SLEEPJ, 2025, 48, 1–19
https://doi.org/10.1093/sleep/zsaf134
Advance Access publication 24 June 2025
Original Article

Original Article

CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics

Samaneh Nasiri[1,2,3,4,*,†], Wolfgang Ganglberger[1,3,†], Thijs Nassi[1,3,5,†], Erik-Jan Meulenbrugge[1,3,†], Valdery Moura Junior[2,3], Manohar Ghanta[2,3], Aditya Gupta[2,3], Katie L. Stone[6,7], Magnus Ruud Kjaer[8], Oliver Sum-Ping[8], Emmanuel Mignot[8], Dennis Hwang[9], Lynn Marie Trotti[10], Gari D. Clifford[4,11], Umakanth Katwa[3,12], Haoqi Sun[1,3,‡], Robert J. Thomas[3,13,‡] and M. Brandon Westover[1,3,‡]

[1]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA,
[2]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA,
[3]Department of Neurology, Harvard Medical School, Boston, MA, USA,
[4]Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA, USA,
[5]Cardiovascular and Respiratory Physiology Group, University of Twente, Enschede, NL, USA,
[6]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA,
[7]Epidemiology and Biostatistics, California Pacific Medical Center Research Institute, San Francisco, CA, USA,
[8]School of Medicine, Stanford University, Palo Alto, CA, USA,
[9]Kaiser Permanente, San Bernardino County Sleep Disorders Center. San Bernardino, CA, USA,
[10]Department of Neurology and Emory Sleep Center, Emory University School of Medicine, Atlanta, GA, USA,
[11]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA,
[12]Division of Sleep Medicine, Boston Children's Hospital, Boston, MA, USA and
[13]Department of Medicine, Division of Pulmonary Critical Care & Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

*Corresponding author. Samaneh Nasiri, Department of Biomedical Informatics, Emory School of Medicine, 100 Woodruff Circle Atlanta, GA 30322, USA. Email: snasiri@emory.edu.
†Equal contribution, co-first authors
‡Equal contribution, co-senior author

Abstract

Study Objectives: To develop and validate a Complete Artificial Intelligence Sleep Report system (CAISR), a system for comprehensive automated sleep analysis, including sleep staging, arousal detection, apnea identification, and limb movement analysis.

Methods: We utilized a large diverse dataset from four cohorts (MGH, MESA, MrOS, SHHS) comprising 25,749 participants to develop CAISR. Following American Academy of Sleep Medicine (AASM) guidelines, CAISR performs four tasks: it stages sleep into five categories (Wake, NREM 1, NREM 2, NREM 3, REM), detects arousals, detects and classifies breathing events (Obstructive Apnea, Central Apnea, Mixed Apnea, Hypopnea, and RERA), and detects limb movements and categorizes them as periodic or isolated. We tested CAISR against multiple datasets independently annotated by multiple experts, including UPenn (69 subjects, 6 experts), BITS (98 subjects, 3 experts), and Stanford (100 subjects, three experts). Sleep staging and arousal detection were accomplished using customized deep neural networks, while breathing event detection and classification and limb movement analysis were accomplished using rule-based signal processing approaches. We quantified CAISR performance with three metrics: Cohen's Kappa, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). To determine whether CAISR performed on par with human experts, we compared expert inter-rater reliability (IRR) with algorithm-expert IRR.

Results: The CAISR model showed strong overall performance across the four tasks: sleep staging, arousal detection, apnea detection, and limb movement detection. In sleep staging, the model achieved AUROC values ranging from 0.82 to 0.97 and AUPRC values between 0.63 and 0.90 across the BITS, Stanford, and Penn datasets, indicating high classification accuracy. The Kappa agreement analysis showed that in the BITS and Stanford datasets, CAISR outperformed human experts, with non-overlapping confidence intervals indicating superiority (Kappa values around 0.7 to 0.8 for CAISR vs. experts). In the Penn dataset, the model's performance was comparable to experts, with overlapping confidence intervals suggesting non-inferiority. For arousal detection, the model maintained reliable performance, with AUROC values ranging from 0.83 to 0.94 and AUPRC values from 0.67 to 0.85, and Kappa analysis showing overlapping confidence intervals, indicating comparable performance to experts in both the BITS and Stanford datasets (Kappa values for CAISR around 0.6 to 0.75). In apnea detection, including the detection of obstructive, central, and mixed apnea, the CAISR model achieved competitive results in the BITS dataset with AUROC values between 0.81 and 0.95 and AUPRC values between 0.58 and 0.82, but in the Stanford dataset, it underperformed compared to human experts, as shown by non-overlapping confidence intervals and lower Kappa values (around 0.55 to 0.65). Finally, in limb movement detection, the model demonstrated superior performance

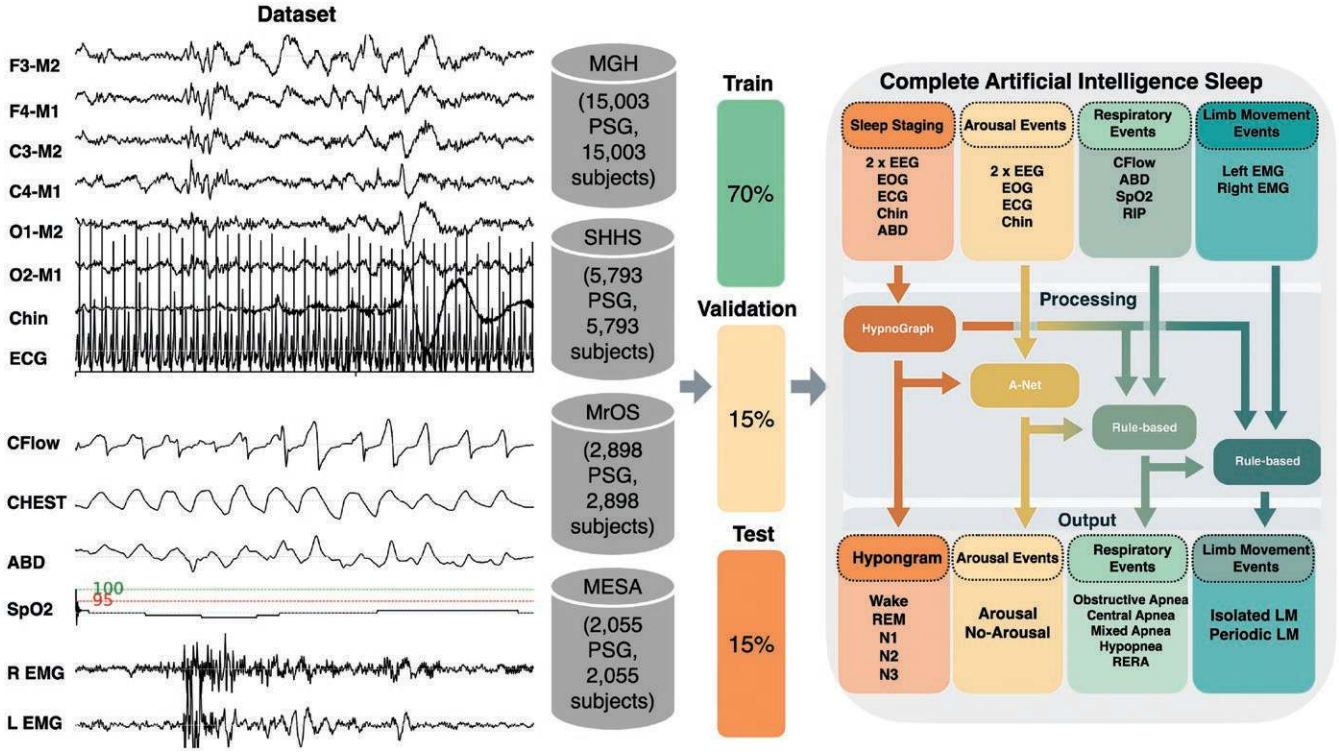
Submitted for publication: January 6, 2025; **Revised:** May 6, 2025

in the BITS dataset, with AUROC values of 0.9 to 0.96 and AUPRC values between 0.75 and 0.85, and Kappa analysis indicating significantly higher reliability compared to experts (CAISR Kappa around 0.8, with non-overlapping confidence intervals). In the Stanford dataset, CAISR's performance was comparable to experts, with overlapping confidence intervals suggesting non-inferiority (Kappa values around 0.65 to 0.7). Overall, the CAISR model consistently exhibited high classification performance and reliability across tasks, often matching or surpassing expert-level performance, with particularly strong results in sleep staging and limb detection.

Conclusions: The CAISR model demonstrated high classification accuracy and reliability across sleep staging, arousal, apnea, and limb movement detection tasks, matching or surpassing human expert performance. Human errors and systematic biases in the annotation of micro-events during sleep, such as arousal and apnea detection, likely contributed to variability in expert performance, while the CAISR model showed more consistent results, reducing the impact of these biases and increasing overall reliability across task.

Key words: sleep staging; arousal detection; apnea detection; limb movement; inter-rater reliability; deep learning; transfer learning; rule-based model; few-shot learning

Graphical Abstract



Statement of significance

CAISR represents a clinically significant milestone in sleep analysis, demonstrating an automated system that performs at or above the level of human experts across all tasks in conventional clinical polysomnography analysis. This represents the first model to comprehensively match or outperform human review across sleep staging, arousal detection, apnea identification, and limb movement analysis. CAISR has the potential to standardize sleep scoring across institutions could lead to more consistent patient care and accelerate sleep research by providing a reliable, scalable tool for large-scale data analysis.

Introduction

Sleep is a vital biological process for maintaining health [1–4]. Sleep plays numerous health-restorative functions, including memory consolidation, immune system regulation, and hormonal regulation [5–7]. Analysis of sleep physiological signals is essential for diagnosing and managing a range of medical conditions, including sleep disorders, neurological diseases, and mental health disorders [1, 8–10].

The conventional clinical approach to sleep analysis involves manual examination by technician or physician experts of polysomnography (PSG) recordings, which include electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), electromyography (EMG), and respiratory signals [11, 12]. Conventional PSG analysis includes four tasks: Sleep staging (classifying each 30-second epoch Wake, N1, N2, N3, or REM) [11, 13]; detecting the onset and offset times of arousals (brief awakenings or shifts to lighter sleep stages); detecting and classifying five types of breathing disturbances (obstructive apnea, central apnea, mixed apnea, hypopnea, and respiratory effort-related arousal [RERA]); and detecting and classifying limb movements and limb-movement patterns (periodic limb movements [PLMs] vs. isolated limb movements). Detailed rules for

scoring each of these tasks are provided in scoring guidelines from the American Academy of Sleep Medicine (AASM) [14, 15].

However, manual PSG scoring has significant limitations. Scoring is labor-intensive, requiring about 2 hours to score a typical overnight PSG recording [16, 17]. This creates an access bottleneck, as the demand for sleep studies outstrips the availability of trained sleep specialists and clinics. Manual scoring is also prone to inter- and intra-rater variability, leading to inconsistencies in diagnosis and treatment recommendations [18–20]. These also impede the development and clinical application of novel automated methods that require accurate event labels as training data. These challenges highlight the need for reliable automated scoring methods that can increase efficiency and consistency, and expand access to sleep diagnostics [21–24].

Automated systems have the potential to standardize sleep analysis across institutions, reduce healthcare costs, and improve patient care by providing more timely and accurate sleep assessments [25–27]. Previous studies on automated sleep analysis, utilizing a range of models, have demonstrated high accuracy on isolated tasks on selected datasets but have not integrated all tasks into a single system [16, 22, 28, 29]. MSED [30] proposed a multi-task model for detecting arousal, limb movement, and apnea. However, their approach for apnea detection was limited to a binary classification, focusing solely on detecting apnea and hypopnea, without incorporating respiratory events, including various clinically meaningful phenotypes, such as obstructive, central, mixed apnea, hypopnea, and respiratory effort-related arousals (RERAs). Additionally, MSED did not integrate sleep staging into its system. The model was trained and validated using the all-male MrOS dataset, which may limit its generalizability to broader clinical populations. Here, we present the Complete Artificial Intelligence Sleep Report (CAISR) system, incorporating sleep staging, arousal detection, apnea identification, and limb movement analysis in one unified system (**Figure 1**) trained on a considerably larger/diverse dataset. We rigorously evaluate CAISR against multiple large, independently annotated datasets, demonstrating its ability to achieve and surpass expert-level accuracy in comprehensive sleep analysis. This comprehensive approach is a major step forward in automating sleep diagnostics.

Methods

Ethical approval

This retrospective data analysis study was conducted under IRB protocol number (BIDMC: # 2016P000058, MGH: # 2013P001024), with the MGH and BIDMC IRBs granting a waiver of consent.

Datasets

CAISR was developed and evaluated using PSG records from 26,048 participants from 7 independent datasets. Unless otherwise noted, the PSGs contained at minimum the following signals: Two EEG, two EOG (E1-M2, E2-M1), one chin EMG, thoracic and abdominal inductance plethysmography effort belts, nasal pressure, thermistor, oxygen saturation, leg EMG, and electrocardiogram (ECG). Except where otherwise noted (MGH dataset whereas it is scored with the 4% hypopnea rule), all data were scored according to the AASM criteria in effect at the time the study was conducted, as the manual scoring guidelines have evolved over time [31]; sleep stages were scored for consecutive 30-second epochs into classes (Wake (W), N1, N2, N3, R); respiratory events were marked with the beginning and end of each event and categorized (obstructive apnea, central apnea, mixed apnea, hypopnea, respiratory-effort-related arousal [RERA]); arousal events and limb movement events were marked with the beginning and end of each event. Hypopneas were scored per AASM recommendations, that is, a hypopnea event must meet a ≥ 3% oxygen desaturation from pre-event baseline and/or the event is associated with an arousal. Details of the eight datasets used in this study are as follows.

1. The "MGH" cohort comprises 15,003 PSG recordings from 15,003 patients recorded in the Massachusetts General Hospital (MGH) sleep laboratory. Each recording was scored by one of seven sleep technicians and subsequently edited by a sleep physician as considered needed [32]. The alternative AASM-accepted definition of hypopneas was used for scoring, that is, hypopneas need to meet a ≥ 4% oxygen desaturation from pre-event baseline.
2. The Sleep Heart Health Study (SHHS) is a multi-center cohort study investigating the health consequences of sleep-disordered breathing. A total of 6,441 men and women aged 40 years and older were enrolled between November 1, 1995, and January 31, 1998 ("Visit 1"). A second polysomnogram ("Visit 2") was obtained from 3,295 participants [33] between January 2001—June 2003. Each PSG was scored by one sleep technician. The PSGs do not contain nasal pressure signals. Of the 6441 total SHHS participants, we included data from the 5793 who opted to make their data available.
3. The Multi-Ethnic Study of Atherosclerosis (MESA) includes 6,814 men and women from diverse racial/ethnic backgrounds. The MESA Sleep ancillary study, conducted between 2010 and 2013, included one PSG recording from 2,055 individuals [34, 35]. Each PSG was scored by one sleep technician.
4. The Osteoporotic Fractures in Men (MrOS) Sleep Study includes 2,898 participants who underwent in-home PSG between 2003 and 2005 [36]. Each PSG was scored by one sleep technician.
5. The Beth Israel Triple Scored (BITS) cohort consists of 98 PSG recordings from 98 subjects evaluated at the Beth Israel Deaconess Medical Center (BIDMC) sleep laboratory. All data were independently scored by three experienced sleep-scoring technicians from the BIDMC sleep laboratory.
6. The UPenn Data cohort comprises 69 research PSGs from 69 individuals, conducted in clinical sleep lab settings [37]. Each PSG was independently sleep-staged by seven sleep technicians from three different sleep labs (University of Pennsylvania, St. Luke's Hospital, Stanford University). For this dataset, only sleep staging annotations were available.
7. The Stanford dataset consists of 100 PSG recordings from 100 individuals from the Stanford sleep lab. Each PSG was independently scored for all sleep event types (sleep staging, arousal, apnea events, and limb movements) by three sleep technicians.

Table 1 summarizes each dataset. The datasets are categorized based on the number of independent scorers (single-scored vs multi-scored) and by which event types are scored (sleep stages, arousals, breathing events, limb movements). Single-scored PSGs, annotated by a single rater, include sleep staging (MGH, SHHS, MrOS, MESA); arousals (MGH, MrOS, MESA); apnea (MGH, MESA, MrOS); and limb movements (MGH). Multi-scored datasets, where PSG data were independently annotated by multiple

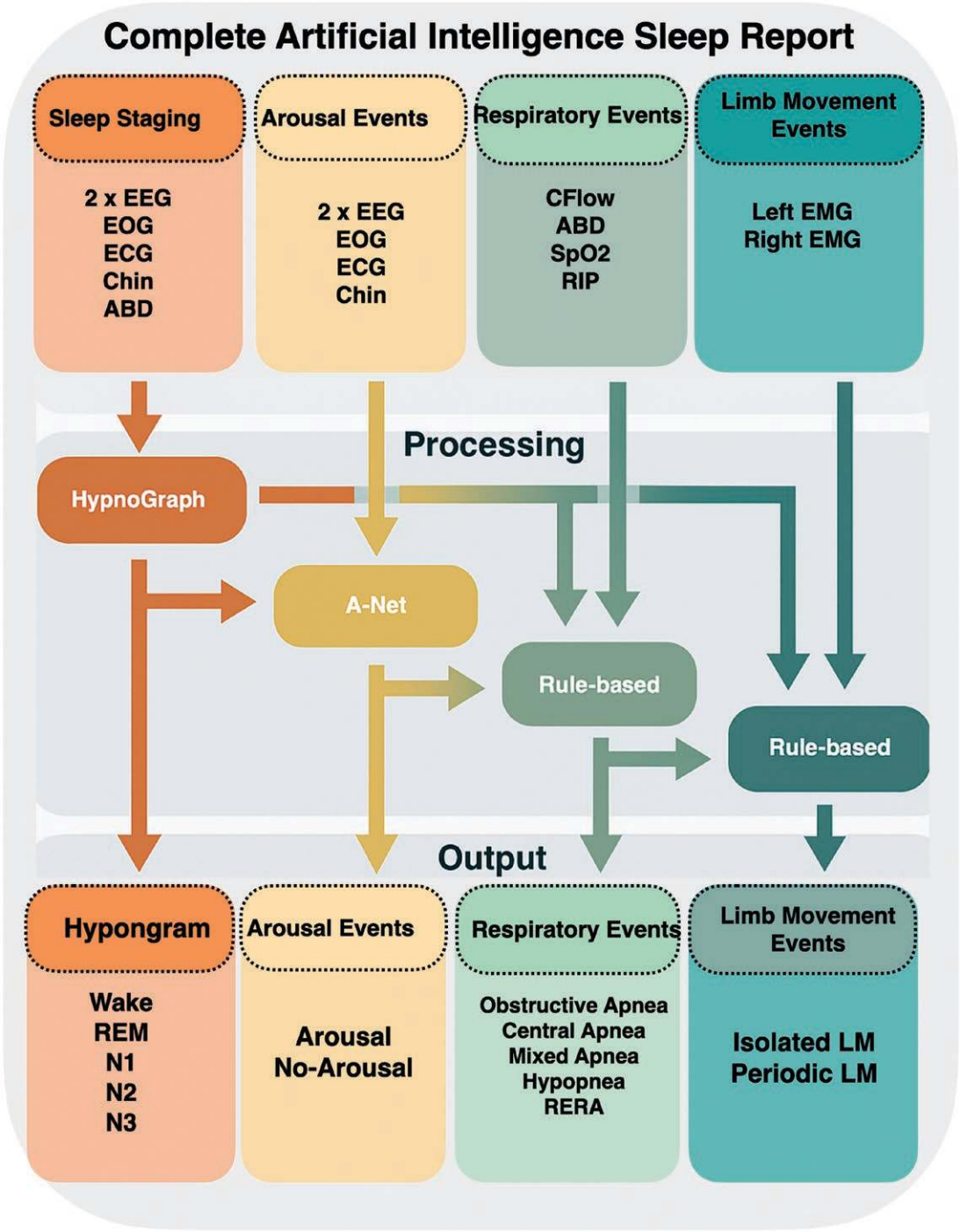


Figure 1. Workflow of the Complete Artificial Intelligence Sleep Report System (CAISR). This figure illustrates the comprehensive operational process of CAISR. The system integrates various physiological sleep signals, including EEG, EOG, EMG, and respiratory signals, to perform sleep staging, arousal detection, apnea identification, and limb movement analysis. The flowchart highlights the data processing pipeline, starting from raw PSG data input, through preprocessing and feature extraction, to the application of deep neural networks and rule-based algorithms. CAISR's predictions are then validated against expert (gold) and super-expert (platinum) annotations, ensuring robust performance across diverse datasets.

raters, include sleep staging (UPenn, BITS, Stanford), arousals (BITS, Stanford), breathing (BITS, Stanford), and limb movement (BITS, Stanford) datasets. Please note that the validation datasets (BITS, Stanford, MGH) used six EEG channels (C3, C4, F3, F4, O1, O2), which provide at least three EEG leads (frontal, central, and occipital) for human experts to classify sleep stages. The extra channels, such as frontal and occipital EEG leads in BITS, MGH, and Stanford datasets, are utilized when available. For the sleep staging task, we used C3-M2 and C4-M1, but these channels can be randomly selected from the available EEG channels. For the arousal task, we randomly chose two EEG channels from the available options for model training and evaluation. Similarly, for other modalities such as EOG and EMG, we selected channels based on availability while ensuring consistency across datasets. The MGH, BITS, and Stanford datasets are accessible through The Human Sleep Project (HSP) at https://bdsp.io/content/hsp/2.0/.

Table 1. Baseline demographics

	MGH	SHHS	MrOS	MESA	BITS	PENN	Stanford
N subjects	15003	5793	2898	2055	98	69	100
Age (Mean, std)	(52, 17)	(63, 11)	(76, 5)	(69, 9)	(55, 17)	(51, 4)	(43, 18)
Age 95% CI	(18, 81)	(42, 83)	(68, 88)	(56, 88)	(23, 83)	(41, 58)	(40, 47)
Sex (% Female)	43	52	0	54	36	100	48
Race (White %)	76	85	91	36	32	54	54
Race (Black or African American %)	6	9	3	28	4	46	4
Race (Asian %)	3	-	3	12	2	-	17
Race (Other %)	12	7	3	24	-	-	17
Race (Unavailable %)	3	-	-	-	61	-	8
Type cohort	Sleep laboratory, attended PSG	Community-dwelling, home-recording	Community-dwelling, home-recording	Community-dwelling, home-recording	Sleep laboratory, attended PSG	Sleep laboratory, attended PSG	Sleep laboratory, attended PSG
N raters	1	1	1	1	3	6	3
N EEG Channels	6	2	2	2	6	2	6
Epworth Sleepiness Scale (Mean, SD)	-	-	-	-	8.1 (5.2)	8.3 (5.5)	-
ICD-based comorbidities (% of subjects)	BITS Sleep Disorders: 90%, Neuropsychiatric Disorders: 33%, Respiratory Conditions: 12%, Cardiovascular Conditions: 62%, Endocrine/Metabolic Conditions: 24%, Neurodegenerative Disorders: 14%, Other Neurological Conditions: 35%, Headache & Migraine Disorders: 12% Stanford Sleep Disorders: 98%, Neuropsychiatric Disorders: 29%, Respiratory Conditions: 5%, Cardiovascular Conditions: 29%, Endocrine/Metabolic Conditions: 13%, Neurodegenerative Disorders: 15%, Other Neurological Conditions: 46%, Headache & Migraine Disorders: 19%, Cognitive Disorders: 6%, Cerebrovascular Disorders: 3%, Neuromuscular Disorders: 7% Seizure & Epileptic Disorders: 5%						

[*]MESA collected race "Chinese," not "Asian." MESA collected "Hispanic" as part of race questionnaire.

MESA, MrOS, and SHHS datasets can be accessed via The National Sleep Research Resource (NSRR) website [38].

Event labels for model training: single-expert labels and enhanced labels

Labels from single-scored datasets were used for training models. Here, we describe the training labels for each task. All labels were defined based on American Academy of Sleep Medicine (AASM) guidelines [39].

Sleep staging.

Thirty-second nonoverlapping epochs of PSG recordings were assigned one of 5 mutually exclusive stage labels: W, N1, N2, N3, REM. Labels for model training came from single-scored datasets and were used as-is; no efforts were made to correct errors in data labels.

Arousal detection.

Arousal detection involves identifying the beginning and end of each arousal event. We identified systematic inaccuracies in the positions and durations of manually scored arousal events, attributable to standardized event lengths and misplacement of event windows due to the use of "hot keys" in conventional sleep scoring software. Additionally, the percentage of apneas with related arousals reached percentages as low as 45% for certain datasets suggesting a systematic tendency to omit scoring of arousals when paired with apneas. To improve label quality for model *training*, we developed a label-adjustment procedure, described in the Supplemental Material. Label adjustment was used only to enhance model *training* data; it was not applied to testing data.

Respiratory event detection and classification.

Respiratory events included obstructive apneas, central apneas, mixed apneas, hypopneas, and RERAs. This is both a detection task (identifying the beginning and end of each event) and a classification task (identifying the type of event). In the MGH dataset, hypopneas were scored according to the 4% desaturation rule. Following a recent update to AASM criteria, hypopneas meeting the 3% desaturation or arousal criteria were additionally identified using an automated method [40]. Scored RERA events that overlapped with these newly added 3% desaturation arousal hypopneas were subsequently reclassified as hypopneas per AASM scoring guidelines. Labels from the MESA, MROS, and SHHS datasets remained unchanged.

Limb movement event detection.

For limb movement detection, we identified the start and end points of all limb movements and then classified them as isolated or periodic limb movements based on AASM scoring guidelines.

This is a detection task, as it involves identifying the beginning and end of each event.

Labels for model *evaluation*: single-expert, multi-expert, and platinum labels

In cases where human scorers exhibit systematic biases, algorithms that exceed human performance can paradoxically appear to perform poorly relative to human experts taken as the gold standard. Several factors can compromise the quality of sleep data labels, including systematic bias among raters trained within the same institution, scoring fatigue, and random variation in scoring behavior between raters [41]. To address this, we created "platinum" labels for each scoring task. Platinum labeling is labor intensive; thus, it was feasible to create platinum labels only for one of the testing datasets: the BITS dataset. BITS was chosen because it had comprehensive labels for all tasks from multiple (3) experienced experts, allowing comparison of experts and algorithms to platinum labels. Examples of arousals, respiratory events, and limb movement that were added to the platinum labels are shown in Supplementary **Figures S15**, **S16, and S17** in the Supplemental Material, respectively.

Exhaustive platinum labeling was done for arousals, breathing, and limb movement events. This was done by members of the research team performing multiple manual and visual labeling rounds with multiple quality checks. Arousals were reviewed by four members of the research team (EJM, SN, MBW, RJT). Breathing events were reviewed by three members of the research team (TN, MBW, RJT). Limb movement events were reviewed by three members of research team (SN, MBW, RJT).

Signal preprocessing

All PSG recordings were resampled to a uniform sampling frequency (fs) of 200 Hz using polyphase filtering [22]. Further preprocessing operations specific to each analysis task are explained in the Supplemental Material.

Model training

Model architecture selection

. The literature was reviewed to select promising candidate model architectures [21, 22, 24, 40, 42–44]. The selected models were trained from scratch to identify the most promising model architecture for further development and inclusion in CAISR. PSG recordings from 8000 individuals from the MGH dataset [32] were divided into training, validation, and test sets consisting of PSG recordings from 5000, 1500, and 1500 distinct individuals, respectively. The held-out testing PSGs were employed to select the top-performing model. The best-performing model was selected for further development.

Training of the selected model architecture was conducted using four single-scored datasets: MGH, SHHS, MrOS, and MESA, comprising 25,749 PSGs from 25,749 individuals. Each dataset was divided into training (at least 70%), validation (up to 15%), and testing (up to 15%) subsets, on a per-subject basis (**Figure S1** in the Supplemental Material). The training subsets were used to train the models, while the validation subsets were employed to monitor performance throughout training. The datasets represent a diverse spectrum of people, including healthy individuals, patients with sleep-related and non-sleep-related disorders, men and women, and individuals from diverse age groups, BMI ranges, and ethnic backgrounds. Collected over several decades, primarily in the United States, these datasets were gathered from various geographical locations and used different hardware, sampling rates, and filters.

We employed the following techniques to develop models for the various sleep analysis tasks. Detailed information on model architectures, training procedures, and algorithms is provided in the Supplemental Material.

Sleep staging.

After evaluating candidate model architectures, ProductGraphSleepNet [21] was selected for further development. This model combines several components: Spatial Attention (SpAtt), Product Graph Learning (PGL), Attentive Graph Convolutional Network (AGC), Bidirectional Gated Recurrent Unit (BiGRU), Graph-wise Attention (GwAT), and a fully connected layer. The model takes two EEG channels as inputs, such as C3-M2 and C4-M1, Cz-Oz, or any other two available EEG channels, one electrooculogram (EOG, E1-M2), chin electromyogram (EMG), abdominal and thoracic respiratory effort, and electrocardiogram (ECG).

Arousal detection.

For arousal detection, we developed ArousalNet (A-Net), inspired by U-Sleep [22] but with expanded input channels and other minor changes for enhanced generalizability. The training utilized a novel batch selection approach to ensure a balanced representation of sleep stages and arousal events.

Respiratory event detection.

A rule-based model was developed based on AASM scoring guidelines. The model's hyperparameters were fine-tuned using random search.

Limb movement detection.

The CAISR-PLM automatic detector was adapted from the Ferri model [42, 45], which was developed based on AASM rules. Our limb movement detection algorithm, named CAISR-PLM, is a rule-based model that uses unsupervised techniques. This means that CAISR-PLM operates purely based on predefined rules (the ASSM rule) and does not involve any learning process. Since there are no parameters to adjust through training, and no loss function to optimize, it is entirely rule-based and does not require training on data. Unlike the original Ferri model, CAISR-PLM incorporates Variable Amplitude Thresholding (VAT) to address signal quality issues [46].

Model evaluation

The three multi-scored datasets (UPenn, BITS, and Stanford) were used for model evaluation; no data from these sources were involved in model building, training, or hyperparameter selection. These datasets enabled the estimation of CAISR's generalizability to new clinical cohorts. Inter-rater reliability (IRR) analysis was conducted by comparing CAISR to human experts.

Model calibration ("fine-tuning")

Model performance can suffer at test time due to differences in the nature of the data or scoring procedures used in a particular test dataset [47]. We applied this approach only to the BITS dataset, as the model's performance matched or outperformed human raters for the other datasets. For the UPenn and Stanford datasets, we used the model as is, as its performance was comparable to or exceeded that of the human raters. The BITS dataset was the only one where the model's performance was lower than that of the human raters. Fine-tuning addresses this challenge by allowing a model trained on a broad dataset to adjust to the specific characteristics of new data [48–50]. This approach

requires fewer labeled examples and less computational effort than training a new model from scratch. We developed a model of a calibration process involving selective adjustment through transfer learning utilizing half of the testing dataset, followed by testing on the remaining half. This process included fine-tuning specific layers, implementing a majority voting mechanism, and hyperparameter tuning. Details are provided in the Supplemental Material.

Performance evaluation: event-level and aggregated metrics

Event classification metrics.

We used two approaches to evaluate model performance: sample-based and event-based evaluations. The sample-based approach treats every data point at the original sampling frequency as possessing a label. The event-based approach groups contiguous samples into single events. For instance, a 10-second apnea event is treated not as multiple individual samples but as one event. The analysis then focuses on whether events overlap between different raters. In our analysis, we consider any overlap as a true positive. This means that for a given predicted event, if it exhibits any degree of overlap with a true event, it is classified as a true positive. This approach is more lenient compared to stricter criteria where a higher Intersection over Union (IoU) threshold is required for a prediction to be considered a true positive. By considering any overlap, we aim to capture all potential true events, which can be particularly useful in scenarios where even partial detection of an event is valuable. This method may increase the sensitivity of the detection system, ensuring that fewer true events are missed, although it may also result in a higher number of false positives depending on the context and application. True negatives in this approach are defined by the average duration observed for such events. The main text reports result from event-based evaluations. Sample-based evaluation results are provided in the Supplemental Material.

To assess model performance, we employed the following event-level performance metrics, which focus on the detection and classification of individual events such as respiratory events, arousal events, or sleep stage classifications every 30 seconds, as opposed to aggregate metrics, which summarize statistics across an entire night:

- Confusion matrix: Quantifies how well two raters agree when classifying events. For PSG data scored by multiple human experts, we also use confusion matrices to compare model-expert agreement against inter-expert agreement [48–50]. Please note that similar to any overlap was considered a true positive.
- Cohen's κ: Quantifies the level of agreement between two raters, while accounting for the possibility of chance agreement. We calculate κ both among all pairs of experts and between each expert and the model outputs [51–54].
- Receiver operating characteristic (ROC) curves and precision-recall curves (PR): Quantify the ability of a model to discriminate between classes. We compute ROC and PR curves for each expert's scoring regarded as ground truth. We also calculate "Experts Under the Curve (EUC)": The percentage of experts' ROC or PR operating points that fall below the model's curve. EUC measures the degree to which a model matches or exceeds expert performance [55–57].

Aggregate metrics.

We also use aggregate metrics, including the arousal index, apnea-hypopnea index (AHI), and limb movement index, which summarize the frequency of specific events (events per hour of sleep). These metrics count the number of events, regardless of their timing or duration, and are not event-level measures. These aggregated measures are evaluated using the intraclass correlation coefficient (ICC) between experts with experts and between experts with algorithm. Additionally, we conducted a Bland-Altman analysis to evaluate the agreement between CAISR and expert annotations across different datasets. This method assesses systematic bias by comparing the mean difference between CAISR and expert annotations against their average values. It helps identify potential over- or underestimation of event counts by the model and highlights variations in agreement across different event types.

Statistical analysis

To estimate the precision of both event-level and aggregate performance metric measurements, we calculated 95% confidence intervals (CIs) using 10,000 rounds of bootstrapping. We employed the following criteria to summarize the performance of CAISR compared to expert raters:

- **Superiority:** The entire 95% CI of the algorithm-expert metric exceeds that of the expert-expert metric.
- **Equivalence (Non-inferiority):** The 95% CI of the algorithm-expert metric overlaps with that of the expert-expert metric.
- **Inferiority:** The entire 95% CI of the algorithm-expert metric lies below that of the expert-expert metric.

Results

Overall performance

Figure 2 summarizes overall agreement results between CAISR and human scorers across cohorts and tasks.

For sleep staging (**Figure 2A**), the median agreement with experts exceeded 0.8 for single-scored datasets. CAISR outperformed experts across all multi-scored datasets and demonstrated higher agreement than two experts on the BITS dataset, matching the level of another rater in that dataset.

For arousal detection (**Figure 2B**), CAISR's ICC ranged between 0.6 and 0.72 for single-scored datasets. On the BITS multi-scored data, algorithm-expert agreement was below expert-expert agreement, while CAISR exceeded experts when judged against the platinum-standard, suggesting that experts achieve high agreement in part because of systematic errors which CAISR does not share. The performance of CAISR on arousal detection for the Stanford multi-scored dataset was comparable to human experts.

For respiratory event detection, CAISR's ICC on the apnea-hypopnea index (**Figure 2C**) ranged between 0.68 and 0.89 for single-scored datasets, matched or exceeded experts on the BITS multi-scored dataset, and on the Stanford multi-scored data was comparable to one expert and inferior to two.

For limb movement detection (**Figure 2D**), CAISR had an ICC of 0.36 for single-scored datasets and showed varying levels of agreement with experts on the BITS and Stanford datasets, with κ values ranged from 0.40 to 0.52, compared to an inter-expert agreement of 0.87 for the BITS dataset and 0.76 for the Stanford dataset. When evaluated against the platinum labels, CAISR

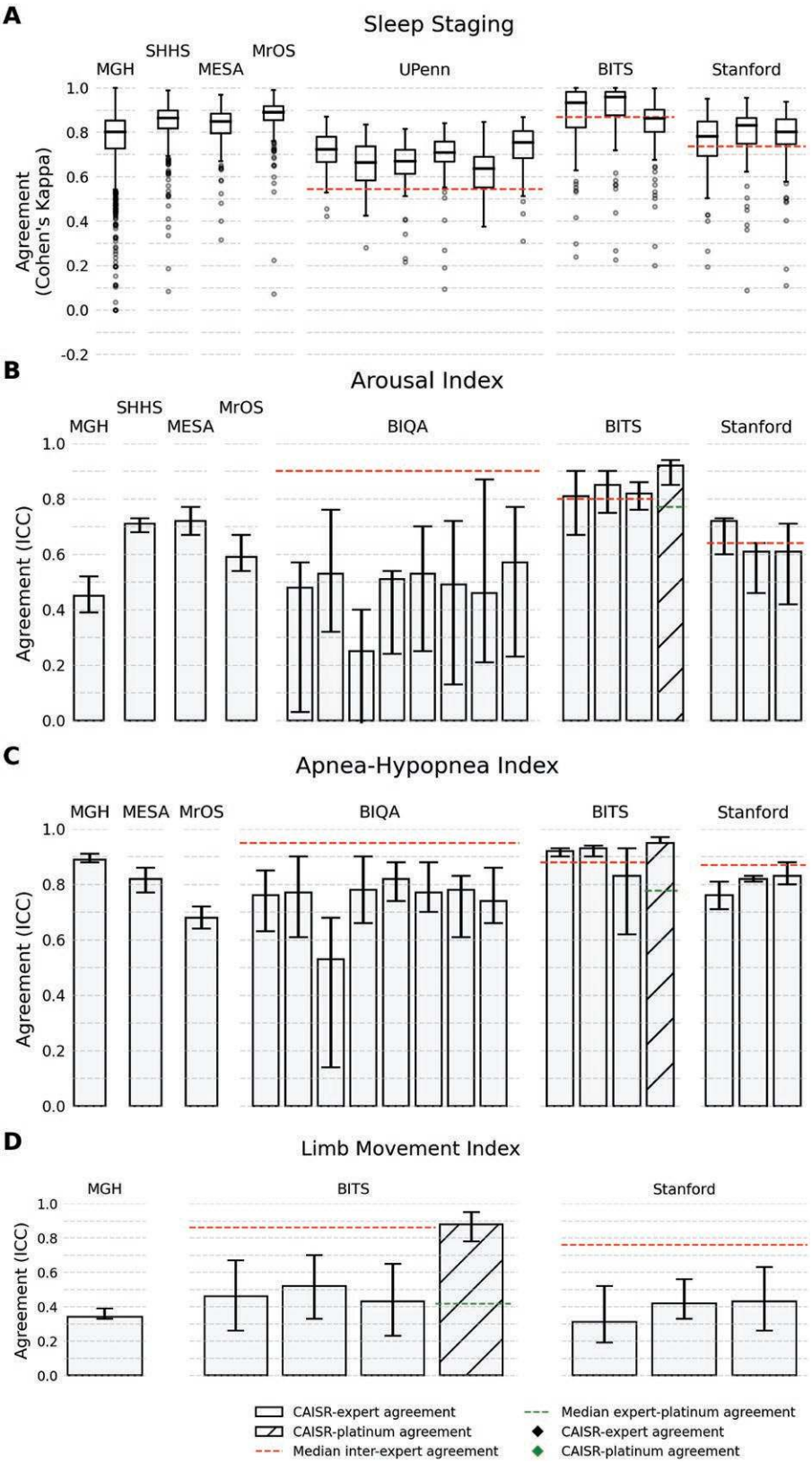


Figure 2. Summary of CAISR results across all cohorts and tasks **(A)** Kappa for CAISR sleep staging across all cohorts, showing the distribution for each subject within each cohort. Boxplots depict data distribution, with boxes extending from the first (Q1) to the third quartile (Q3), a median line, whiskers up to 1.5x the interquartile range (IQR), and fliers beyond the whiskers. **(B)** CAISR results across all cohorts, detailing performance in detecting arousal events **(C)** CAISR results across all cohorts for detecting apnea events **(D)** CAISR results across all cohorts for detecting limb movements. Bar plots show median ICC values with 95% confidence intervals, comparing CAISR performance against experts. The hatched-filled bars represent results against the platinum labels.

significantly outperformed the experts, achieving a κ of 0.89, while median expert-platinum agreement is 0.41. This indicates that CAISR is more accurate in detecting limb movements, while the expert agreement is influenced by systematic errors that CAISR overcomes.

To assess the impact of participant covariates (e.g. age, sex, and clinical indicators) on model performance, we conducted separate univariate ordinary least-squares regression analyses. We categorized ICD codes into clinically relevant groups (e.g. cerebral degeneration, headache disorders) based on our classification schema (see Table S1 in the Supplemental Material). In each model, the mean Cohen's kappa (expert-agreement score) served as the dependent variable, while each covariate (plus an intercept) was included as a predictor. This analysis was performed across four tasks (stage, respiration, arousal, limb) and multiple cohorts (BITS, Penn, Stanford). To correct for multiple comparisons, we applied the Benjamini–Hochberg procedure to control the false discovery rate, considering adjusted p-values < 0.05 statistically significant. Out of the 78 regressions performed (four tasks × cohorts × covariates), 12 exhibited a nominally significant p-value before correction. However, after applying the Benjamini–Hochberg procedure, only five comparisons remained significant. Notably, age showed a small but statistically significant positive association (corrected $p < 0.05$) with model performance (respiratory-event classification) in both the BITS and Stanford cohorts, as well as with arousal classification in the BITS cohort. In the Stanford data, cardiovascular conditions and diabetes were additional significant covariates. The corresponding coefficients were modest in magnitude (e.g. 0.002–0.005 for age, 0.13 and 0.19 for cardiovascular and diabetic conditions respectively), and the R^2 values for these regressions did not exceed 0.21. All other covariates, including sex, Epworth sleepiness scale (ESS), and various clinical indicators, failed to reach significance once multiple-testing corrections were applied. Overall, these results suggest that although certain participant characteristics (especially age) may influence model performance, the strength of these effects is relatively small. All results are reported in Table 2. Stratified analyses revealed three key deviations: (i) respiratory-event κ was lower in the youngest age tertile at Stanford, (ii) respiratory performance was modestly reduced in females within the Stanford cohort, and (iii) sleep-staging κ was higher in Black compared to White participants in the UPenn cohort. All other strata—age, Epworth Sleepiness Scale (ESS), sex, race, and comorbidities—showed no significant differences, highlighting the model's overall robustness across diverse demographic and clinical groups (see Tables S2–S6 in the Supplementary Material).

Detailed performance evaluation on multi-scored datasets

Sleep staging.

Figures 3 presents inter-rater reliability analysis (IRR) results of CAISR sleep staging on the UPenn, Stanford, and BITS datasets, compared to human experts. The top section shows ROC and PR curves, with CAISR achieving AUC-ROC values ranging from 0.82 to 0.97 and AUC-PR values between 0.63 and 0.9. The bottom section shows a bar plot summarizing the median κ agreement: the blue bars represent the median Kappa agreement between rater pairs across different sleep stage, while the black bars represent the median Kappa agreement between CAISR and each human scorer. 95% confidence intervals were calculated using 10,000 bootstrap samples. On all datasets, CAISR outperformed experts (CAISR ≥ Experts). The heatmaps of Cohen's κ values (**Figures S2A**, **S3A, and S4A** in the Supplementary Material) show substantial agreement between CAISR and experts, with κ values ranging from 0.66 to 0.91 across datasets, similar to experts' agreement with each other. Box plots (**Figures S2B**, **S3B, and S4B** in the Supplementary Material) highlight CAISR's narrower distribution of κ values, underscoring the model's reliability, while the experts display more variability. The ROC and precision-recall curves (Supplementary **Figures S2C**, **S3C, and S4C**) for different sleep stages (Wake, REM, N1, N2, N3) reveal CAISR's high AUC values (ranging from 0.82 to 0.99) and AUC-PR values (from 0.39 to 0.97), along with high EUC values, indicating excellent accuracy and reliability.

Arousal detection.

Figure 4 presents IRR results for arousal event detection, overall and stratified by sleep stage. The top section shows ROC and PR curves across sleep stages for the BITS and Stanford multi-scored datasets. The bottom section shows a bar plot for the median agreement (κ) between CAISR and experts. In both datasets, the performance of CAISR and experts is comparable (AI ~ Experts). These findings indicate that CAISR performs similarly to human experts in detecting arousal events, across sleep stages and datasets. Supplementary **Figures S5** (BITS Cohort) and **S6** (Stanford Cohort) in the supplementary material present the detailed IRR analysis of the CAISR arousal detection model compared to multiple human experts across two datasets. In Panels **S5A** and **S6A**, the distributions of Cohen's κ values across subjects show that the agreement between the platinum label and CAISR of 0.78 [0.56, 0.91] matching the expert—platinum label performance range of 0.76-0.79 for the BITS dataset. The Cohen's κ values for expert-CAISR and expert-expert comparisons on the Stanford Cohort are overlapping with the respective ranges of 0.61-0.67 and 0.63-0.77. Panels **S5B** and **S6B** show the subject-wise Cohen's κ distributions, with CAISR outperforming experts when compared to the platinum labels on the BITS cohort and showing on-par performance on the Stanford Cohort. Panels **S5C** and **S6C** show the ROC and PR curves per Sleep stage for arousal detection, displaying an area under the curve ranging from 0.96-0.98 in the ROC curves and 0.74–0.88 for the PRC curves of the BITS Cohort. The area under the curve for the Stanford cohort ranged respectively from 0.81–0.98 and 0.48–0.82, all displaying comparable performance between CAISR and the experts. Panels **S5D** and **S6D** show the intra-class correlation coefficient for the Arousal index. **S5D** shows a higher platinum label—expert agreement. The platinum-CAISR agreement of 0.92 [0.85,0.92] overlaps with the platinum-expert range of 0.81–0.94 for the BITS Cohort. For the Stanford Cohort, expert-CAISR performance (0.61–0.72) overlaps with expert–expert performance (0.62–0.68).

Supplementary Figure S11 and S12 present inter-rater reliability analyses for arousal detection across two independent PSG datasets (BITS: $N = 98$; Stanford: $N = 100$), comparing three human experts with the CAISR Arousal detector. In these figures, we employ a sample-based approach, treating each data point at the original sampling frequency as an independently labeled instance. Supplementary Figure S11A shows the confusion matrix quantifying inter-rater agreement for arousal detection in BITS PSG recordings ($N = 98$) using Cohen's Kappa (95% CI), while **S12A** presents analogous results for the Stanford dataset ($N = 100$). Both **S11B** and **S12B** display the distribution of Cohen's Kappa values when alternately treating each rater (including CAISR) as ground truth, demonstrating consistent agreement patterns across datasets. The ROC/PR curves in **S11C** and **S12C** further validate CAISR's detection

performance, with the sample-based analysis confirming reliable arousal detection at the original sampling resolution. Together, these results establish CAISR's robustness as an automated detector across independent PSG datasets.

Respiratory event detection and classification

Figure 5 compares CAISR and human experts in detecting respiratory events, including obstructive apnea (OA), central apnea (CA), mixed apnea (MA), hypopnea (HY), and respiratory effort-related arousal (RERA). The top section presents ROC and PR curves for the BITS and Stanford datasets. The bottom section compares the reliability of CAISR and experts using median agreement (κ). On the Stanford dataset, CAISR-expert agreement is lower than expert-expert agreement (CAISR < Experts). On the platinum-labeled BITS dataset, CAISR and experts show comparable reliability with overlapping confidence intervals (CAISR ~ Experts), indicating non-inferiority.

Table 2. Effects of participant covariates on model performance across tasks and cohorts

Task	Cohort	Covariate	Coefficient	Model fit (R^2)	*p*-value (uncorrected)	*p*-value	Significant
Stage	BITS	age	0.001	0.023	0.136	0.483	FALSE
	BITS	sex	−0.027	0.009	0.364	0.804	FALSE
	BITS	ess	0.001	0.002	0.634	0.908	FALSE
	BITS	Sleep Disorders	−0.082	0.025	0.119	0.442	FALSE
	BITS	Neuropsychiatric Disorders	0.0	0.0	0.99	0.999	FALSE
	BITS	Respiratory Conditions	−0.045	0.008	0.37	0.804	FALSE
	BITS	Cardiovascular Conditions	−0.023	0.005	0.49	0.852	FALSE
	BITS	Endocrine/Metabolic Conditions	−0.033	0.008	0.379	0.804	FALSE
	BITS	Neurodegenerative Disorders	−0.043	0.009	0.351	0.804	FALSE
	BITS	Other Neurological Conditions	0.008	0.001	0.82	0.959	FALSE
	BITS	Headache & Migraine Disorders	−0.026	0.003	0.603	0.908	FALSE
	PENN	age	0.002	0.011	0.39	0.804	FALSE
	PENN	ess	−0.004	0.066	0.033	0.286	FALSE
	Stanford	age	0.0	0.001	0.75	0.932	FALSE
	Stanford	sex	0.032	0.017	0.204	0.637	FALSE
	Stanford	Neuropsychiatric Disorders	−0.019	0.005	0.502	0.852	FALSE
	Stanford	Cardiovascular Conditions	−0.025	0.008	0.368	0.804	FALSE
	Stanford	Endocrine/Metabolic Conditions	−0.051	0.018	0.184	0.597	FALSE
	Stanford	Neurodegenerative Disorders	0.005	0.0	0.895	0.968	FALSE
	Stanford	Other Neurological Conditions	−0.043	0.03	0.088	0.428	FALSE
	Stanford	Headache & Migraine Disorders	−0.001	0.0	0.966	0.999	FALSE
Resp	**BITS**	**age**	**0.004**	**0.16**	**0.0**	**0.002**	**TRUE**
	BITS	sex	0.058	0.027	0.109	0.442	FALSE
	BITS	ess	−0.003	0.007	0.424	0.804	FALSE
	BITS	Sleep Disorders	0.148	0.056	0.02	0.224	FALSE
	BITS	Neuropsychiatric Disorders	−0.02	0.002	0.629	0.908	FALSE
	BITS	Respiratory Conditions	−0.035	0.003	0.572	0.908	FALSE
	BITS	Cardiovascular Conditions	0.058	0.022	0.15	0.509	FALSE
	BITS	Endocrine/Metabolic Conditions	−0.011	0.001	0.802	0.959	FALSE
	BITS	Neurodegenerative Disorders	0.026	0.002	0.646	0.908	FALSE
	BITS	Other Neurological Conditions	0.016	0.002	0.699	0.908	FALSE
	BITS	Headache & Migraine Disorders	−0.053	0.008	0.383	0.804	FALSE
	Stanford	**age**	**0.005**	**0.21**	**0.0**	**0.0**	**TRUE**
	Stanford	sex	0.07	0.04	0.048	0.312	FALSE
	Stanford	Neuropsychiatric Disorders	0.075	0.038	0.055	0.328	FALSE
	Stanford	**Cardiovascular Conditions**	**0.127**	**0.104**	**0.001**	**0.022**	**TRUE**
	Stanford	**Endocrine/Metabolic Conditions**	**0.186**	**0.118**	**0.0**	**0.013**	**TRUE**
	Stanford	Neurodegenerative Disorders	−0.02	0.002	0.694	0.908	FALSE
	Stanford	Other Neurological Conditions	0.025	0.005	0.49	0.852	FALSE
	Stanford	Headache & Migraine Disorders	−0.072	0.026	0.113	0.442	FALSE

Table 2. Continued

Task	Cohort	Covariate	Coefficient	Model fit (R^2)	*p*-value (uncorrected)	*p*-value	Significant
Arousal	**BITS**	**age**	**0.002**	**0.088**	**0.003**	**0.049**	**TRUE**
	BITS	sex	0.025	0.008	0.369	0.804	FALSE
	BITS	ess	−0.005	0.033	0.074	0.412	FALSE
	BITS	Sleep Disorders	0.008	0.0	0.865	0.963	FALSE
	BITS	Neuropsychiatric Disorders	−0.025	0.006	0.433	0.804	FALSE
	BITS	Respiratory Conditions	0.032	0.005	0.499	0.852	FALSE
	BITS	Cardiovascular Conditions	0.032	0.011	0.308	0.804	FALSE
	BITS	Endocrine/Metabolic Conditions	−0.004	0.0	0.91	0.968	FALSE
	BITS	Neurodegenerative Disorders	−0.019	0.002	0.666	0.908	FALSE
	BITS	Other Neurological Conditions	0.003	0.0	0.919	0.968	FALSE
	BITS	Headache & Migraine Disorders	−0.074	0.026	0.113	0.442	FALSE
	Stanford	age	0.001	0.007	0.424	0.804	FALSE
	Stanford	sex	0.056	0.043	0.038	0.3	FALSE
	Stanford	Neuropsychiatric Disorders	−0.005	0.0	0.856	0.963	FALSE
	Stanford	Cardiovascular Conditions	0.01	0.001	0.753	0.932	FALSE
	Stanford	Endocrine/Metabolic Conditions	−0.006	0.0	0.891	0.968	FALSE
	Stanford	Neurodegenerative Disorders	−0.104	0.073	0.007	0.091	FALSE
	Stanford	Other Neurological Conditions	0.013	0.002	0.627	0.908	FALSE
	Stanford	Headache & Migraine Disorders	−0.06	0.031	0.081	0.419	FALSE
Limb	BITS	age	0.003	0.053	0.024	0.232	FALSE
	BITS	sex	0.103	0.042	0.045	0.312	FALSE
	BITS	ess	0.003	0.004	0.52	0.863	FALSE
	BITS	Sleep Disorders	0.04	0.002	0.66	0.908	FALSE
	BITS	Neuropsychiatric Disorders	0.001	0.0	0.991	0.999	FALSE
	BITS	Respiratory Conditions	−0.019	0.001	0.824	0.959	FALSE
	BITS	Cardiovascular Conditions	0.03	0.003	0.598	0.908	FALSE
	BITS	Endocrine/Metabolic Conditions	−0.027	0.002	0.68	0.908	FALSE
	BITS	Neurodegenerative Disorders	0.036	0.002	0.65	0.908	FALSE
	BITS	Other Neurological Conditions	−0.016	0.001	0.785	0.957	FALSE
	BITS	Headache & Migraine Disorders	−0.135	0.026	0.118	0.442	FALSE
	Stanford	age	0.001	0.007	0.418	0.804	FALSE
	Stanford	sex	0.06	0.011	0.294	0.804	FALSE
	Stanford	Neuropsychiatric Disorders	0.0	0.0	0.999	0.999	FALSE
	Stanford	Cardiovascular Conditions	0.012	0.0	0.847	0.963	FALSE
	Stanford	Endocrine/Metabolic Conditions	0.072	0.007	0.415	0.804	FALSE
	Stanford	Neurodegenerative Disorders	0.097	0.014	0.238	0.715	FALSE
	Stanford	Other Neurological Conditions	0.062	0.012	0.279	0.804	FALSE
	Stanford	Headache & Migraine Disorders	0.025	0.001	0.731	0.932	FALSE

In the BITS cohort (**Figure S7A** in the Supplementary Material), CAISR achieved an overall Cohen's κ of 0.74 [0.53–0.83] when compared to platinum-labeled data, closely aligning with expert performance values (range: 0.76–0.79). In the Stanford cohort (Supplementary **Figure S8A**), CAISR's median Cohen's κ was 0.58 [0.13–0.73], slightly lower but still comparable to experts. **Figures S7B and S8B** in the Supplementary Material highlight the inter-subject agreement between CAISR and experts, showing a strong agreement between CAISR and experts in both cohorts. For **Figure S7C** and **Figure S8C** in the Supplementary Material, the ROC curves show that both CAISR and experts' operating points are tightly clustered near the upper left corner, indicating high sensitivity and specificity across event types such as obstructive apnea, central apnea, and hypopnea. Similarly, in the PR curves, the CAISR black square is near the upper right corner, closely aligned with the experts' markers, signifying comparable precision and recall. The exception is in the detection of mixed apnea, where CAISR shows slightly lower PR performance, and RERA detection, where accuracy varies more widely. The high intraclass correlation coefficients for AHI, approximately 0.95 in the BITS

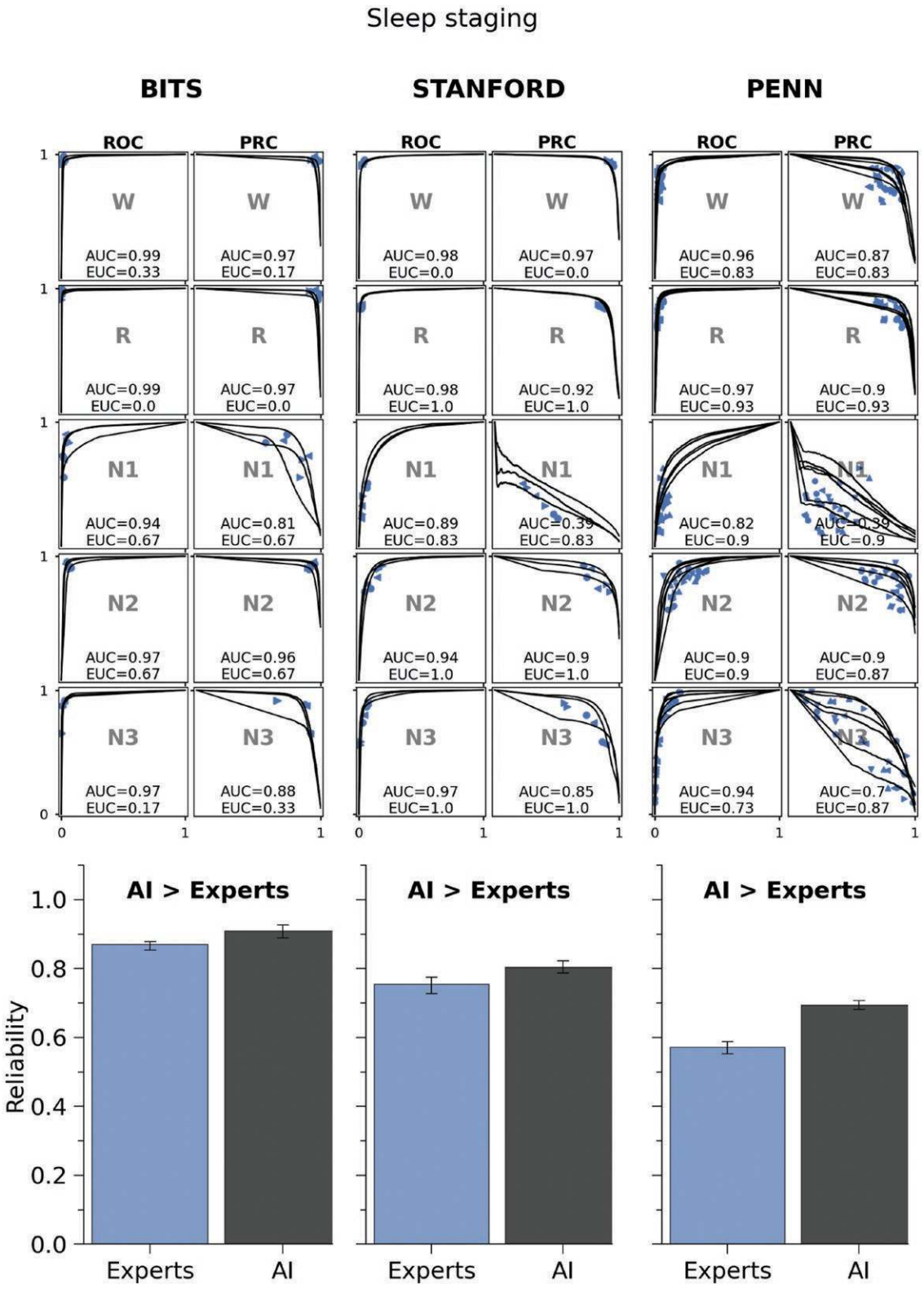


Figure 3. Inter-rater reliability analysis results for sleep staging. Top: ROC and precision-recall curves for the CAISR sleep staging model and six experts across different sleep stages (Wake, REM, N1, N2, N3). The CAISR model achieves high AUC values (0.82 to 0.97) and AUC-PR values (0.63 to 0.9), indicating robust performance. Bottom: Barplot summarizing the median Kappa agreement between rater pairs across different sleep per dataset. 95% confidence intervals were computed using 10,000 bootstrap samples. Non-overlapping CIs indicate superiority of either the AI model or the expert, while overlapping CIs suggest non-inferiority of the AI model compared to the expert.

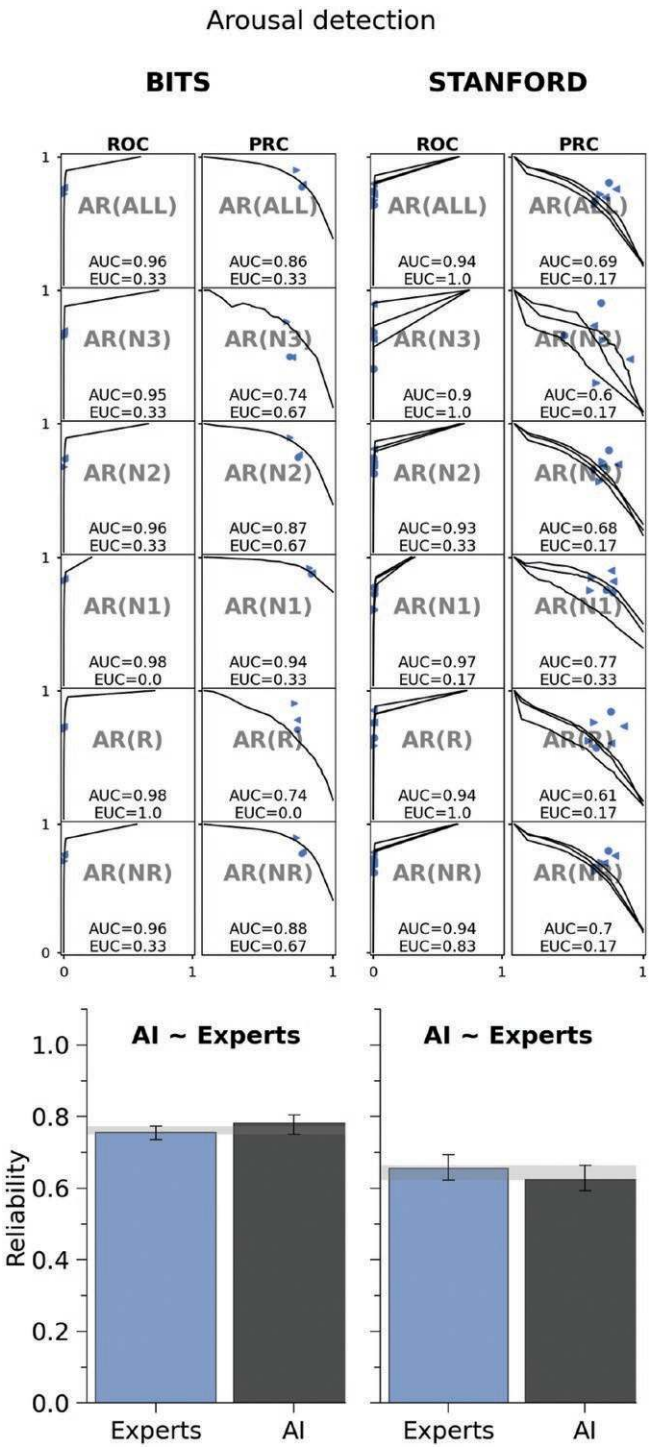


Figure 4. Inter-rater reliability analysis results for arousal event detection. Top: ROC and precision-recall curves for the CAISR arousal event detection model, overall and stratified by sleep stage. Bottom: Barplot summarizing the median Kappa agreement between rater pairs across different arousal classes per dataset. 95% confidence intervals were computed using 10,000 bootstrap samples. Non-overlapping CIs indicate superiority of either the AI model or the expert, while overlapping CIs suggest non-inferiority of the AI model compared to the expert.

cohort (**Figure S7D** in the Supplementary Material) and 0.8 in the Stanford cohort (**Figure S8D** in the Supplementary Material), further support CAISR's comparable performance to expert scorers.

Supplementary Figures S13 and S14 show the inter-rater reliability analysis for respiratory event detection in BITS ($N = 98$) and Stanford ($N = 100$) PSG recordings, respectively. Both figures demonstrate: (1) moderate to strong agreement between three human experts and the CAISR breathing detector (Cohen's $\kappa = 0.33$–0.67), with CAISR maintaining $\kappa \geq 0.39$ against all experts; (2) consistent performance when alternately treating each rater as reference (Panel B); and (3) comparable detection capability through ROC/PR curve analyses (Panel C). The sample-based approach, analyzing each data point at original sampling frequency, confirms CAISR's reliability as an automated respiratory event detector across independent datasets.

Limb movement detection task

Figure 6 compares CAISR with experts in detecting limb movements. The top section shows ROC and PR curves for the BITS and Stanford datasets. On the Stanford dataset, expert-expert agreement is higher than expert-CAISR agreement. However, on the platinum-labeled BITS data, CAISR outperformed experts (AI > Experts).

Supplementary **Figures S9 and S10** in the supplementary material present the inter-rater reliability results for CAISR on limb movement detection across the BITS and Stanford datasets, respectively. In both cases, CAISR shows moderate agreement with most experts, with Cohen's κ values ranging from 0.43 to 0.58, as seen in Supplementary **Figures S9A and S10A**. The distribution of κ values across subjects (Supplementary **Figures S9B and 10B**) reveals a slightly lower median agreement between CAISR and the experts on the Stanford dataset. The ROC and PR curves (Supplementary **Figures S9C and S10C**) also indicate that CAISR's performance is comparable to that of the experts, albeit with mildly lower agreement for individual limb movement events. The intraclass correlation coefficient (ICC) values (Supplementary **Figures S9D and S10D**) further confirm moderate reliability, with values ranging from 0.43 to 0.52.

The discrepancy in IRR between data with standard expert labels vs platinum labels underscores a critical issue: human raters are systematically poor at accurately marking limb movements. The tedious nature of manually reviewing full PSGs makes limb movement detection prone to variability and systematic errors. As shown in Supplementary **Figure S18**, experts frequently miss events or incorrectly mark boundaries. CAISR's rule-based model demonstrates higher accuracy and consistency in identifying the start and end points of limb movement events. Thus, the lower performance does not represent a weakness but reflects the difficulty of the task and limitations of human raters.

The Bland-Altman analysis in Supplementary Figure S22 revealed that while agreement for respiratory events and arousals remained relatively stable, discrepancies increased at higher event counts, suggesting a potential bias in detecting frequent events. Limb movement detection exhibited greater variability, likely due to differences in scoring methodologies (movements near respiratory events depend on the accuracy of respiratory annotation) across datasets, including variations in event definitions and thresholds used by different annotators.

Competing models

We evaluated CAISR against published state-of-the-art models using the BITS and Stanford triple-scored datasets; the results are presented in Table 3 for BITS and Table 4 for Stanford, respectively. On the BITS dataset, CAISR outperformed the U-Sleep [22] on sleep staging (κ: 0.88–0.90 vs. 0.70–0.71), WaveNet [40]

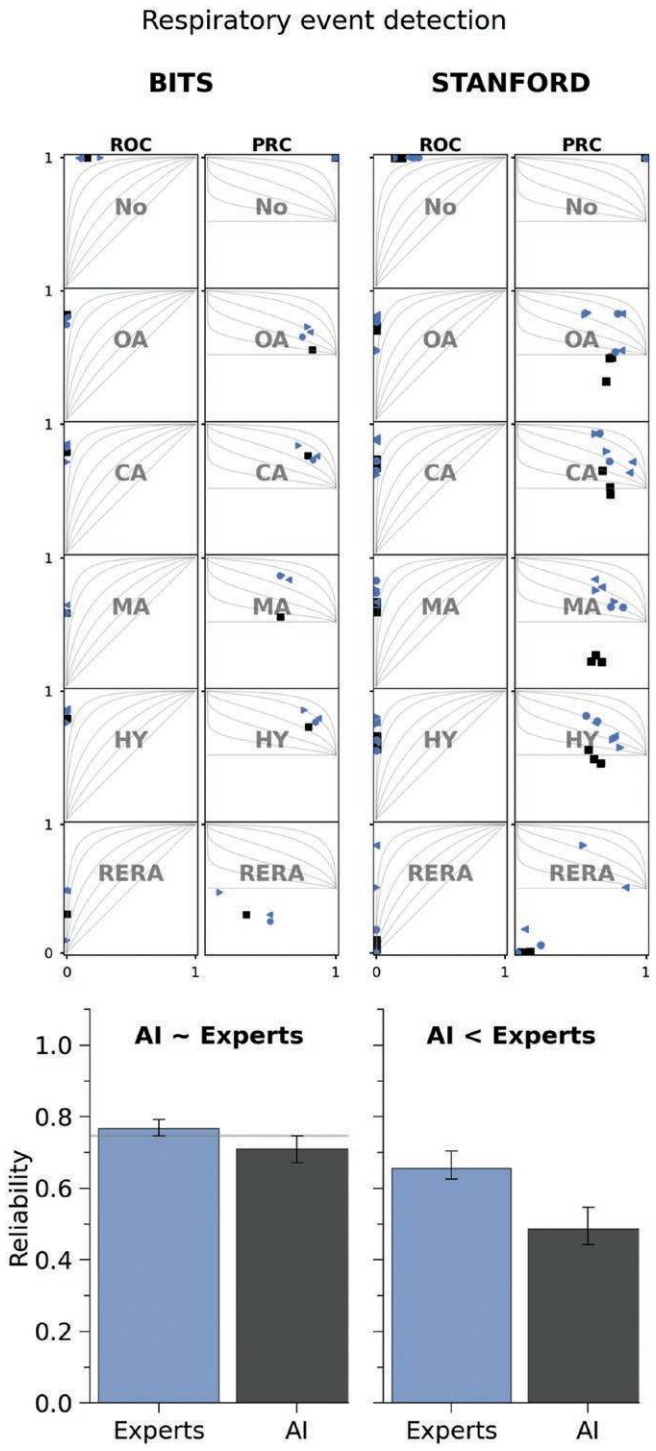


Figure 5. Inter-rater reliability analysis results for respiratory event detection. Top: ROC and precision-recall curves for the CAISR respiratory event detection model per respiratory class (obstructive apnea, central apnea, mixed apnea, hypopnea, RERA). Bottom: Barplot summarizing the median Kappa agreement between rater pairs across different apnea classes per dataset. 95% confidence intervals were computed using 10,000 bootstrap samples. Non-overlapping CIs indicate superiority of either the AI model or the expert, while overlapping CIs suggest non-inferiority of the AI model compared to the expert.

on respiratory event detection (ICC: 0.83–0.93 vs. 0.38–0.50), and the Multimodal Arousal Detector (MAD) by Brink-Kjaer et al [58] on arousal detection (ICC: 0.81–0.85 vs. 0.35–0.56). For limb movement detection, CAISR achieved comparable results on the BITS dataset to the Ferri Original model [42] (ICC: 0.45–0.52 vs. 0.46–0.51). On the Stanford dataset, CAISR demonstrated superior performance in sleep staging (κ: 0.77–0.80 vs. 0.74–0.76) and respiratory event detection (ICC: 0.76–0.83 vs. 0.62–0.71). For limb movement detection the original Ferri model was superior (ICC: 0.31–0.43 vs. 0.55–0.68). As platinum labels for the Stanford dataset are not available, we randomly selected 10 examples to investigate why CAISR shows lower performance compared to the original Ferri algorithm [42, 45]. As shown in Supplementary **Figure S18**, many limb movement annotations are incorrectly placed, and experts often miss events.

Discussion

In this study, we developed and validated CAISR, the first model trained and tested on diverse data sets capable of fully automated comprehensive scoring of clinical polysomnography (PSG). CAISR achieves or exceeds human expert-level performance across all major PSG scoring tasks, including sleep staging, arousal detection, respiratory event detection and classification, and limb movement analysis. By addressing key challenges such as inter-rater variability and the labor-intensive nature of manual scoring, CAISR holds significant potential for clinical application. Its deployment could particularly benefit remote and underserved regions where access to sleep testing services is limited, while also helping to alleviate the workload of sleep specialists in high-demand tertiary care centers. Moreover, CAISR's scalability can advance sleep research by enabling the analysis of sleep on a larger scale than ever before.

The design and validation of CAISR were undertaken with care to avoid common sources of bias and error [54, 59, 60]. We employed a large and diverse dataset comprising 25,749 PSGs, a subset annotated by seven independent human experts representing a wide range of health conditions and pathologies. This rigorous methodology ensures that CAISR's generalizability to real-world clinical scenarios is robust. The validation dataset, independent from the training data, included 3,860 individuals and was representative of diverse clinical conditions and populations. Additionally, the experts who participated in the validation process were different from those involved in the model's development, further reducing potential biases. Platinum-standard labels, developed for each task, helped address systematic errors commonly seen in clinical annotations. The model was tested using multi-center datasets, with data scored by different experts and using various equipment configurations. To enhance its adaptability, we explored model fine-tuning with small sets of target-domain samples (the BITS dataset), improving CAISR's ability to conform to specific clinical environments. Specifically, we applied this method on the BITS dataset, using a two-fold cross-validation strategy where each fold included 49 subjects. This was done across all tasks—sleep staging, arousal, apnea, and limb movement detection. After fine-tuning, CAISR's performance improved by 2–5%, allowing it to surpass human experts, despite the high agreement among the three independent human raters in the BITS dataset. While this approach helps CAISR adapt to different clinical environments, it also raises the question of whether matching lab-specific scoring is always the best practice. Thus, fine-tuning not only provides a clear performance boost but also highlights the need for careful consideration of standardization across clinical scoring. Although performance showed some variation across different cohorts, there was no consistent

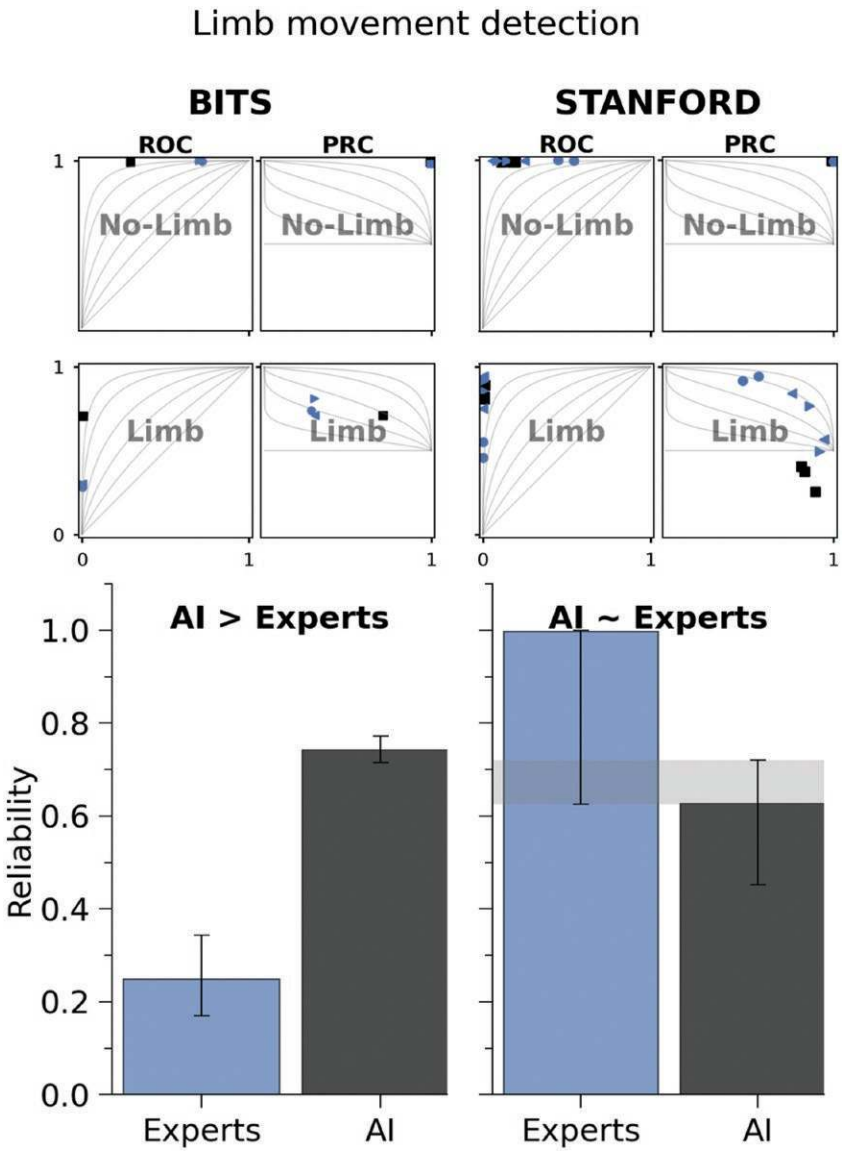


Figure 6. Summary results limb movement detection. Top: ROC and precision-recall curves for the CAISR limb movement event detection model per limb movement class (limb movement and no-limb movement). Bottom: Barplot summarizing the median Kappa agreement between rater pairs across different limb movement classes per dataset. 95% confidence intervals were computed using 10,000 bootstrap samples. Non-overlapping CIs indicate superiority of either the AI model or the expert, while overlapping CIs suggest non-inferiority of the AI model compared to the expert.

evidence of superior or inferior performance in laboratory-based versus home-based PSG data. The lack of a standardized set of expert scorers across these cohorts complicates direct comparisons. Additionally, variable inter-rater agreement underscores the difficulty of establishing a uniform ground truth in different scoring environments. Overall, our findings did not reveal a systematic advantage or disadvantage for CAISR in either setting.

CAISR is the first AI system to offer a fully automated, clinically relevant assessment of PSG data, providing more nuanced and complete analyses than previous models. It detects and categorizes sleep-disordered breathing events, aiding in the diagnosis of different subtypes of sleep apnea. Furthermore, CAISR's ability to identify arousals and limb movements offers crucial insights into sleep quality and potential causes of sleep fragmentation. This level of granularity and comprehensiveness has not been achieved by any prior system, making CAISR a groundbreaking tool in automated sleep analysis.

A noteworthy finding in this study was the identification of substantial systematic errors in clinical scoring, particularly in the detection of arousals, breathing events, and limb movements. This was evident when comparing expert-expert agreement against expert-CAISR agreement and by assessing CAISR's performance using platinum-standard datasets. These errors likely result from the challenges inherent in manually scoring lengthy PSG recordings, a process that is both time-consuming and prone to human fatigue. Based on discussions with clinicians, this likely reflects the idea that human scorers do not focus on precisely measuring the start and end times; instead, they aim to ensure that the event is considered "counted," regardless of its exact timing. While inter-expert agreement on standard clinical datasets was high, CAISR initially showed lower agreement with experts. However, when evaluated against platinum-standard labels—more accurate annotations—CAISR matched or exceeded expert performance. This discrepancy reveals a critical

Table 3. Comparison of competing models vs. CAISR across tasks for the BITS dataset

		Kappa (Sleep Staging), ICC (Resp, Arousal, Limb)		
		Expert 1	**Expert 2**	**Expert 3**
Stage	CAISR	0.88 (0.85, 0.90)	0.90 (0.87, 0.92)	0.84 (0.81, 0.86)
	U-Sleep[22]	0.70 (0.67, 0.73)	0.71 (0.69, 0.74)	0.70 (0.66, 0.73)
Resp	CAISR	0.92 (0.86, 0.94)	0.93 (0.89, 0.95)	0.83 (0.77, 0.90)
	WaveNet[40]	0.38 (0.18, 0.52)	0.39 (0.31, 0.51)	0.50 (0.31, 0.59)
Arousal	CAISR	0.81 (0.78, 0.90)	0.85 (0.74, 0.90)	0.82 (0.71, 0.87)
	MAD[59]	0.35 (0.32, 0.45)	0.36 (0.25, 0.44)	0.56 (0.45, 0.66)
Limb	CAISR	0.46 (0.24, 0.66)	0.52 (0.42, 0.58)	0.45 (0.26, 0.65)
	Ferri's Model[43]	0.46 (0.29, 0.75)	0.51 (0.32, 0.77)	0.46 (0.29, 0.68)

Table 4. Comparison of competing models vs. CAISR across tasks for the Stanford dataset

		Kappa (Sleep Staging), ICC (Resp, Arousal, Limb)		
		Expert 1	**Expert 2**	**Expert 3**
Stage	CAISR	0.77 (0.75, 0.78)	0.80 (0.79, 0.82)	0.79 (0.77, 0.80)
	U-Sleep[22]	0.75 (0.73, 0.76)	0.75 (0.74, 0.76)	0.74 (0.73, 0.76)
Resp	CAISR	0.76 (0.71, 0.79)	0.82 (0.76, 0.87)	0.83 (0.76, 0.91)
	WaveNet[40]	0.71 (0.55, 0.78)	0.62 (0.51, 0.73)	0.66 (0.50, 0.76)
Arousal	CAISR	0.64 (0.50, 0.71)	0.52 (0.32, 0.63)	0.62 (0.50, 0.78)
	MAD[59]	N/A	N/A	N/A
Limb	CAISR	0.31 (0.24, 0.53)	0.42 (0.26, 0.66)	0.43 (0.28, 0.69)
	Ferri's Model[43]	0.55 (0.43, 0.66)	0.65 (0.38, 0.75)	0.68 (0.45, 0.75)

insight: high inter-expert agreement may reflect shared biases rather than actual accuracy. CAISR's superior performance on the platinum-labeled data suggests that it overcomes human limitations, particularly in tasks that are tedious and error-prone. This has significant implications for sleep medicine, as CAISR could enhance the reliability and quality of sleep study interpretations, especially in high-volume clinical settings where human expertise is stretched thin. Besides, the superior performance of rule-based systems for respiratory event and limb movement detection can be attributed to their ability to leverage domain-specific knowledge and well-defined heuristics. These systems are designed to capture specific patterns in physiological signals that are strongly associated with events such as apnea, hypopnea, or limb movements, based on expert-defined thresholds. rule-based systems provide more transparency and interpretability, as the decision-making process is based on explicit, predefined criteria, which makes it easier to understand why a particular event is detected.

CAISR stands out from previous approaches, which typically focus on isolated tasks, such as sleep staging [16, 22, 28, 29], arousal detection [43, 58, 61], or apnea identification [40, 62, 63]. By integrating all major sleep scoring tasks into a single unified system, CAISR represents a comprehensive solution for automated sleep analysis. One of its key strengths is the use of an exceptionally large and diverse dataset of over 18,017 PSG recordings for training, allowing it to generalize effectively across different populations and clinical environments. It is important to consider that, ideally, the models should be periodically retrained or fine-tuned with new data to ensure they remain up-to-date and maintain accuracy. The frequency of updates could depend on the volume of new data, with updates occurring at regular intervals (e.g. quarterly or annually) or triggered when a certain amount of newly labeled data is available. This approach would help ensure the model stays aligned with current trends and patterns in the data. Moreover, CAISR has undergone extensive validation on multiple independent datasets, including platinum-labeled datasets that address the systematic biases of traditional PSG scoring. This rigorous validation, often lacking in other models, underscores CAISR's readiness for clinical adoption. Another notable innovation is the systematic comparison of expert-expert agreement versus expert-CAISR agreement, along with an objective comparison against platinum-standard labels. This thorough methodology ensures that CAISR's performance is assessed accurately, without merely reproducing the biases of routine clinical practices.

Despite its strengths, CAISR has several limitations that warrant further exploration. First, pediatric PSGs were not included in this study; ongoing work aims to address this gap. Second, although CAISR performed well across diverse datasets, it has not been systematically evaluated in populations with severe neurological or psychiatric conditions, such as stroke survivors or patients with neurodegenerative diseases. The absence of data from these populations may limit its generalizability to such cases. Third, CAISR's reliance on datasets predominantly from North America raises concerns about potential regional biases. Further investigation is needed to ensure that CAISR performs robustly in underrepresented populations, including those in low-resource settings and across different ethnic groups. Fourth, while CAISR was designed to handle various channel configurations, data quality issues or missing channels could impact its performance. In addition, the manually added 3%/arousal hypopneas for the MGH datasets may have introduced slight variability during the fine-tuning of

the flow reduction detection parameters. To address this, we optimized these parameters separately from those used for detecting desaturation and arousal events. This separation helped maintain robust generalizability across datasets as demonstrated in the BITS dataset. Fifth, while the model architecture was carefully selected based on current best practices, there remains a wide range of deep learning models and transfer learning strategies that could enhance performance further. Sixth, while the base CAISR model demonstrates strong performance, fine-tuning is necessary for improved accuracy on datasets that were excluded during the training of CAISR. Variations in rater bias, measurement systems, and internal training/education methods introduce bias that requires small adjustments to CAISR to align with the scoring styles of different datasets. Future efforts should focus on eliminating the need for fine-tuning to ensure broader applicability while omitting the introduction of the systematic clinical bias as present in non-platinum-labeled datasets. Finally, CAISR follows the American Academy of Sleep Medicine (AASM) scoring guidelines, which have their own limitations. For example, periodic limb movements associated with respiratory events are not scored (and thus is also dependent on precision of scoring of respiratory abnormality), while there is some data suggesting an independent importance of limb movements [64]. Besides, visual scoring of sleep stages must be done in 30-second increments, as this is the standard method used by human scorers, even though the brain state may not change at fixed 30-second intervals. Future development could expand beyond these conventional rules to incorporate more advanced, less constrained analyses of EEG, EMG, respiratory, autonomic, and oximetry data, enhancing its ability to handle diverse sleep data, even in the absence of EEG signals.

Conclusions

In this study, the Complete Artificial Intelligence Sleep Report (CAISR) system demonstrated expert-level performance across all domains of clinical polysomnography scoring, including sleep staging, arousal detection, respiratory event classification, and limb movement identification. CAISR's ability to integrate and automate these tasks marks a significant advancement in the field, offering a reliable and scalable solution for sleep diagnostics. By minimizing inter-rater variability and accelerating the scoring process, CAISR holds promise for improving access to accurate sleep diagnostics in both remote areas and high-volume clinical settings. Moreover, its rigorous validation on diverse datasets underscores its potential to standardize sleep analysis and enhance research on sleep disorders at scale.

Supplementary material

Supplementary material is available at *SLEEP* online.

Disclosure statement

Financial disclosure: This work was supported by grants from the NIH (R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119) and AWS. This work also made use of data from the Sleep Heart Health Study (SHHS), which was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The work also made use of the National Sleep Research Resource, which was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

Non-financial disclosure: Dr. Westover is a co-founder, scientific advisor, and consultant to Beacon Biosignals and has a personal equity interest in the company. Beacon Biosignals provided no support for this work. R.J. Thomas discloses (1) patent and license/royalties from MyCardio, LLC, for the ECG-spectrogram; (2) patent and license/royalties from DeVilbiss-Drive for an auto-CPAP algorithm; and (3) consulting for Jazz Pharmaceuticals, Guidepoint Global, and GLG Councils. Dr. Trotti is a member of the Board of Directors of the American Academy of Sleep Medicine. Any opinions expressed are those of the authors and do not reflect statements of the AASM or its Board.

Data availability

We have made the MGH dataset publicly available at https://bdsp.io/content/hsp/2.0/. All other datasets are available assuming the individual researcher and use-case is eligible for a given dataset as determined by the third-party dataset license holders listed for each dataset individually in the Supplementary Material.

Code availability

The codebase developed in-house for training the CAISR system is publicly available on GitHub at https://github.com/bdsp-core/CAISR-App. This software features a command-line interface that allows users to initialize, train, and evaluate models without modifying the underlying code. The CAISR system is designed to streamline sleep data analysis tasks using Docker containers for enhanced usability. Users can easily download pre-built Docker images from our website. For transparency and customization, the Python code used to create these images is also provided. This allows users to adapt the system to their own datasets or analysis preferences and rebuild the Docker images with minimal effort.

References

1. Foster RG. Sleep, circadian rhythms and health. *Interface Focus* 2020;**10**(20190098):20190098. doi:10.1098/rsfs.2019.0098
2. Krueger JM, Rector DM, Roy S, Van Dongen HPA, Belenky G, Panksepp J. Sleep as a fundamental property of neuronal assemblies. *Nat Rev Neurosci.* 2008;**9**:910–919. doi:10.1038/nrn2521
3. Luyster FS, Strollo PJ, Zee PC, Walsh JK; Boards of Directors of the American Academy of Sleep Medicine and the Sleep Research Society. Sleep: A Health Imperative. *Sleep.* 2012;**35**:727–734. doi:10.5665/sleep.1846
4. Wetter TC, Collado-Seidel V, Pollmächer T, Yassouridis A, Trenkwalder C. Sleep and periodic leg movement patterns in drug-free patients with Parkinson's disease and multiple system atrophy. *Sleep.* 2000;**23**:361–367.
5. Baranwal N, Yu PK, Siegel NS. Sleep physiology, pathophysiology, and sleep hygiene. *Prog Cardiovasc Dis.* 2023;**77**:59–69. doi:10.1016/j.pcad.2023.02.005
6. Desai D, Momin A, Hirpara P, Jha H, Thaker R, Patel J. Exploring the role of circadian rhythms in sleep and recovery: a review article. *Cureus* 2024;**16**:6. doi:10.7759/cureus.61568

7. Irwin MR, Opp MR. Sleep health: reciprocal regulation of sleep and innate immunity. *Neuropsychopharmacology.* 2017;**42**:129–155. doi:10.1038/npp.2016.148
8. Gottlieb DJ, Punjabi NM. Diagnosis and management of obstructive sleep apnea: a review. *JAMA.* 2020;**323**(1389):1389–1400. doi:10.1001/jama.2020.3514
9. Ramar K, Malhotra RK, Carden KA, *et al*. Sleep is essential to health: an American Academy of Sleep Medicine position statement. *J Clin Sleep Med.* 2021;**17**:2115–2119. doi:10.5664/jcsm.9476
10. The L. Waking up to the importance of sleep. *The Lancet* 2022;**400**(973):10357.
11. Malhotra, R. K. & Avidan, A. Y. Sleep Stage Scoring. In: Thomas RJ, Bhat S, Chokroverty S, eds. *Atlas of Sleep Medicine* Cham: Springer International Publishing; 2023: 125–163. doi:10.1007/978-3-031-34625-5_7
12. Reis MJF, Vago EL, Frange C, Coelho FMS. Objective assessment of sleep. In Frange C, Coelho FMS, eds. *Sleep Medicine and Physical Therapy* . Cham: Springer International Publishing; 2022: 401–410. doi:10.1007/978-3-030-85074-6_36
13. Parker KP, Bliwise DL, Ribeiro M, *et al*. Sleep/wake patterns of individuals with advanced cancer measured by ambulatory polysomnography. *J Clin Oncol.* 2008;**26**:2464–2472. doi:10.1200/JCO.2007.12.2135
14. Berry RB, Abreu AR, Krishnan V, Quan SF, Strollo PJ, Malhotra RK. A transition to the American Academy of Sleep Medicine–recommended hypopnea definition in adults: initiatives of the Hypopnea Scoring Rule Task Force. *J Clin Sleep Med.* 2022;**18**:1419–1425. doi:10.5664/jcsm.9952
15. Caples SM, Anderson WM, Calero K, Howell M, Hashmi SD. Use of polysomnography and home sleep apnea tests for the longitudinal management of obstructive sleep apnea in adults: an American Academy of Sleep Medicine clinical guidance statement. *J Clin Sleep Med.* 2021;**17**:1287–1293. doi:10.5664/jcsm.9240
16. Phan H, Mikkelsen K, Chen OY, Koch P, Mertins A, De Vos M. SleepTransformer: automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans Biomed Eng.* 2022;**69**:2456–2467. doi:10.1109/TBME.2022.3147187
17. Nasiri, S. & Clifford, G. D. *Attentive Adversarial Network for Large-Scale Sleep Staging*.
18. Hermans LW, Huijben IA, van Gorp H, *et al*. Representations of temporal sleep dynamics: Review and synthesis of the literature. *Sleep Med Rev.* 2022;**63**(101611):101611. doi:10.1016/j.smrv.2022.101611
19. Yan R, Li F, Wang X, Ristaniemi T, Cong F. Automatic Sleep Scoring Toolbox and Its Application in Sleep Apnea. In Obaidat, MS, ed. *E-Business and Telecommunications.* **vol. 1247**. Cham: Springer International Publishing; , 2020: 256–275.
20. Zan H, Yildiz A. Multi-task learning for arousal and sleep stage detection using fully convolutional networks. *J Neural Eng.* 2023;**20:**056034. doi:10.1088/1741-2552/acfe3a
21. Einizade A, Nasiri S, Sardouie SH, Clifford GD. ProductGraphSleepNet: Sleep staging using product spatio-temporal graph learning with attentive temporal aggregation. *Neural Netw.* 2023;**164**:667–680. doi:10.1016/j.neunet.2023.05.016
22. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *Npj Digit. Med.* 2021;**4**(72):72. doi:10.1038/s41746-021-00440-5
23. Nasiri S, Ganglberger W, Sun H, Thomas RJ, Westover MB. Exploiting labels from multiple experts in automated sleep scoring. *Sleep.* 2023;**46**:zsad034. doi:10.1093/sleep/zsad034
24. Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc.* 2018;**25**:1643–1650. doi:10.1093/jamia/ocy131
25. Guillot A, Thorey V. RobustSleepNet: transfer learning for automated sleep staging at scale. *IEEE Trans Neural Syst Rehabil Eng.* 2021;**29**:1441–1451. doi:10.1109/TNSRE.2021.3098968
26. Gunter KM, Brink-Kjaer A, Mignot E, Sorensen HBD, During E, Jennum P. SViT: A spectral vision transformer for the detection of REM sleep behavior disorder. *IEEE J. Biomed. Health Inform.* 2023;**27**:4285–4292. doi:10.1109/JBHI.2023.3292231
27. Jin Z, Jia K. SAGSleepNet: A deep learning model for sleep staging based on self-attention graph of polysomnography. *Biomed. Signal Process. Control* 2023;**86**(105062):105062. doi:10.1016/j.bspc.2023.105062
28. Goshtasbi N, Boostani R, Sanei S. SleepFCN: a fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Trans Neural Syst Rehabil Eng.* 2022;**30**:2088–2096. doi:10.1109/TNSRE.2022.3192988
29. Vallat R, Walker MP. An open-source, high-performance tool for automated sleep staging. *eLife* 2021;**10**:e70092. doi:10.7554/eLife.70092
30. Zahid AN, Jennum P, Mignot E, Sorensen HBD. MSED: a multimodal sleep event detection model for clinical sleep analysis. *IEEE Trans Biomed Eng.* 2023;**70**:2508–2518. doi:10.1109/TBME.2023.3252368
31. Malhotra RK. AASM Scoring Manual 3: a step forward for advancing sleep care for patients with obstructive sleep apnea. *J Clin Sleep Med.* 2024;**20**:835–836. doi:10.5664/jcsm.11040
32. Westover, M. B. *et al*. *The Human Sleep Project*. BDSP doi:10.60508/QJBV-HG78
33. Newman AB. Progression and regression of sleep-disordered breathing with changes in weight: the sleep heart health study. *Arch Intern Med.* 2005;**165**(2408):2408. doi:10.1001/archinte.165.20.2408
34. Chen X, Wang R, Zee P, *et al*. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA). *Sleep.* 2015;**38**(6):877–888. doi:10.5665/sleep.4732
35. Yeboah J, Redline S, Johnson C, *et al*. Association between sleep apnea, snoring, incident cardiovascular events and all-cause mortality in an adult population: MESA. *Atherosclerosis.* 2011;**219**:963–968. doi:10.1016/j.atherosclerosis.2011.08.021
36. Paudel ML, Taylor BC, Ancoli-Israel S, *et al*.; Osteoporotic Fractures in Men (MrOS) Study. Rest/activity rhythms and mortality rates in older men: MROS sleep study. *Chronobiol Int.* 2010;**27**:363–377. doi:10.3109/07420520903419157
37. West LC, Summers M, Tang S, *et al*. Evaluation of consensus sleep stage scoring of dysregulated sleep in Parkinson's disease. *Sleep Med.* 2023;**107**:236–242. doi:10.1016/j.sleep.2023.04.031
38. Zhang G-Q, Cui L, Mueller R, *et al*. The National Sleep Research Resource: towards a sleep data commons. *J Am Med Inform Assoc.* 2018;**25**:1351–1358. doi:10.1093/jamia/ocy064
39. Berry RB, *et al*. The AASM manual for the scoring of sleep and associated events. *Rules Terminol. Tech. Specif. Darien Ill. Am. Acad. Sleep Med.* 2012.
40. Nassi TE, Ganglberger W, Sun H, *et al*. Automated scoring of respiratory events in sleep with a single effort belt and deep neural networks. *IEEE Trans Biomed Eng.* 2022;**69**:2094–2104. doi:10.1109/TBME.2021.3136753
41. Lacourse K, Yetton B, Mednick S, Warby SC. Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data. *Sci Data.* 2020;**7**(190):190. doi:10.1038/s41597-020-0533-4
42. Ferri R, Manconi M, Rundo F, *et al*. A data-driven analysis of the rules defining bilateral leg movements during sleep. *Sleep.* 2016;**39**:413–421. doi:10.5665/sleep.5454

43. Li H, Guan Y. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Commun Biol*. 2021;**4**(1):18.
44. Zhang H, Wang X, Li H, Mehendale S, Guan Y. Auto-annotating sleep stages based on polysomnographic data. *Patterns (New York, N.Y.)* 2022;**3**:100371. doi:10.1016/j.patter.2021.100371
45. Ferri R, Marelli S, Ferini-Strambi L, *et al*. An observational clinical and video-polysomnographic study of the effects of clonazepam in REM sleep behavior disorder. *Sleep Med*. 2013;**14**:24–29. doi:10.1016/j.sleep.2012.09.009
46. Moore H, Leary E, Lee S-Y, *et al*. Design and validation of a periodic leg movement detector. *PLoS One*. 2014;**9**:e114565. doi:10.1371/journal.pone.0114565
47. Ganglberger W, Nasiri S, Sun H, *et al*. Refining sleep staging accuracy: transfer learning coupled with scorability models. *Sleep*. 2024;**47**(11):zsae202. doi:10.1093/sleep/zsae202
48. Sun Q, Liu Y, Chua TS, Schiele B. Meta-Transfer Learning for Few-Shot Learning. Preprint at http://arxiv.org/abs/1812.02391 (2019).
49. Tsimpoukelli, M. *et al*. Multimodal Few-Shot Learning with Frozen Language Models. Preprint at http://arxiv.org/abs/2106.13884 (2021).
50. Shen Z, Liu Z, Qin J, Savvides M, Cheng K-T. Partial Is Better Than All: Revisiting Fine-tuning Strategy for Few-shot Learning. *Proc. AAAI Conf. Artif Intell*. 2021;**35**:9594–9602.
51. Basner M, Griefahn B, Penzel T. Inter-rater agreement in sleep stage classification between centers with different backgrounds. *Somnologie - Schlafforschung und Schlafmedizin*. 2008;**12**:75–84. doi:10.1007/s11818-008-0327-y
52. Danker-Hopfe H, Kunz D, Gruber G, *et al*. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res*. 2004;**13**:63–69. doi:10.1046/j.1365-2869.2003.00375.x
53. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*. 2000;**23**:901–908.
54. Younes M, Raneri J, Hanly P. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. *J Clin Sleep Med*. 2016;**12**:885–894. doi:10.5664/jcsm.5894
55. Jing J, Ge W, Hong S, *et al*. Interrater Reliability of Expert Electroencephalographers Identifying Seizures and Rhythmic and Periodic Patterns in EEGs. *Neurology*. 2023;**100**:e1750–e1762. doi:10.1212/WNL.0000000000207127
56. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. 2015;**68**:855–859. doi:10.1016/j.jclinepi.2015.02.010
57. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*. 2015;**10**:e0118432. doi:10.1371/journal.pone.0118432
58. Brink-Kjaer A, Olesen AN, Peppard PE, *et al*. Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness. *Clin Neurophysiol*. 2020;**131**:1187–1203. doi:10.1016/j.clinph.2020.02.027
59. Khalid F, Ayache M, Auckley D. The differential impact of respiratory event scoring criteria on CPAP eligibility in women and men. *J Clin Sleep Med*. 2021;**17**:2409–2414. doi:10.5664/jcsm.9448
60. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med*. 2022;**18**:193–202. doi:10.5664/jcsm.9538
61. Pourbabaee, B., Howe-Patterson, M., Patterson, M. & Benard, F. SleepNet: Automated Sleep Analysis via Dense Convolutional Neural Network Using Physiological Time Series. Preprint at http://arxiv.org/abs/1903.04377 (2019).
62. Almarshad MA, Al-Ahmadi S, Islam MS, BaHammam AS, Soudani A. Adoption of transformer neural network to improve the diagnostic performance of oximetry for obstructive sleep apnea. *Sensors (Basel, Switzerland)* 2023;**23**:7924. doi:10.3390/s23187924
63. Levy J, Álvarez D, Del Campo F, Behar JA. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. *Nat Commun*. 2023;**14**:4881. doi:10.1038/s41467-023-40604-3
64. Zinchuk A, Srivali N, Qin Li, *et al*. Association of periodic limb movements and obstructive sleep apnea with risk of cardiovascular disease and mortality. *J. Am. Heart Assoc*. 2024;**13**:e031630. doi:10.1161/JAHA.123.031630