBoost results were unavailable due to the absence of hypnodensity data for the CAP dataset.

model, however, significantly improved performance in an end-to-end fashion. This highlights the HTN's capacity to grasp both localized sequential sleep transitions and broader contextual patterns in sleep hypnograms, enhancing diagnostic performance. Notably, initializing the HTN with pretrained SleepGPT parameters yielded the highest performance, achieving a balanced accuracy of 96.27%, with a sensitivity of 98.91% and a specificity of 93.62%. This pretrained HTN demonstrated a marked improvement in sensitivity over its non-pretrained counterpart, with an increase of over 10% on the CAP dataset.

Similar trends were observed in the MNC dataset, which involved distinguishing Type-1 narcolepsy (T1N) patients from non-T1N individuals (including other hypersomnia patients and healthy controls), an arguably more challenging task than distinguishing general sleep-disorder patients from health controls. The XGBoost classifier trained on hypnogram and hypnodensity features achieved balanced accuracies of 84.39% and 85.16%, respectively. Although the BaseNet model performed reasonably well in diagnosing T1N patients, achieving a balanced accuracy of 79.69%, it struggled with sensitivity, identifying only 67.07% of T1N subjects. The HTN model trained from scratch produced a balanced accuracy of 85.49%, a sensitivity of 78.05%, and a specificity of 92.92%, matching the performance of the XGBoost classifier trained on hypnodensity features. When fine-tuned with pretrained SleepGPT parameters, the HTN model further boosted performance, achieving a balanced accuracy of 92.81%, a sensitivity of 91.46%, and a specificity of 94.15%. This improvement underscores the benefits of the SleepGPT model's self-supervised pretraining on large-scale sleep stage datasets. Importantly, the pretrained SleepGPT model demonstrated the potential for fine-tuning on smaller datasets, which is crucial in sleep medicine, where data scarcity often presents a challenge.

Table 4: Performance of the proposed SleepGPT model on the sleep disorder diagnosis task. All results are reported in terms of balanced accuracy (**BACC**) (%), sensitivity (**SENS**) (%), and specificity (**SPEC**) (%). **Hypnogram** and **Hypnodensity**: two empirical feature-based XGBoost classifiers; **BaseNet**: a baseline neural network; **From scratch**: the proposed hierarchical transformer network (HTN) trained from scratch; **Pretrained**: the HTN model incorporating pretrained SleepGPT parameters.

Method	**CAP** (normal vs. abnormal)			**MNC** (T1N vs. others)		
	BACC	**SENS**	**SPEC**	**BACC**	**SENS**	**SPEC**
Hypnogram [53]	90.91	95.65	86.17	84.39	82.93	85.85
Hypnodensity [26]	-	-	-	85.16	81.71	88.62
BaseNet	84.97	86.96	82.98	79.69	67.07	92.31
From scratch	89.77	88.04	91.49	85.49	78.05	92.92
Pretrained	**96.27**	**98.91**	**93.62**	**92.81**	**91.46**	**94.15**

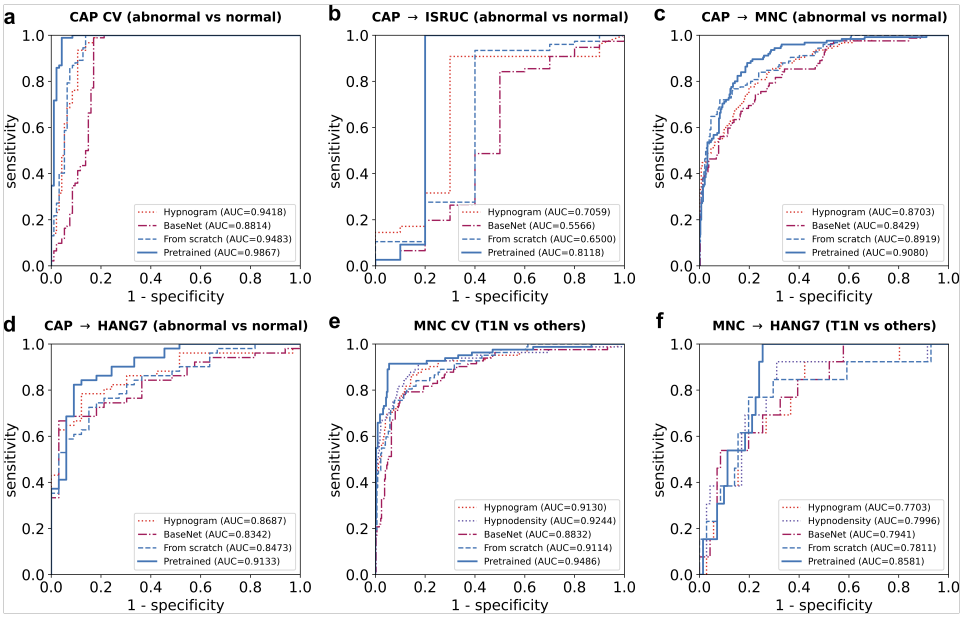

Fig. 2: Receiver operating characteristic (ROC) curves for the proposed SleepGPT-powered models and baseline methods in sleep disorder diagnosis on the CAP, ISRUC, MNC, and HANG7 datasets are presented. The area under the ROC curve (AUC) for each method is provided in the legend. The curves depict the performance of distinguishing between abnormal and normal sleep: (**a**) cross-validated (CV) on the CAP dataset, (**b**) blindly validated on the ISRUC dataset, (**c**) blindly validated on the MNC dataset, and (**d**) blindly validated on the HANG7 dataset. Additionally, they include distinguishing Type-1 narcolepsy (T1N) from others: (**e**) cross-validated (CV) on the MNC dataset, and (**f**) blindly validated on the HANG7 dataset. **Hypnogram** and **Hypnodensity**: two empirical feature-based XGBoost classifiers; **BaseNet**: a baseline neural network; **From scratch**: the proposed hierarchical transformer network (HTN) trained from scratch; **Pretrained**: the HTN model incorporating pretrained SleepGPT parameters.

Testing generalization of the sleep disorder diagnosis models

We assessed the out-of-sample generalizability of SleepGPT in diagnosing abnormal sleep by evaluating the prediction performance of the models trained from the CAP dataset on the ISRUC [51], MNC [26], and HANG7 [52] dataset. The ISRUC sleep dataset contains whole-night PSG recordings of obstructive sleep apnea (OSA) patients ($N = 76$) and healthy controls ($N = 10$). As shown in Table 5, on this unseen dataset, the pretrained HTN model achieved a balanced accuracy of 90.00%, with a sensitivity of 100% and a specificity of 80.00% when distinguishing OSA patients from healthy controls. The model also demonstrated strong performance on the MNC dataset, achieving a balanced accuracy of 83.72%, with a sensitivity of 88.00% and a specificity of 79.43% in discriminating healthy controls ($N = 282$) from patients with

T1N or other hypersomnia ($N = 125$). On the HANG7 dataset, which contains 51 patients with narcolepsy and 33 healthy controls, the HTN model achieved a balanced accuracy of 84.05%, with a sensitivity of 86.27% and a specificity of 81.82%.

For the more challenging task of differentiating T1N patients from other hypersomnia patients and healthy controls, the HTN model trained on the MNC dataset was externally validated using the HANG7 dataset. As shown in Table 6, the HTN model fine-tuned with SleepGPT parameters achieved a balanced accuracy of 84.18%, with a sensitivity of 92.31% and a specificity of 76.06%, outperforming the compared methods.

The ROC curves of each method across the generalization datasets are shown in Fig. 2. These results significantly outperform the compared methods, demonstrating the robustness and generalizability of the SleepGPT model in diagnosing sleep disorders.

Table 5: Generalization performance on the ISRUC [51], MNC [26], and HANG7 dataset [52] with model trained from the CAP [50] dataset when classifying abnormal vs normal sleep. All results are reported in terms of balanced accuracy (**BACC**) (%), sensitivity (**SENSI**) (%), and specificity (**SPECI**) (%). **Hypnogram** and **Hypnodensity**: two empirical feature-based XGBoost classifiers; **BaseNet**: a baseline neural network; **From scratch**: the proposed hierarchical transformer network (HTN) trained from scratch; **Pretrained**: the HTN model incorporating pretrained SleepGPT parameters.

Method	CAP → ISRUC			CAP → MNC			CAP → HANG7		
	BACC	**SENSI**	**SPECI**	**BACC**	**SENSI**	**SPECI**	**BACC**	**SENSI**	**SPECI**
Hypnogram [53]	80.39	90.79	70.00	77.72	82.40	73.05	79.59	80.39	78.79
BaseNet	67.11	84.21	50.00	74.92	67.07	**82.77**	72.99	82.35	63.64
From scratch	76.71	93.42	60.00	79.94	84.00	75.89	76.65	74.51	78.79
Pretrained	**90.00**	**100.00**	**80.00**	**83.72**	**88.00**	79.43	**84.05**	**86.27**	**81.82**

Visualization of the SleepGPT-powered sleep disorder diagnosis model

To provide insights into the global feature extractor of the HTN model, a visualization of the attention weights from the global feature extractor on the CAP dataset is presented in Fig. 3 for one healthy subject and five subjects with distinct sleep disorders—insomnia, narcolepsy, nocturnal frontal lobe epilepsy (NFLE), periodic leg movements (PLMs), and REM behavior disorder (RBD). By focusing on the attention weights assigned to the *CLS* token (a special token that outputs a global representation of the entire sequence, detailed in the Methods section) within the final transformer layer, we can observe how each model identifies and prioritizes salient features in the data. The results clearly demonstrate the effectiveness of the global feature extractor in capturing atypical sleep patterns associated with sleep disorders. For instance, in the case of the insomnia subject, characterized by small N3 and REM

Table 6: Generalization performance on the HANG7 dataset [52] with model trained from the MNC [26] dataset when classifying Type-1 narcolepsy (T1N) vs others. All results are reported in terms of balanced accuracy (**BACC**) (%), sensitivity (**SENSI**) (%), and specificity (**SPECI**) (%). **Hypnogram** and **Hypnodensity**: two empirical feature-based XGBoost classifiers; **BaseNet**: a baseline neural network; **From scratch**: the proposed hierarchical transformer network (HTN) trained from scratch; **Pretrained**: the HTN model incorporating pretrained SleepGPT parameters.

Method	**MNC → HANG7**		
	BACC	**SENSI**	**SPECI**
Hypnogram [53]	72.32	61.54	**83.10**
Hypnodensity [26]	75.46	69.23	81.69
BaseNet	69.12	69.23	69.01
From scratch	75.08	76.92	73.24
Pretrained	**84.18**	**92.31**	76.06

sleep ratios and direct transitions from REM sleep to wakefulness, the global feature extractor allocates more attention to the segments marked by these transitions (Fig. 3b). Likewise, the global transformer encoder of HTN highlights segments with short wake-to-REM sleep latency or abrupt transitions from wakefulness to REM sleep for the narcolepsy subject (Fig. 3c). Furthermore, the self-attention module identifies segments featuring abnormal stage transitions, short N3 and REM sleep durations, direct shifts from REM sleep to wakefulness, and frequent toggling between deep (N3) or REM sleep and wakefulness (Fig. 3d-f). Conversely, for subjects without pathologies, the attention weights are more evenly dispersed across the hypnogram segments (Fig. 3a). The above abnormal patterns in sleep stage sequences are indicative of sleep disorders and may serve as potential biomarkers for sleep disorder diagnosis.

We also show the attention weights obtained by the global feature extractor of the HTN model on the MNC dataset in Fig. 4 for a healthy control, a hypersomnia patient, and a T1N patient. The attention weights are more evenly distributed across the hypnogram segments for the healthy control, while for the narcolepsy patient, there is heightened attention to segments featuring short wake-to-REM latency or direct transitions from wakefulness to REM sleep. Since we are classifying T1N patients from other hypersomnia patients and healthy controls, the attention weights are more focused on segments similar to those of the narcolepsy patient. These segments include short wake-to-REM latency, direct transitions from wakefulness to REM sleep, and dissociated REM sleep. The self-attention module of the HTN model effectively identifies the most discriminative sleep patterns, thereby enhancing the classification performance for T1N.

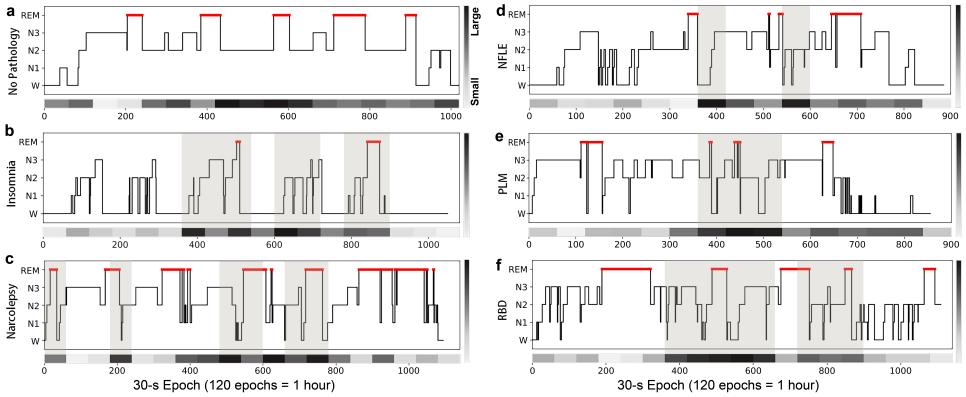

Fig. 3: Visualization of the learned attention weights of the global transformer encoder on the CAP dataset for sleep disorder diagnosis (abnormal vs. normal). (a) The attention weights are more evenly distributed among the segments of the sleep hypnogram for a healthy subject. (b) There is a small N3 and REM sleep ratio and direct transitions from REM sleep to wakefulness for the insomnia subject. (c) Segments with short REM sleep latency or direct transitions into REM sleep from wakefulness are highlighted by the global transformer encoder of the HTN for the narcolepsy subject. (d-f) Segments with abnormal stage transitions, such as short N3 and REM sleep duration, a direct transition from REM sleep to wakefulness, and frequent switches between deep sleep, i.e., N3 and REM sleep, and wakefulness, are identified by the self-attention module for the subjects with NFLE, PLMs, and RBD, respectively. The gray-shaded areas highlight the segments with abnormal sleep patterns.

Discussion

This study introduces SleepGPT, a novel sleep language model adapted from the GPT architecture and trained in a self-supervised manner on a comprehensive sleep stage dataset. This effort parallels recent advances in biological sequence modeling, as seen in studies using transformer-based models for protein folding [40–42] and cell type annotation [38], and expands these applications into the domain of sleep medicine. SleepGPT offers several advantages for efficient sleep assessment: it consistently enhances existing sleep staging methods, effectively captures sleep stage transition dynamics, and integrates as a feature extractor within hierarchical transformer networks to improve sleep disorder diagnosis. Additionally, it identifies interpretable abnormal sleep patterns, potentially providing mechanistic insights into sleep disorders. Taken together, these findings highlight SleepGPT's potential as a scalable, clinically translatable artificial intelligence (AI)-powered solution for automated sleep assessment.

The integration of SleepGPT consistently enhances sleep staging performance, although the degree of improvement varies across models and datasets. Notably, models lacking memory mechanisms, which fail to utilize contextual information from

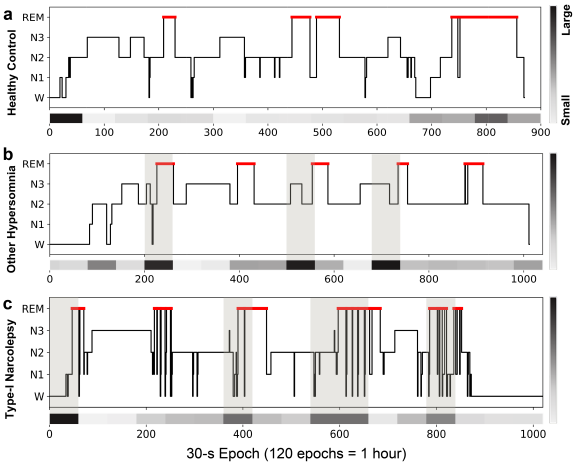

Fig. 4: Visualization of the learned attention weights of the global transformer encoder on the MNC dataset for sleep disorder diagnosis (Type-1 narcolepsy vs. (other hypersomnia + healthy control)). (a) The attention weights are more evenly distributed among the segments of the sleep hypnogram for a healthy subject. (b) The attention weights for a hypersomnia subject focus on deep sleep stages (N3) in the late half of sleep. (c) Segments with abnormal stage transitions, such as short REM sleep latency, and dissociated REM sleep, are identified by the self-attention module for the narcolepsy subject. The gray-shaded areas highlight the segments with the most discriminative sleep patterns.

physiological signals and sleep stages, exhibit more substantial improvement. In comparison, models with sequential components like LSTMs show more moderate, likely due to their inherent ability to capture temporal dependencies in sleep-related data, but still consistent gains. SleepGPT contributes to the sleep staging task by leveraging its understanding of natural sequential dependencies between sleep stages, learned through pretraining on large hypnogram datasets. This contextual correction layer enables SleepGPT to refine predictions from existing staging models, adjusting misclassifications based on surrounding stages. For example, if a model misclassifies a REM stage as wakefulness, SleepGPT can reclassify it by recognizing that the sequence context supports a REM classification. Additionally, SleepGPT's attention mechanisms focus on critical stage transitions, identifying inconsistent patterns, such as abrupt shifts from deep sleep (N3) directly to wakefulness, that may signal staging errors. The benefits of SleepGPT are particularly pronounced in blinded evaluations on independent datasets, where data heterogeneity often hinders the generalizability of sleep staging models. By effectively capturing the intrinsic sequential patterns of sleep stages, SleepGPT enhances model generalizability across diverse datasets.

The relatively low staging accuracy for N1 sleep observed across all models is consistent with findings in the literature [53]. N1 sleep, being a transitional and highly variable stage, is notably challenging to classify accurately due to its subtle features and overlap with both wakefulness and N2 stages. Studies have reported that human

scorers also struggle with N1 consistency, often showing significant inter-rater variability [56]. This inherent ambiguity makes N1 classification particularly difficult for automated models, which rely on training data labeled by human experts. Despite these challenges, the SleepGPT models achieves consistent improvements across sleep stages. We anticipate even greater performance improvements as the volume of training data for SleepGPT increases.

In contrast to traditional sleep disorder diagnosis approaches that rely on handcrafted features derived from hypnogram analysis (e.g., stage latencies, durations, and transitions), our proposed hierarchical transformer network (HTN) directly models and categorizes sleep stage sequences in an end-to-end fashion. This hierarchical architecture allows the model to capture both local and global contextual information within sleep hypnograms, enabling more accurate classification of sleep stage sequences, especially those with varying lengths, compared to the baseline model. By initializing the HTN's local feature extractor with pretrained SleepGPT parameters and subsequently fine-tuning it, we can extract highly informative features for sleep disorder diagnosis. Notably, the HTN effectively addresses the challenge of subject-level sleep disorder diagnosis in weakly supervised learning scenarios, where only session-level labels are available and sleep disorder symptoms may not be consistently present throughout the entire sleep session [37]. Leveraging the self-attention mechanism in its global transformer module, the HTN achieves accurate subject-level classification and identifies potential biomarkers indicative of specific sleep disorders. These include, but are not limited to: (1) reduced REM sleep ratio and frequent REM sleep-to-wake transitions, suggesting insomnia; (2) shortened REM sleep latency and direct transitions from wakefulness to REM sleep, characteristic of narcolepsy; and (3) brief REM sleep duration and frequent shifts between deep sleep (N3 and REM) and wakefulness, signaling other sleep abnormalities. These insights hold promise for developing novel diagnostic methodologies. Moreover, the HTN's adaptability allows for its potential application to other sleep-related brain disorders, such as depression, anxiety, dementia, and Parkinson's disease [1, 2], e.g., by replacing sleep disorder labels with corresponding condition labels. This flexibility is particularly valuable given the frequent comorbidity of sleep disorders with other neurological or psychiatric conditions. Additionally, information in sleep stage sequences or hypnograms is much simpler than that in PSG data, which is often noisy and complex. Diagnosis from sleep stage sequences has the advantage of alleviating inter-subject or inter-cohort heterogeneity. However, PSG data may also be fused with the sequential features of sleep stages to further improve the performance of sleep disorder diagnosis.

On the CAP dataset, the pretrained HTN model effectively identifies nearly all abnormal subjects, with only 1.09% of abnormal subjects being erroneously classified as normal. This high sensitivity is particularly crucial in sleep disorder diagnosis, where false positives (misclassifying normal as abnormal) are generally preferable to false negatives, as they lead to further medical investigation, while undetected abnormalities could have serious consequences. Moreover, previous literature indicates that sensitivity and specificity for T1N are 75-90% and 90-98%, respectively [26, 57–59]. The performance achieved by the pretrained HTN model on the MNC dataset is comparable to or exceeds these reported values. These results suggest that analyzing a

single-night PSG can be as effective as the PSG-MSLT gold standard, which involves a 24-hour procedure and is expensive. Consequently, our model offers a promising and cost-effective alternative as a screening tool for T1N, potentially reducing the need for MSLT by reliably identifying individuals likely to have T1N. This approach could streamline the pathway for patients requiring further diagnostic evaluation. Furthermore, Stephansen et al. [26] reported a sensitivity of 91% and specificity of 96% for diagnosing T1N on the test data from the full MNC dataset, with replication set sensitivity and specificity at 93% and 91%. However, while their study utilized the entire MNC dataset from nine cohorts, our study analyzed only a subset of the MNC dataset (SSC, DHC, and CNC) due to missing sleep staging labels in the other six cohorts.

The strong performance of SleepGPT-based models in sleep staging with wearable EEG and in disorder identification opens exciting possibilities for real-time, at-home sleep monitoring. Wearable devices powered by SleepGPT could provide immediate feedback on sleep quality, identify potential sleep disturbances as they occur, and offer personalized advice for improving sleep. Such real-time monitoring could be invaluable for individuals with chronic sleep disorders, allowing for timely interventions and adjustments to treatment plans. Furthermore, continuous sleep monitoring could facilitate early detection of sleep problems, potentially preventing them from escalating into more serious health issues.

While our experiments demonstrate the effectiveness of SleepGPT, it is important to acknowledge certain limitations. The SHHS dataset used for pretraining SleepGPT is relatively small compared to the massive text corpora used to train traditional GPT models. Expanding the pretraining dataset could further enhance model performance. Future research should also explore fine-grained sleep disorder classification by leveraging larger and more diverse datasets to gain deeper insights into specific sleep stage patterns and transitions associated with particular disorders. Furthermore, while hypnograms provide valuable information on sleep macrostructure, they have limited ability to capture microstructural events like sleep spindles and K-complexes [28]. Integrating additional data modalities, such as high-resolution EEG or other physiological signals, may enhance the accuracy and granularity of sleep disorder diagnosis. Finally, further validation of SleepGPT across a broader range of sleep datasets is essential to corroborate its generalizability and robustness before it could be translated for use in clinical practice.

In summary, this study presents SleepGPT, a novel sleep language model based on the GPT architecture, for efficient sleep assessment. Extensive evaluation across multiple publicly available sleep datasets as well as fully blinded replications on independent datasets demonstrate that SleepGPT significantly improves sleep staging accuracy and exhibits promising efficacy in classifying abnormal sleep patterns. This novel approach for modeling sleep architecture opens new avenues for sleep data analysis, providing a path towards automated diagnosis and personalized treatment of sleep disorders.

Methods

Datasets

This study utilizes three distinct types of datasets: sleep datasets for SleepGPT pre-training, sleep staging, and sleep disorder diagnosis. A total of six publicly accessible sleep datasets are employed in the experiments. Table 7 provides a comprehensive summary of these datasets, and further details are provided below.

Table 7: Overview of the involved sleep datasets. **BMI** = body mass index, **AHI** = apnea-hypopnea index.

Dataset	Subjects / Sessions	Recording Duration	Age (AVG±STD)	Sex (% Male)	BMI (AVG±STD)	AHI (AVG±STD)	Health Conditions
SHHS [45, 60]	5793 / 5793	Overnight	63.1 ± 11.2	47.6	28.2 ± 5.1	17.9 ± 16.1	Sleep-disordered breathing, heart diseases, and others
SleepEDF [46, 61]	78 / 153	Around 9h	59 ± 22.1	46.4	-	-	Healthy subjects
MASS [47]	200 / 200	Overnight	40.6 ± 19.4	48.5	-	≤ 20	Healthy subjects
Physio2018 [48, 61]	994 / 994	7.7h	55 ± 14.3	67.0	33 ± 7.8	19 ± 14.6	Sleep disorders, healthy subjects
BOAS [49]	128 / 128	Overnight	42.2 ± 19.0	40.6	23.8 ± 3.2	-	Healthy subjects
CAP [50, 61]	108 / 108	8–10h	45.2 ± 19.7	61.1	-	-	Sleep disorders (n=92), healthy subjects (n=16)
ISRUC [51]	86 / 86	Overnight	49.7 ± 15.7	59.4	-	-	Sleep apnea (n=76), healthy subjects (n=10)
MNC-CNC [26, 60]	77 / 77	Overnight	28.5 ± 16.9	51.3	23.2 ± 11.5	5.34 ± 1.51	Type-1 narcolepsy (n=54), healthy subjects (n=23)
MNC-DHC [26, 60]	79 / 79	Overnight	33.4 ± 14.8	50.0	24.8 ± 4.9	-	Type-1 narcolepsy (n=21), hypersomnia (n=38), healthy subjects (n=20)
MNC-SSC [26, 60]	251 / 251	Overnight	45.4 ± 13.8	59.4	23.9 ± 6.5	13.7 ± 0.7	Type-1 narcolepsy (n=7), hypersomnia (n=5), healthy subjects (n=239)
HANG7 [52]	84 / 84	8h	24.5 ± 9.6	47.6	22.72 ± 3.65	-	Type-1 narcolepsy (n=13), other narcolepsy (n=38), healthy subjects (n=33)

Datasets for SleepGPT Pretraining

Training a transformer-based language model typically requires an extensive text corpus, often encompassing millions or even billions of web pages or documents. However, a sleep dataset of comparable scale, complete with sleep stage annotations, is not available. Fortunately, the increasing advancements in sleep medicine research, coupled with the research community's open data policy, have produced several publicly accessible sleep datasets with sleep stage annotations. The Sleep Heart Health Study (SHHS) database, a multi-center cohort study examining the cardiovascular and other consequences of sleep-disordered breathing [45, 60], is a noteworthy example. This database comprises two rounds of PSG records: Visit 1 (SHHS-1) and Visit 2 (SHHS-2). In this work, we use the SHHS-1 cohort, which encompasses 5,793 subjects aged between 39 and 90 years, to train the SleepGPT model. Notably, to ensure alignment with the AASM scoring standard [3], we merge the N3 and N4 stages into the N3 stage while discarding the MOVEMENT and UNKNOWN epochs, as the SHHS-1 database was manually scored following the R&K guidelines [62].

Datasets for Sleep Staging

The proposed SleepGPT model's performance on sleep staging tasks is evaluated with two widely used sleep datasets, namely, the Sleep-EDF and the Montreal Archive of Sleep Studies (MASS). The Physio2018 dataset serves as a benchmark to evaluate the generalization performance of SleepGPT in enhancing sleep staging. Table 8 provides a detailed summary of these datasets.

Table 8: Number of subjects, EEG channels, and sleep stage distribution of the sleep staging datasets

Datasets	Subjects	EEG channel	W	N1	N2	N3	REM	Total
SleepEDF	78	Fpz-Cz	69824	21522	69132	13039	25835	199352
MASS	200	C4-A1/C3-A2	31184	19359	107930	30383	40184	229040
Physio2018	994	C3-A2	157945	136978	377870	102592	116877	892262
BOAS	128	C4/AF7	19137	4462	72181	5225	18754	120095

SleepEDF Dataset: We utilize the 2018 version of the SleepEDF Expanded dataset [46, 61]. This collection comprises data from 78 healthy Caucasian subjects aged 25 to 101 years. Each subject contributed two consecutive day-night PSG recordings, except for subjects 13, 36, and 52, where one recording was lost due to device failure. Consequently, the dataset contains 153 overnight recordings. Sleep experts manually scored the epochs based on the R&K standard [62], assigning each 30-second PSG epoch to one of eight categories: {W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN}. To align with convention, the N3 and N4 stages were merged into the N3 stage, while the MOVEMENT and UNKNOWN epochs were excluded. Notably,

the SleepEDF-20 dataset was not assessed because it is a subset of this particular version of SleepEDF.

MASS Dataset: Derived from different hospital-based sleep laboratories, the MASS database comprises whole-night recordings from 200 subjects (97 males and 103 females) aged 18 to 76 years [47]. The annotation process involved sleep experts adhering to either the AASM standard [3] (for the SS1 and SS3 subsets) or the R&K standard [62] (for the SS2, SS4, and SS5 subsets). In alignment with the previously mentioned datasets, we harmonized the R&K annotations with the five sleep stages {W, N1, N2, N3, REM} according to the AASM standard. Epochs initially spanning 20 seconds were extended to 30 seconds by incorporating the 5-second segments before and after them.

Physio2018 Dataset: The PhysioNet 2018 Challenge dataset, also known as the Physio2018 dataset, comprises 1,985 polysomnographic recordings provided by the Computational Clinical Neurophysiology Laboratory (CCNL) and the Clinical Data Animation Center (CDAC) at Massachusetts General Hospital (MGH). This dataset was used in the 2018 PhysioNet Challenge [48, 61] to detect sleep arousals. We used the training set for our experiments, which included 944 subjects aged 18 to 90. Sleep experts manually scored the recordings according to the American Academy of Sleep Medicine (AASM) guidelines [3], annotating five sleep stages: W, N1, N2, N3, and REM. This dataset is employed to blindly validate the sleep staging model derived from the MASS dataset.

BOAS Dataset: The Bitbrain Open Access Sleep (BOAS) dataset serves to bridge the gap between gold-standard clinical sleep monitoring and emerging wearable EEG technologies [49]. This dataset comprises data from 128 nights, during which healthy participants were simultaneously monitored using both a Brain Quick Plus Evolution PSG system by Micromed and a Bitbrain wearable EEG headband. The Micromed PSG system collected EEG signals from electrodes placed at F3, F4, C3, C4, O1, and O2, following the international 10-20 system. In contrast, the Bitbrain headband recorded EEG signals from the frontal AF7 and AF8 electrode sites. Both systems utilized a sampling rate of 256 Hz. Sleep staging was independently annotated by three expert scorers following the American Academy of Sleep Medicine (AASM) criteria [3], with a consensus label established by a fourth expert. The BOAS dataset enables the evaluation of the SleepGPT-based model's ability to achieve sleep staging accuracy comparable to PSG using wearable EEG data.

Datasets for Sleep Disorder Diagnosis

Two sleep-disorder-related datasets were identified for the sleep disorder diagnosis analysis, including the CAP Sleep Database and the Mignot Nature Communications (MNC) dataset, which offer both sleep stage annotations and sleep disorder labels.

CAP Dataset: The CAP Sleep Database is a collection of 108 polysomnographic recordings contributed by the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy [50, 61]. It contains data from seven groups of patients with distinct sleep disorders, as well as a healthy control group without any medical, neurological, or psychiatric conditions. The details of these groups are summarized in Table 9. Well-trained neurologists who are sleep experts manually scored the sleep recordings based

on the R&K rules, categorizing the epochs into sleep stages 1-4, wake, REM sleep, and movement artifacts. To adhere to the AASM standard, we harmonized the R&K annotations, obtaining five sleep stages: {W, N1, N2, N3, REM}. Notably, the CAP Sleep Database has been widely used in numerous studies focusing on sleep disorder diagnosis.

Table 9: Sleep disorder groups in the CAP Sleep Database.

Sleep disorder	No. of subjects
Bruxism (BRUX)	2
Sleep-disordered breathing (SBD)	4
Insomnia (INS)	9
Narcolepsy (NARCO)	5
Nocturnal frontal lobe epilepsy (NFLE)	40
Periodic leg movements (PLMs)	10
REM behavior disorder (RBD)	22
No pathology (N)	16

Note that the CAP Sleep Database is inherently imbalanced. Several groups contain an extremely limited number of subjects. For instance, the BRUX and SBD groups consisted of only 2 and 3 subjects, respectively. To facilitate subject-level diagnoses, we conducted binary classification experiments on the CAP Sleep Database, specifically distinguishing abnormal from normal cases. Furthermore, after incorporating the first-session data of 78 subjects (excluding subjects 36 and 52, who had only one session available) from the SleepEDF Expanded database into the normal group, the comprehensive dataset included 186 subjects, 94 of whom were labeled as normal and 92 of whom were labeled as abnormal.

ISRUC Dataset: The ISRUC-Sleep dataset [51] contains full-night PSG recordings, each approximately eight hours in duration, collected at the Sleep Medicine Centre of Coimbra University Hospital (CHUC) between 2009 and 2013. Data were acquired non-invasively using a SomnoStar Pro multi-channel system, with sensors placed according to the international 10–20 standard. The dataset includes recordings from both healthy adults and subjects with sleep disorders under medication, divided into three groups: 1) 100 subjects with one session each; 2) 8 subjects with two sessions for longitudinal studies; 3) 10 healthy subjects with one session, used for comparison with sleep disorder patients. Among the dataset, 76 obstructive sleep apnea (OSA) patients and 10 healthy controls are included to blindly validate the sleep disorder diagnosis model derived from the CAP dataset.

MNC Dataset: The Mignot Nature Communications (MNC) dataset comprises raw polysomnography data collected from an automated sleep staging project utilizing neural networks [26, 60]. It encompasses data from ten distinct cohorts recorded at twelve sleep centers across three continents: the patient-based Stanford Sleep Cohort (SSC), the population-based Wisconsin Sleep Cohort (WSC), the patient-based Inter-scorer Reliability Cohort (IS-RC), the Jazz Clinical Trial Sample (JCTS), the patient-based Korean Hypersomnia Cohort (KHC), the patient-based Austrian Hypersomnia

Cohort (AHC), the patient-based Italian Hypersomnia Cohort (IHC), the patient-based Danish Hypersomnia Cohort (DHC), the patient-based French Hypersomnia Cohort (FHC), and the patient-based Chinese Narcolepsy Cohort (CNC). The study received approval from institutional review boards, and informed consent was obtained from all participants. Trained sleep-scoring technicians manually annotated all sleep studies according to the AASM Scoring Manual [3]. Additionally, the subjects were divided into the Type-1 narcolepsy (T1N; with either low CSF hypocretin-1 levels or clear cataplexy), other hypersomnia (OHS), or non-narcolepsy control (NNC) group based on multiple sleep latency test (MSLT) results, cataplexy symptoms, and human leukocyte antigen (HLA) results (if available). Further information about the MNC dataset can be found in [26].

However, as diagnostic results were only available for the SSC, DHC, and CNC cohorts, we exclusively utilized these three cohorts for the sleep disorder diagnosis task. The dataset compiled from these cohorts encompassed 407 subjects, among whom 82 were in the T1N group, 43 were in the OHS group, and 282 were in the NNC group (see Table 10). To facilitate comparison with previous research, we adhered to the methodology outlined in [26] and conducted binary classification experiments on the MNC dataset to discriminate T1N subjects from all other subjects, i.e., OHS and NNC subjects. This dataset is also employed to blindly assess the generalizability of the sleep disorder diagnosis model derived from the CAP dataset.

Table 10: Sleep disorder groups in the MNC dataset.

Sleep disorder	**No. of subjects**			
	CNC	**DHC**	**SSC**	**Total**
Type-1 narcolepsy (T1N)	54	21	7	82
Other hypersomnia (OHS)	0	38	5	43
Non-narcolepsy control (NNC)	23	20	239	282
Total	77	79	251	407

HANG7 Dataset: In the HANG7 dataset [52], 84 participants aged 11 to 57 years (mean age: 24.5 ± 9.6), including 44 females and 40 males, were recruited to collect polysomnography recordings at the Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine. The study was conducted at Zhejiang University with Institutional Review Board approval, and written consent was obtained from all participants or their caregivers. PSG recordings were collected following the AASM sleep standards [3] and manually scored by experienced sleep technicians. Each participant's PSG recording covered one full night, from approximately 21:00 to 5:00 the next morning, totaling about 8 hours. Of the 84 participants, 13 were diagnosed with T1N (with clear cataplexy), 38 were other narcolepsy, and 33 were healthy controls. The dataset was used to blindly validate the generalizability of the sleep disorder diagnosis models derived from the CAP (abnormal vs. normal) and MNC (T1N vs. others) datasets.

Problem Formulation

Sleep Staging as Speech Recognition

A general speech recognition system consists of an acoustic model and a language model. The acoustic model $\mathcal{M}_1$ predicts the most likely word sequence given speech audio features X via

$$P(Y;\mathcal{M}_1) = P(Y|X;\mathcal{M}_1), \tag{1}$$

while the language model $\mathcal{M}_2$ is built based on a large-scale text corpus to capture the sequential characteristics of the word sequence via

$$P(Y;\mathcal{M}_2) = P(y_1, y_2, \cdots, y_T) = \prod_{t=1}^{T} P(y_t|y_1, y_2, \cdots, y_{t-1}). \tag{2}$$

The trained language model is then used to rectify the word sequence predicted by the acoustic model to improve the speech recognition performance as follows [63–65]:

$$\begin{aligned} Y^\star &= \arg\max_Y P(Y;\mathcal{M}_1,\mathcal{M}_2) \\ &= \arg\max_Y P(Y|X;\mathcal{M}_1)P(Y;\mathcal{M}_2). \end{aligned} \tag{3}$$

In the automated sleep staging process, features are extracted from a sleep data epoch, and a sleep staging model is then applied to predict the most likely sleep stage for that epoch. The model, whether it is a traditional machine learning or deep learning model, is similar to the acoustic model in a speech recognition system. It is designed and trained on a large-scale sleep dataset to achieve satisfactory prediction accuracy on unseen sleep epochs. However, due to various factors such as architectural limitations, training inadequacies, and data scarcity, imperfections in the PSG-based sleep staging model may occur, leading to inaccurate predictions of sleep stages. The inter-scorer reliability for sleep stage scoring is reported to be 82.6% on average, which closely aligns with that achieved by machine learning-based automated staging systems [26]. Given the analogous nature of text and sleep stage sequences, emulating the speech recognition system paradigm may improve sleep staging performance. This involves training a sleep language model that can capture the inherent sequential characteristics of sleep stages to assume a role akin to that of a natural language model (see Fig. **S5**).

Sleep Disorder Diagnosis as Text Classification

It is believed that sleep architecture, i.e., the distribution of sleep stages, is strongly related to the quality of sleep. Certain sleep disorders exhibit disrupted sleep architecture and atypical sleep stage transitions. Examples include shortened deep sleep stages in cases of insomnia, and short REM sleep latency or immediate transitions from wakefulness to REM sleep in narcolepsy. An overnight sleep stage sequence has the form $\{\mathrm{W}, \cdots, \mathrm{N1}, \cdots, \mathrm{N2}, \cdots, \mathrm{N3}, \cdots, \mathrm{REM}, \cdots\}$. Such sequences are plotted as hypnograms in Fig. 1 (top-left). They are akin to concise news pieces or textual documents with a vocabulary of only five words. In turn, the sleep stage sequence encodes the

sequential attributes and transitional patterns characterizing the progression between sleep stages. The integration of a sequential model enables these characteristics to be identified and described.

Hence, a viable approach is to treat sleep disorder diagnosis as a long text classification task. Here, the trained sleep language model is harnessed to extract features from the sleep stage sequences (hypnograms), enabling the subsequent classification of these sequences as associated with a sleep disorder or not. This is an innovative method for sleep disorder diagnosis using only the sleep stages obtained from either human experts or automated sleep staging models.

The SleepGPT model

Recently, the success of transformer-based language models (LMs), including bidirectional encoder representations from transformers (BERT) and generative pretrained transformers (GPT), has been noted across various natural language processing (NLP) tasks, such as machine translation, question-answering, and text generation [43, 44, 66]. BERT, which functions as a bidirectional LM, is trained to predict masked words based on neighboring words and to perform next-sentence prediction. Conversely, GPT operates as a decoder-only transformer LM, autoregressively predicting the next character or word based on preceding ones. Both models undergo self-supervised training on extensive text corpora. The success of ChatGPT has illustrated its proficient ability to capture the inherent sequential attributes of natural language. This ability has been extensively utilized to improve performance in speech recognition and diverse NLP tasks, such as text classification.

Transformers consist of multiple transformer blocks, typically including a multi-head self-attention layer, a feed-forward layer, and residual and normalization layers [67]. GPT employs causal attention in its self-attention layer to ensure exclusive attention to preceding words. This concept has been applied in sequential sleep staging models that rely on prior sleep epochs for current sleep stage predictions. It has proven to be especially significant in online sleep staging systems, where only preceding epochs are accessible during the current prediction task. Hence, GPT is selected to capture the sequential traits of sleep stages. The SleepGPT model adopts the architecture of the GPT-2 language model developed by OpenAI [43, 44]. As illustrated in Fig. **S4**, the main part of the model is composed of a series of n transformer decoder blocks, marked by the masked multi-head self-attention layers. Token embedding and position embedding layers are employed to map the sleep stage tokens into a vector space and to infuse positional information into each token's sequence placement, respectively. The final transformer block's output is input to a linear layer to obtain the probability distribution of the next sleep stage.

The model undergoes self-supervised training through the autoregressive forecast of the most likely sleep stage based on preceding sleep stages. For each subject or sleep session, the overnight sleep stage annotations are structured as a long sequence $\mathcal{U} = \{u_1, ..., u_N\}$, where $u_i \in \{0, 1, 2, 3, 4\}$ represents the five sleep stages $\{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ as integers ranging from 0 to 4. Training samples are derived from the sequence using a sliding window of dimensions K, passing over the sequence with a stride of 1, where K signifies the sample sequence length. Consequently, the

overnight stage sequence yields $N-K+1$ overlapping instances. The $(i-1)$th instance of the K sleep stages, denoted as $\mathbf{U}_{i-1} = \{u_{i-K}, ..., u_{i-1}\}$, constitutes the input and is fed into the model for predicting the target; the target is the ith instance $\mathbf{U}_i = \{u_{i-K+1}, ..., u_i\}$:

$$h_0 = \mathbf{U}_{i-1}\mathbf{W}_e + \mathbf{W}_p \tag{4}$$

$$h_l = \text{transformer\_block}(h_{l-1}), \ \forall l \in [1, L] \tag{5}$$

$$P(\mathbf{U}_i) = \text{softmax}(h_n \mathbf{W}_s^T) \tag{6}$$

where L is the number of transformer layers, $\mathbf{W}_e$ is the token embedding matrix, $\mathbf{W}_p$ is the position embedding matrix, and $\mathbf{W}_s$ is the weight of the classification head. The model is trained with the objective of minimizing the cross-entropy loss, thereby reducing the disparity between the predicted sleep stage and the actual ground truth:

$$\mathcal{L}(\mathcal{U}) = -\sum_{i=1} \log P(\mathbf{U}_i | \mathbf{U}_{i-1}; \Theta), \tag{7}$$

where Θ is the parameter of the GPT sequential model.

Sleep Staging with SleepGPT

Automated sleep staging is an active research topic within the realm of sleep medicine; it aims to automate the prediction of sleep stages for individual sleep PSG data epochs, typically spanning 30 seconds. Several excellent deep learning models have been proposed for this purpose. Most of them share a common architecture, comprising a CNN to extract intra-epoch features and an RNN to incorporate the contextual information in adjacent PSG data epochs[9, 11–14, 23]. Among cutting-edge sleep staging models, XSleepNet[23] employs two network streams to learn from multi-view inputs (e.g., both raw signals and time-frequency images) for sleep staging. By adapting the contributions of the two views on time to perform joint feature learning during training, XSleepNet outperforms the single-view baselines and multi-view baselines with a simple fusion strategy. However, while contextual information within PSG signals is incorporated into the above deep learning models, they neglect the inherent sequential traits and transition patterns within sleep stages.

To use the trained SleepGPT model to improve sleep staging performance, we treat the sleep staging task as a speech recognition task and follow the pipeline in Fig. **S5**. The sleep staging model (SSM) is used to predict the most likely sleep stage given a PSG data epoch or preceding epochs, while the SleepGPT model is used to rectify the predicted sleep stage given past stages. That is, the output logits of the SSM $P_{\text{SSM}}(y)$ and those of the sleep language model (SLM) $P_{\text{SLM}}(y)$ are weighted by a factor α to obtain the final sleep stage prediction:

$$P(y) = \alpha P_{\text{SSM}}(y|\mathbf{x}) + (1-\alpha) P_{\text{SLM}}(y|\mathbf{y}_-), \tag{8}$$

where $\mathbf{x}$ is the set of input data epochs, including the current and preceding epochs depending on whether a memory staging model is used, and $\mathbf{y}_-$ is the set of preceding

sleep stages (context). Notably, the hyperparameter α governs the relative influences of these two models, ultimately steering the sleep stage prediction toward the highest probability outcome.

Sleep Disorder Diagnosis with SleepGPT

Sleep disorder diagnosis involves discerning the presence and specific type of sleep disorder in an individual. This task is typically conducted either by sleep specialists or by ML models that leverage PSG data for automated assessment. ML-based approaches often include PSG data feature extraction, followed by the classification of sleep disorders. However, the high dimensionality of PSG data, coupled with a lack of labeled instances, makes it challenging to train an ML model with robust generalizability. The situation is even worse in the case of deep learning models, which commonly demand an extensive volume of labeled data to achieve optimal model performance.

Pretrained language models, such as GPTs trained on expansive text corpora, can capture the intrinsic sequential characteristics of natural language. These pretrained sleep language models can function as feature extractors, processing text sequences to subsequently enable classification. Notably, these pretrained models can be further fine-tuned using limited-scale datasets, significantly enhancing the classification performance. This strategy, referred to as transfer learning, effectively addresses the challenge of limited sample size, and it has a proven track record of success across numerous NLP applications.

Drawing inspiration from this pretraining and fine-tuning paradigm, we leverage the pretrained SleepGPT model as a feature extractor. We replace the subsequent stage prediction layer with a classifier and perform comprehensive fine-tuning to facilitate sleep disorder classification. Nevertheless, a challenge arises for the SleepGPT model in handling long sequences: the sequence length limitation. A whole-night (8-hour) sleep stage sequence comprises 960 time steps, significantly exceeding the sequence length restriction of the SleepGPT model. We address this by segmenting the overnight sequence into shorter sections. Initially, the SleepGPT model is employed to extract local context features from these short sleep stage segments. Subsequently, another sequential model (a transformer encoder) is used to capture the global contextual features from the SleepGPT output [67]. Finally, the resulting global context features are fed into a classification head to predict sleep disorder labels. The configuration of the hierarchical transformer network (HTN) is depicted in Fig. **S6**. This architecture, which is tailored for lengthy sequence classification, is borrowed from the hierarchical attention network (HAN) and hierarchical transformers utilized for long text classification [68, 69].

Additionally, to facilitate mini-batch training, we pad the short sleep stage sequences within a batch to achieve a uniform length. Subsequently, a mask matrix is employed to exclude the padded values during the computation of attention weights and the loss calculation. The loss function involves cross-entropy loss, which compares the predicted sleep disorder labels with the ground truth:

$$\mathcal{U} = \{\mathbf{U}_i\},\ \forall i \in [1, N] \tag{9}$$

$$\mathbf{z}_i = \text{SleepGPT}(\mathbf{U}_i) \tag{10}$$

$$\mathbf{V} = \text{transformer\_encoder}(\mathbf{Z}) \tag{11}$$

$$P(\mathcal{U}) = \text{softmax}(\mathbf{W}_c\mathbf{v}_0 + b_c) \tag{12}$$

$$\mathcal{L}(\mathcal{U}) = -\sum_{i=1} \log P(\mathcal{U}; \Phi), \tag{13}$$

where $\mathbf{U}_i$ is the ith segment, and N represents the number of segments within a sleep stage sequence. $\mathbf{z}_i$ denotes the ith local feature vector generated by SleepGPT. Notably, $\mathbf{Z} = \{\mathbf{z}_0, \cdots, \mathbf{z}_N\}$ and $\mathbf{V} = \{\mathbf{v}_0, \cdots, \mathbf{v}_N\}$ denote the input and output sequences of the transformer encoder, respectively. Here, $\mathbf{z}_0$ corresponds to the CLS token, while $\mathbf{v}_0$ represents the global feature vector at the CLS token's output position. The CLS token in transformers is a special token added at the beginning of the input sequence, and its final output serves as a global representation of the entire sequence for tasks like classification [66, 70]. The parameters of the classification head are denoted as $\mathbf{W}_c$ and b_c, and Φ includes the trainable parameters of the entire model.

Baseline Models

To evaluate the efficacy of the SleepGPT model, we utilized two baseline models for sleep staging tasks: TinySleepNet [12] and XSleepNet [23]. TinySleepNet is a lightweight architecture integrating a convolutional neural network (CNN) and long short-term memory (LSTM) components. It extracts local EEG signal features via the CNN, while the LSTM captures sequential dependencies between sleep epochs (refer to Fig. **S7**). Despite its simplicity, TinySleepNet exhibits competitive performance across various publicly available sleep datasets. XSleepNet, on the other hand, is a state-of-the-art sleep staging model that harnesses multi-view inputs, including raw signals and time-frequency images, to increase sleep staging accuracy. The model comprises two network streams, each learning from a distinct view, with a fusion strategy to integrate the contributions of the two views. Additionally, a bidirectional LSTM is employed to capture temporal dependencies within sleep stage sequences. XSleepNet has demonstrated superior performance over single-view and multi-view baselines by employing a simple fusion strategy. Further details regarding the XSleepNet model can be found in [23]. However, to eliminate the impact of numerical precision and device-specific factors, we reimplemented the XSleepNet models using the PyTorch framework. Experiments were conducted with the reimplemented version of XSleepNet, referred to as XSleepNet-reimp.

In the context of sleep disorder diagnosis based on stage sequences, two types of models were employed: empirical feature-based XGBoost classifiers and end-to-end deep neural networks. One XGBoost classifier was trained on statistical features extracted from the sleep stage sequences, including the percentage of each sleep stage, stage latency, stage duration, sleep efficiency, and transition probabilities between stages (as detailed in [53]). Another XGBoost classifier was trained on hypnodensity features, which represent the classification probabilities for each sleep stage, introduced by [26]. On the other hand, deep neural network models were trained directly on the raw sleep stage sequences. As a baseline end-to-end model (BaseNet), we employed the widely recognized fastText model for feasibility validation [71]. The fastText model

consists of an embedding layer followed by averaging operations and a hidden layer (Fig. **S8**). While extremely simple, it often matches the accuracy of deep learning classifiers and is significantly faster in both training and evaluation.

Experimental Setup

Settings for SleepGPT pretraining

Importantly, the sleep stage vocabulary consists of only five "words", namely {W, N1, N2, N3, REM}, so it is significantly smaller than the vocabulary of a typical language model. Consequently, the size of SleepGPT is much smaller than the original GPT model. Specifically, the SleepGPT model architecture comprises 3 stacked transformer blocks, each having 6 attention heads and 48 hidden units. Both the token and position embeddings have a size of 48. Training is performed on the SHHS dataset for 50 epochs, utilizing a batch size of 256. The learning rate is set at 5e-4, and the Adam optimizer is employed, utilizing a cosine learning rate decay schedule. The training is implemented is done via the PyTorch framework and the HuggingFace Transformers library.

Settings for sleep staging

We adapted the original TinySleepNet and XSleepNet by eliminating the temporal LSTM layer, allowing them to make predictions based solely on the current EEG epoch. This exclusion of memory resulted in versions named TinySleepNet-nonseq and XSleepNet-nonseq. We deployed both TinySleepNet-nonseq and XSleepNet-nonseq and their original implementations, TinySleepNet and XSleepNet, for a thorough evaluation of the proposed SleepGPT model. The temporal sequential model has a sequence length of 15, utilizing the current and preceding 14 epochs to forecast the current sleep stage. All sleep staging models underwent 200 training epochs, utilizing a batch size of 256. A learning rate of 5e-4 was employed along with the Adam optimizer, implementing a cosine learning rate decay schedule. To determine a reasonable value for the weight α in Eq. 8, which influences the balance between the sleep staging model and the SleepGPT model, a grid search was conducted on an independent validation set. The evaluation of model performance involved ten-fold cross-validation, conducted at either the subject or session level, aligning with the evaluation protocols commonly adopted in related research. Independent validation experiments were also conducted to assess the generalization performance of the models across different sleep datasets.

Settings for sleep disorder diagnosis

Given that a sleep cycle typically spans approximately 90 minutes, multiple SleepGPT models were trained, each with distinct context sizes: 30, 60, 120, and 180 epochs (equal to 15, 30, 60, and 90 minutes, respectively). The context size directly influences the segment lengths of the sleep stage sequences used in our sleep disorder diagnosis HTN model. Despite the use of larger local context sizes, noteworthy performance gains were not observed. Instead, these larger sizes led to increased training time and memory utilization. Consequently, a context size of 60 epochs (30 minutes) was

selected for both the SleepGPT model and the HTN model. In the HTN architecture, the local feature extraction SleepGPT module adopts a configuration identical to that of the pretraining phase, facilitating the direct application of the pretrained model. For global feature extraction through the transformer encoder, one transformer block consisting of 6 attention heads and 48 hidden units is sufficient. Moreover, to validate the efficacy of pretraining, we trained the hierarchical sleep disorder diagnosis model entirely from scratch. Additionally, we evaluated the previously mentioned fastText model as a baseline in the context of the sleep disorder diagnosis task, facilitating direct comparison. Training was conducted for 100 epochs on both the CAP and MNC datasets, employing a batch size of 16. The learning rate was set at 5e-5, with optimization performed by the Adam optimizer with a cosine learning rate decay schedule. Furthermore, following previous protocols, both ten-fold cross-validation and independent-sample test were utilized to evaluate model performance.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The SHHS [45] and MNC [26] 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 SleepEDF [46], Physio2018 [48], and CAP [50] datasets are available from PhysioNet at https://physionet.org/content/sleep-edfx/1.0.0/, https://physionet.org/content/challenge-2018/1.0.0/, and https://physionet.org/content/capslpdb/1.0.0/, respectively. The MASS [47] dataset is available at http://ceams-carsm.ca/mass/. The BOAS [49] dataset can be accessed at OpenNeuro. The ISRUC [51] dataset can be accessed at https://sleeptight.isr.uc.pt/. Access to the HANG7 dataset is governed by data-use agreements, and it is therefore not publicly available.

Code availability

The models and source codes for reproducing the results reported in this paper can be accessed at https://github.com/yuty2009/sleepgpt.

Acknowledgements

This work was supported in part by STI2030-Major Projects under Grant 2022ZD0211700, the National Natural Science Foundation of China under Grant 62376098, 62276102, and U22A20293, and GuangDong Basic and Applied Basic Research Foundation 2024A1515011983.

Author contributions

T.Y. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. Z.G. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. R.H., F.W., M.L., J.Y., Z.Y., J.Z., Y.X., H.J., W.L., G.D., Z.G., Y.W, J.L., Y.Z., and M.J. contributed to the analysis and interpretation of the data. Y.L. contributed to the analysis and interpretation of the data, and the drafting of the manuscript. J.X. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. W.W. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. All authors reviewed the manuscript.

Competing interests

W.W. reports equity from Alto Neuroscience. None of the other authors has financial disclosures to report.

Additional information

Additional study and data information is contained in the supplementary material.

References

[1] Mahowald, M.W., Schenck, C.H.: Insights from studying human sleep disorders. Nature **437**(7063), 1279–85 (2005) https://doi.org/10.1038/nature04287

[2] Zheng, N.S., Annis, J., Master, H., Han, L., Gleichauf, K., Ching, J.H., Nasser, M., Coleman, P., Desine, S., Ruderfer, D.M., Hernandez, J., Schneider, L.D., Brittain, E.L.: Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the all of us research program. Nature Medicine **30**(9), 2648–2656 (2024) https://doi.org/10.1038/s41591-024-03155-8

[3] Iber, C., Ancoli-Israel, S., Chesson, A.L., Quan, S.: The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. American Academy of Sleep Medicine (2007)

[4] MacLean, A.W., Lue, F., Moldofksy, H.: The reliability of visual scoring of alpha eeg activity during sleep. Sleep **18**(7), 565–9 (1995)

[5] Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., Saletu, B., Schmidt, A., Dorffner, G.: Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. J. Sleep Res. **18**(1), 74–84 (2009) https://doi.org/10.1111/j.1365-2869.2008.00700.x

[6] Rosenberg, R.S., Van Hout, S.: The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. J. Clin. Sleep Med. **9**(1), 81–7 (2013) https://doi.org/10.5664/jcsm.2350

[7] Alickovic, E., Subasi, A.: Ensemble svm method for automatic sleep stage classification. IEEE Transactions on Instrumentation and Measurement **67**(6), 1258–1265 (2018) https://doi.org/10.1109/TIM.2018.2799059

[8] Li, X., Cui, L., Tao, S., Chen, J., Zhang, X., Zhang, G.-Q.: HyCLASSS: A hybrid classifier for automatic sleep stage scoring. IEEE Journal of Biomedical and Health Informatics **22**(2), 375–385 (2018) https://doi.org/10.1109/JBHI.2017.2668993

[9] 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**(11), 1998–2008 (2017) https://doi.org/10.1109/TNSRE.2017.2721116

[10] Memar, P., Faradji, F.: A novel multi-class eeg-based sleep stage classification system. IEEE Trans. Neural Syst. Rehabil. Eng. **26**(1), 84–95 (2018) https://doi.org/10.1109/TNSRE.2017.2776149

[11] Mousavi, S., Afghah, F., Acharya, U.R.: SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. PLoS One **14**(5), 0216456 (2019) https://doi.org/10.1371/journal.pone.0216456

[12] Supratak, A., Guo, Y.: TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In: The 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 641–644 (2020). https://doi.org/10.1109/EMBC44109.2020.9176741

[13] Phan, H., Andreotti, F., Cooray, N., Chen, O.Y., De Vos, M.: Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Trans. Biomed. Eng. **66**(5), 1285–1296 (2019) https://doi.org/10.1109/TBME.2018.2872652

[14] Phan, H., Andreotti, F., Cooray, N., Chen, O.Y., De Vos, M.: SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans. Neural Syst. Rehabil. Eng. **27**(3), 400–410 (2019) https://doi.org/10.1109/TNSRE.2019.2896659

[15] Phan, H., Lorenzen, K.P., Heremans, E., Chén, O.Y., Tran, M.C., Koch, P., Mertins, A., Baumert, M., Mikkelsen, K.B., De Vos, M.: L-SeqSleepNet: Whole-cycle long sequence modeling for automatic sleep staging. IEEE Journal of Biomedical and Health Informatics **27**(10), 4748–4757 (2023) https://doi.org/10.1109/JBHI.2023.3303197

[16] Phan, H., Mikkelsen, K., Chén, O.Y., Koch, P., Mertins, A., De Vos, M.: SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Transactions on Biomedical Engineering **69**(8), 2456–2467 (2022) https://doi.org/10.1109/TBME.2022.3147187

[17] Dai, Y., Li, X., Liang, S., Wang, L., Duan, Q., Yang, H., Zhang, C., Chen, X., Li, L., Li, X., Liao, X.: MultiChannelSleepNet: A transformer-based model for automatic sleep stage classification with psg. IEEE Journal of Biomedical and Health Informatics **27**(9), 4204–4215 (2023) https://doi.org/10.1109/JBHI.2023.3284160

[18] Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P.M., Guo, Y.: Mixed neural network approach for temporal sleep stage classification. IEEE Trans. Neural Syst. Rehabil. Eng. **26**(2), 324–333 (2018) https://doi.org/10.1109/TNSRE.2017.2733220

[19] Perslev, M., Jensen, M.H., Darkner, S., Jennum, P.J., Igel, C.: U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging. Curran Associates Inc., Red Hook, NY, USA (2019)

[20] Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C.K., Li, X., Guan, C.: An attention-based deep learning approach for sleep stage classification with single-channel eeg. IEEE Trans. Neural Syst. Rehabil. Eng. **29**, 809–818 (2021) https://doi.org/10.1109/TNSRE.2021.3076234

[21] Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P.J., Igel, C.: U-Sleep: resilient high-frequency sleep staging. npj Digital Medicine **4**(1), 72 (2021) https://doi.org/10.1038/s41746-021-00440-5

[22] Jia, Z., Lin, Y., Wang, J., Ning, X., He, Y., Zhou, R., Zhou, Y., Lehman, L.-w.H.: Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **29**, 1977–1986 (2021) https://doi.org/10.1109/TNSRE.2021.3110665

[23] Phan, H., Chen, O.Y., Tran, M.C., Koch, P., Mertins, A., De Vos, M.: XSleepNet: Multi-view sequential model for automatic sleep staging. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5903–5915 (2022) https://doi.org/10.1109/TPAMI.2021.3070057

[24] Seo, H., Back, S., Lee, S., Park, D., Kim, T., Lee, K.: Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. Biomedical Signal Processing and Control **61**, 102037 (2020) https://doi.org/10.1016/j.bspc.2020.102037

[25] Zhou, W., Zhu, H., Shen, N., Chen, H., Fu, C., Yu, H., Shu, F., Chen, C.,

Chen, W.: A lightweight segmented attention network for sleep staging by fusing local characteristics and adjacent information. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 238–247 (2023) https://doi.org/10.1109/TNSRE.2022.3220372

[26] Stephansen, J.B., Olesen, A.N., Olsen, M., Ambati, A., Leary, E.B., Moore, H.E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y.L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S.C., Kim, T.W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S.T., Schweitzer, P.K., Kushida, C., Peppard, P.E., Sorensen, H.B.D., Jennum, P., Mignot, E.: Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat. Commun. **9**(1), 5229 (2018) https://doi.org/10.1038/s41467-018-07229-3

[27] Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). American Psychiatric Publishing, Washington, DC, USA (2022)

[28] Wikipedia: Hypnogram. https://en.wikipedia.org/wiki/Hypnogram (2023)

[29] Thachayani, M., Loganayagi, M.: Artificial intelligence based classifier for sleep disorder detec-tion using eeg-bci data. Int. J. Comp. Sci. Trends. Technol. **9** (2021)

[30] Sharma, M., Dhiman, H.S., Acharya, U.R.: Automatic identification of insomnia using optimal antisymmetric biorthogonal wavelet filter bank with ecg signals. Comput Biol Med **131**, 104246 (2021) https://doi.org/10.1016/j.compbiomed.2021.104246

[31] Dimitriadis, S.I., Salis, C.I., Liparas, D.: An automatic sleep disorder detection based on eeg cross-frequency coupling and random forest model. J. Neural Eng. **18**(4) (2021) https://doi.org/10.1088/1741-2552/abf773

[32] Phyo, J., Ko, W., Jeon, E., Suk, H.-I.: TransSleep: Transitioning-aware attention-based deep neural network for sleep staging. IEEE Transactions on Cybernetics **53**(7), 4500–4510 (2023) https://doi.org/10.1109/TCYB.2022.3198997

[33] Phan, H., Mikkelsen, K.: Automatic sleep staging of eeg signals: recent development, challenges, and future directions. Physiol. Meas. **43**(4) (2022) https://doi.org/10.1088/1361-6579/ac6049

[34] Liang, S.-F., Kuo, C.-E., Hu, Y.-H., Cheng, Y.-S.: A rule-based automatic sleep staging method. Journal of Neuroscience Methods **205**(1), 169–176 (2012) https://doi.org/10.1016/j.jneumeth.2011.12.022

[35] Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., Jernajczyk, W., Riener, R., Buhmann, J., Achermann, P.: Automatic human sleep stage scoring using deep neural networks. Frontiers in Neuroscience **12** (2018)

[36] Bianchi, M.T., Cash, S.S., Mietus, J., Peng, C.-K., Thomas, R.: Obstructive sleep apnea alters sleep stage transition dynamics. PLOS ONE **5**(6), 1–12 (2010) https://doi.org/10.1371/journal.pone.0011356

[37] Xu, S., Faust, O., Seoni, S., Chakraborty, S., Barua, P.D., Loh, H.W., Elphick, H., Molinari, F., Acharya, U.R.: A review of automated sleep disorder detection. Comput. Biol. Med. **150**, 106100 (2022) https://doi.org/10.1016/j.compbiomed.2022.106100

[38] Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., Yao, J.: scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence **4**(10), 852–866 (2022) https://doi.org/10.1038/s42256-022-00534-z

[39] Born, J., Manica, M.: Regression transformer enables concurrent sequence regression and generation for molecular language modelling. Nature Machine Intelligence **5**(4), 432–444 (2023) https://doi.org/10.1038/s42256-023-00639-z

[40] Hie, B., Zhong, E.D., Berger, B., Bryson, B.: Learning the language of viral evolution and escape. Science **371**(6526), 284–288 (2021) https://doi.org/10.1126/science.abd7331

[41] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023) https://doi.org/10.1126/science.ade2574

[42] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B.: ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 7112–7127 (2022) https://doi.org/10.1109/TPAMI.2021.3095381

[43] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning (2018)

[44] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[45] Quan, S.F., Howard, B.V., Iber, C., Kiley, J.P., Nieto, F.J., O'Connor, G.T., Rapoport, D.M., Redline, S., Robbins, J., Samet, J.M., Wahl, P.W.: The Sleep Heart Health Study: Design, Rationale, and Methods. Sleep **20**(12), 1077–1085 (1997) https://doi.org/10.1093/sleep/20.12.1077

[46] Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A.C., Oberye, J.J.L.:

Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. IEEE Transactions on Biomedical Engineering **47**(9), 1185–1194 (2000) https://doi.org/10.1109/10.867928

[47] O'Reilly, C., Gosselin, N., Carrier, J., Nielsen, T.: Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. Journal of Sleep Research **23**(6), 628–635 (2014) https://doi.org/10.1111/jsr.12169

[48] Ghassemi, M.M., Moody, B.E., Lehman, L.H., Song, C., Li, Q., Sun, H., Mark, R.G., Westover, M.B., Clifford, G.D.: You snooze, you win: the physionet/computing in cardiology challenge 2018. Comput Cardiol (2010) **45** (2018) https://doi.org/10.22489/cinc.2018.049

[49] López-Larraz, E., Sierra-Torralba, M., Clemente, S., Fierro, G., Oriol, D., Minguez, J., Montesano, L., Klinzing, J.G.: "bitbrain open access sleep dataset" (2024) https://doi.org/10.18112/openneuro.ds005555.v1.0.0

[50] Terzano, M.G., Parrino, L., Sherieri, A., Chervin, R., Chokroverty, S., Guilleminault, C., Hirshkowitz, M., Mahowald, M., Moldofsky, H., Rosa, A., Thomas, R., Walters, A.: Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Medicine **2**(6), 537–553 (2001) https://doi.org/10.1016/S1389-9457(01)00149-6

[51] Khalighi, S., Sousa, T., Santos, J.M., Nunes, U.: Isruc-sleep: A comprehensive public dataset for sleep researchers. Computer Methods and Programs in Biomedicine **124**, 180–192 (2016) https://doi.org/10.1016/j.cmpb.2015.10.013

[52] Wang, J., Zhao, S., Zhou, Y., Jiang, H., Yu, Z., Li, T., Li, S., Pan, G.: Narcolepsy diagnosis with sleep stage features using psg recordings. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 3619–3629 (2023) https://doi.org/10.1109/TNSRE.2023.3312396

[53] Vallat, R., Walker, M.P.: An open-source, high-performance tool for automated sleep staging. eLife **10**, 70092 (2021) https://doi.org/10.7554/eLife.70092

[54] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009) https://doi.org/10.1016/j.ipm.2009.03.002

[55] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**, 37–46 (1960) https://doi.org/10.1177/001316446002000104

[56] Rosenberg, R.S., Van Hout, S.: The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. Journal of clinical sleep medicine **9**(1), 81–87 (2013)

[57] Mignot, E., Lammers, G.J., Ripley, B., Okun, M., Nevsimalova, S., Overeem, S., Vankova, J., Black, J., Harsh, J., Bassetti, C., Schrader, H., Nishino, S.: The Role of Cerebrospinal Fluid Hypocretin Measurement in the Diagnosis of Narcolepsy and Other Hypersomnias. Archives of Neurology **59**(10), 1553–1562 (2002) https://doi.org/10.1001/archneur.59.10.1553

[58] Andlauer, O., Moore, I. Hyatt, Hong, S.-C., Dauvilliers, Y., Kanbayashi, T., Nishino, S., Han, F., Silber, M.H., Rico, T., Einen, M., Kornum, B.R., Jennum, P., Knudsen, S., Nevsimalova, S., Poli, F., Plazzi, G., Mignot, E.: Predictors of Hypocretin (Orexin) Deficiency in Narcolepsy Without Cataplexy. Sleep **35**(9), 1247–1255 (2012) https://doi.org/10.5665/sleep.2080

[59] Andlauer, O., Moore, I. Hyatt, Jouhier, L., Drake, C., Peppard, P.E., Han, F., Hong, S.-C., Poli, F., Plazzi, G., O'Hara, R., Haffen, E., Roth, T., Young, T., Mignot, E.: Nocturnal Rapid Eye Movement Sleep Latency for Identifying Patients With Narcolepsy/Hypocretin Deficiency. JAMA Neurology **70**(7), 891–902 (2013) https://doi.org/10.1001/jamaneurol.2013.1589

[60] Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., Redline, S.: The National Sleep Research Resource: towards a sleep data commons. Journal of the American Medical Informatics Association **25**(10), 1351–1358 (2018) https://doi.org/10.1093/jamia/ocy064

[61] Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet. Circulation **101**(23), 215–220 (2000) https://doi.org/10.1161/01.CIR.101.23.e215

[62] Wolpert, E.A.: A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Archives of General Psychiatry **20**(2), 246–247 (1969) https://doi.org/10.1001/archpsyc.1969.01740140118016

[63] Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964 (2016). https://doi.org/10.1109/ICASSP.2016.7472621

[64] Rao, K., Sak, H., Prabhavalkar, R.: Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 193–199 (2017). https://doi.org/10.1109/ASRU.2017.8268935

[65] Karita, S., Soplin, N., Watanabe, S., Delcroix, M., Ogawa, A., Nakatani, T.: Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH **2019-September**, 1408–1412 (2019) https://doi.org/10.21437/

Interspeech.2019-1938 . Publisher Copyright: Copyright © 2019 ISCA; 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019 ; Conference date: 15-09-2019 Through 19-09-2019

[66] Kenton, J.D.M.-W.C., Toutanova, L.K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)

[67] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[68] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-1174 . https://aclanthology.org/N16-1174

[69] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., Dehak, N.: Hierarchical transformers for long document classification. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 838–844 (2019). https://doi.org/10.1109/ASRU46091.2019.9003958

[70] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YicbFdNTTy

[71] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017). https://aclanthology.org/E17-2068