Automated Sleep Stage and Event Detection Algorithms Using Quality-Controlled PSG Annotations

Michiru Kaneda,[1,*] Sho Ogaki,[1] Tomoyuki Nohara,[1] Syuhei Fujita,[1] Naoshi Osako,[1] Tomoko Yagi,[2] Yasuhiro Tomita[3,4,5] and Takanori Ogata[1]

[1]ACCELStars, Inc., 7-3-1 Hongo, Bunkyo-ku, 113-0033, Tokyo, Japan, [2]School of Medical Technology, Kurume University, 67 Asahi-machi, Kurume, 830-0011, Fukuoka, Japan, [3]Sleep and Respiratory Medicine, Toranomon Hospital, 2-2-2 Toranomon, Minato-ku, 105-8470, Tokyo, Japan, [4]Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-0033, Tokyo, Japan and [5]Faculty of Medicine, Juntendo University, 2-1-1 Hongo, Bunkyo-ku, 113-8421, Tokyo, Japan

*Corresponding author. kaneda@accelstars.com

Abstract

Study Objectives To develop machine-learning models for sleep stage classification, arousal detection, and respiratory event detection from polysomnography (PSG), and to evaluate performance against expert scorers.

Methods Overnight PSG recordings were collected from healthy participants and patients with suspected sleep-disordered breathing. Four certified scorers underwent calibration sessions and produced reference annotations for sleep stages, arousals, and respiratory events. In addition, all four scorers annotated a subset of recordings for consensus analyses. Gradient-boosted decision tree models were trained on hand-crafted features from electroencephalography, electrooculography, electromyography, nasal pressure airflow, and oxygen saturation using five-fold cross-validation. Performance was assessed using epoch-wise accuracy, Cohen's κ, F_1-score, event-based precision, recall, and F_1-score based on intersection-over-union matching, and recording-level agreement for total sleep time, arousal index, and apnea-hypopnea index (AHI). Model performance was compared with inter-scorer agreement derived from the consensus dataset.

Results The sleep stage model achieved accuracy 0.840, Cohen's κ 0.791, and F_1-score 0.841, with total sleep time residuals centered near zero and limits of agreement of approximately ± 0.5 h. Arousal detection yielded recall 0.725, precision 0.742, and F_1-score 0.733, with arousal index limits of agreement of about ± 15 events/h. Respiratory event detection achieved recall 0.829, precision 0.807, and F_1-score 0.818, with similar agreement for AHI. In consensus analyses, model performance was comparable to or within the range of inter-scorer agreement.

Conclusions Feature-based decision-tree models can approach human-level accuracy across PSG scoring tasks, highlighting the importance of annotation consistency and supporting their use as decision-support tools in clinical practice.

Key words: Polysomnography; Sleep Stages; Arousal; Sleep Disordered Breathing; Sleep Apnea Syndromes; Machine Learning

Statement of significance

Accurate and consistent scoring of overnight sleep studies remains a major bottleneck in sleep medicine, limiting timely diagnosis and effective treatment. This study demonstrates that carefully designed, interpretable machine learning models trained on rigorously calibrated expert annotations can reach performance comparable to human scorers for sleep stages, arousals, and respiratory events. The findings suggest that the primary barrier to robust automation is not model complexity but the reliability of human labels. Automated scoring based on such models could provide a stable reference for training clinicians, standardizing quality across centers, and enabling large-scale research. Critical next steps include external validation, prospective outcome studies, and integration into routine clinical workflows.

Introduction

Sleep disorders such as insomnia and sleep-disordered breathing (SDB) are highly prevalent conditions associated with increased cardiovascular, metabolic, and neurocognitive risks [1–4]. Accurate evaluation of sleep architecture and respiration is therefore essential for diagnosis and management, with polysomnography (PSG) remaining the gold standard modality [5, 6]. By capturing comprehensive physiological signals, PSG enables precise assessment of sleep physiology and disease severity, supporting optimal clinical and therapeutic decision-making [7].

PSG typically records multiple physiological channels, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), respiratory effort, airflow, and oxygen saturation. From these data, trained scorers manually annotate events such as sleep stages, arousals, and respiratory abnormalities [8]. The American Academy of Sleep Medicine (AASM) manual provides standardized scoring rules that are widely adopted across sleep laboratories [9, 10].

Although these criteria have improved standardization, PSG scoring remains predominantly based on visual inspection by trained scorers. This process is labor-intensive, time-consuming, and dependent on specialized expertise, often limiting diagnostic throughput. Because PSG interpretation inherently depends on human judgment, scoring accuracy is constrained by perceptual and cognitive limitations, resulting in notable inter-scorer variability even under AASM guidelines [7, 11–15]. Variations in institutional practices, scorer training, and interpretation further challenge standardization, limiting both clinical consistency and research reproducibility. Consequently, there has been increasing interest in automated PSG analysis to streamline clinical workflows, reduce scoring burden, and improve inter-scorer reliability.

Earlier studies employed feature-based machine learning approaches such as support vector machines, random forests, and decision trees for automatic sleep stage classification [16–21]. Subsequent works introduced deep learning architectures capable of learning temporal and spectral representations directly from raw data, achieving improved accuracy [22–26]. More recently, interpretable modeling approaches have emerged, including image-based representations and cross-modal

transformer architectures designed to provide human-aligned explanations of PSG predictions [27, 28]. At the same time, traditional machine-learning pipelines have been shown to perform comparably to state-of-the-art deep learning models [29].

Similar progress has occurred in automated detection of arousals and respiratory events. Early automated approaches relied on hand-crafted features derived from EEG, ECG, airflow, and respiratory effort [30–33], whereas later research increasingly adopted deep learning models that directly capture temporal dependencies and event durations from PSG signals [34–38].

Furthermore, a recent hybrid pipeline integrates deep learning and signal processing to analyze sleep staging, arousal, and respiratory events, and additionally detects limb movements within the same framework [39].

Many of these studies rely on publicly available PSG databases, which provide valuable benchmarks for cross-study comparison. However, these datasets differ considerably in their scoring conventions and preprocessing standards. Older databases were annotated according to the Rechtschaffen and Kales (R&K) criteria [40], and although conversion guidelines to the AASM system exist [41], they are not perfectly equivalent. Moreover, studies vary in how they select and segment data—such as the definition of lights-off and lights-on intervals, the treatment of high-noise segments, and the allocation of data into training and test sets—as well as in how they match events for evaluation. Because these methodological details are not standardized across datasets or studies, direct numerical comparison of reported performance metrics is inherently difficult.

In addition to these inconsistencies, the reliability of the human-generated annotations themselves remains uncertain. While annotation variability places an upper bound on achievable model accuracy, improving label consistency—as done in this study—raises this ceiling and enables more reliable automated analysis. As the accuracy of these annotations is not absolute, the true upper limit of model performance remains difficult to define, and small differences in reported accuracy should therefore be interpreted with caution. Recent hypnodensity-based analyses have shown that scorer disagreement often reflects intrinsic ambiguity in the underlying physiological signals rather than simple noise, suggesting that a certain degree of uncertainty is intrinsic to PSG scoring [42–44].

To better account for this limitation, several studies have explicitly examined inter-scorer agreement and incorporated it into model evaluation. Some investigations have compared model accuracy directly with human-level benchmarks by using multiple independent annotations [45, 46], whereas others have generated consensus-based "platinum labels" through adjudication among several scorers to obtain maximally reliable reference standards [39]. Collectively, these findings underscore that the consistency and quality of human annotations strongly influence both model performance and interpretability, highlighting the importance of improving label agreement in datasets used for training and evaluation.

In the present study, a high-quality reference dataset was established through a structured calibration process rather than through adjudicated consensus labeling. Multiple certified sleep scorers first participated in targeted calibration sessions to harmonize their scoring criteria for sleep stages, arousals, and respiratory events. After achieving adequate alignment, each scorer independently annotated the same

PSG recordings under unified scoring guidelines, yielding a set of consistent and quality-controlled labels suitable for algorithm development. A subset of recordings was double-scored to quantify inter-scorer reliability, which was then compared with model performance to assess the extent to which the automated system approximated human-level agreement. This design avoids the variability introduced by uncalibrated datasets while providing a set of consistently generated, high-quality annotations for model development and evaluation.

Using this rigorously standardized annotation set, a unified feature-based machine-learning framework was developed and evaluated for sleep stage classification, arousal detection, and respiratory event detection. Model performance was assessed relative to multiple expert scorers to provide a benchmark grounded in high-quality, consistently generated PSG annotations.

Methods

Ethics Statements

This study was approved by the Ethics Committee of Medical Corporation Kouseikai Sone Clinic (Protocol IDs: SUG08869 and SUG08884). Written informed consent was obtained from all participants. All data were anonymized prior to analysis, and no identifying information is included in this manuscript.

Annotation

All analyses and annotations in this study were performed on PSG data. Annotation of sleep stages, arousals, and respiratory events followed the AASM Scoring Manual [47]. For each recording, the lights-off and lights-on times were first identified, and scoring was performed only within the defined lights-off interval.

A team of four experienced scorers first independently annotated a subset of the PSG data to establish consistent criteria for sleep stage classification, arousal detection, and respiratory event identification. A calibration meeting was then held to harmonize these criteria and resolve systematic differences among scorers, ensuring consistent interpretation across all annotation types.

After calibration, a small set of PSG recordings were annotated by all scorers to create a consensus dataset for sleep stage classification and arousal detection. Once this consensus dataset was established, the remaining PSG recordings were distributed among the scorers, and each scorer independently annotated their assigned data.

Except for the consensus subset, all respiratory event annotations were conducted by a single expert scorer. Respiratory events included hypopnea, obstructive apnea, central apnea, and mixed apnea; however, for the purposes of this study, all were treated collectively as a single class of respiratory events.

Because PSG data were collected over an extended period, annotation sessions were also conducted at different times, which may have introduced minor temporal variations in scoring style.

Dataset

Two types of overnight PSG datasets were used in this study. The first dataset consisted of PSG recordings acquired from healthy participants in a dedicated in-house sleep research laboratory equipped for controlled overnight monitoring. The second dataset consisted of PSG data obtained at a collaborating clinical sleep center from patients referred for evaluation of suspected SDB, primarily obstructive sleep apnea (OSA). All participants were Japanese.

SOMNO BT Plus (SOMNOmedics AG) and Alice 6 LDx (Royal Philips) were used to acquire PSG data in the sleep research laboratory and clinical sleep center, respectively. Table 1 summarizes the sampling frequencies of the PSG channels from which features were extracted.

SOMNO BT Plus (SOMNOmedics AG) and Alice 6 LDx (Royal Philips) were used to acquire PSG data in the sleep research laboratory and clinical sleep center, respectively. Table 1 summarizes the sampling frequencies of the PSG channels from which features were extracted.

Table 1. Sampling frequencies of PSG channels

Channel	SOMNO BT Plus	Alice 6 LDx
Frequency [Hz]		
Electroencephalography	256	500
Electrooculography	256	500
Electromyography	256	500
Nasal pressure airflow	256	100
SpO_2	4	1

A total of 473 recordings from healthy participants and 244 recordings from patients with suspected SDB were used for the sleep stage classification study (Table 2), for which the recordings were sufficiently clear to allow full-night sleep stage classification. From this pool, 412 healthy recordings and 178 SDB recordings were used in the arousal detection study, reflecting the requirement that data quality for the arousal and respiratory event detection studies be sufficient over the entire night (Table 3).

A total of 604 recordings from healthy participants and 245 recordings from patients with suspected SDB were initially obtained. Recordings were screened to include only complete overnight sessions with sufficient signal quality and no major signal loss. For participants with multiple recorded nights, a single night was selected to avoid duplication at the participant level. After this quality control, 473 healthy recordings and 244 SDB recordings were included in the sleep stage classification study (Table 2), as these recordings were sufficiently clear to allow full-night sleep stage classification. From this quality-controlled subset, 412 healthy recordings and 178 SDB recordings were further included in the arousal detection study, reflecting the additional requirement that data quality be sufficient over the entire night for reliable arousal and respiratory event detection (Table 3).

Table 2. PSG datasets for the sleep stage study

Characteristic	Healthy	SDB
Recording period	2022/05/04 –2024/09/30	2023/08/01 –2024/06/24
N participants	473	244
Sex, F/M [%]	44.2/55.6[1]	21.3/78.7
Age [years]	33.4 (13.3) [9, 59]	56.7 (12.2) [24, 85]
BMI [kg/m^2]	21.6 (3.1) [15.6,37.4]	25.5 (4.3) [16.1, 45.6]
TST [h]	6.2 (1.4) [1.5, 9.8]	6.0 (1.0) [2.3, 7.8]

BMI: body mass index; TST: total sleep time.

The Healthy group refers to participants without suspected sleep disorders, while the SDB group includes patients referred for evaluation of suspected sleep-disordered breathing.

For age, BMI, and TST, numbers are presented as mean (standard deviation) [range].

[1]One participant did not answer.

For the respiratory event detection study, only recordings from participants with suspected SDB were used.

For the evaluation of inter-scorer agreement, 27 recordings were used in the consensus study for sleep stage classification and arousal detection, and 5 SDB recordings were used in the respiratory event consensus study (Table 4).

Algorithms

This study employed a unified feature-based machine learning framework for three tasks: sleep stage classification, arousal detection, and respiratory event detection. All tasks were formulated as epoch-wise classification problems based on manually annotated PSG signals. Data were segmented into fixed-length epochs—30 seconds for sleep stage classification and 1 second for arousal and respiratory events. From each epoch, task-specific sets of time-domain and frequency-domain features were extracted from EEG, EOG, EMG, nasal pressure airflow, and SpO_2 signals. Although the PSG recordings were acquired at different native sampling frequencies (Table 1), the feature-extraction procedures operated on time-normalized segments and predefined frequency bands, ensuring that signals from both devices yielded harmonized feature representations.

A hierarchical model structure was used: the estimated sleep stage from the sleep stage model was included as a feature for arousal detection, and both the estimated sleep stage and arousal labels were used as features for respiratory event detection. To capture temporal dependencies in physiological signals, features from the target epoch and neighboring epochs were concatenated to form an extended context window.

Table 3. PSG datasets for the arousal and respiratory event studies

Characteristic	Healthy	SDB
Recording period	2022/05/04 –2024/09/30	2023/08/01 –2024/06/24
N participants	412	178
Sex, F/M [%]	44.4/55.6	19.7/80.3
Age [years]	33.5 (13.2) [9, 59]	56.5 (12.4) [24, 85]
BMI [kg/m^2]	21.6 (3.1) [15.6,32.3]	25.2 (4.3) [16.1, 45.6]
TST [h]	6.2 (1.4) [1.5, 9.8]	6.0 (1.0) [2.3, 7.8]
ArI [1/h]	14.8 (9.1) [2.1, 72.5]	41.1 (17.6) [6.2, 114.5]
AHI [1/h]	-	40.9 (20.6) [3.5, 103.0]

BMI: body mass index; TST: total sleep time; ArI: arousal index; AHI: apnea-hypopnea index.

The Healthy group refers to participants without suspected sleep disorders, while the SDB group includes patients referred for evaluation of suspected sleep-disordered breathing.

For age, BMI, TST, ArI, and AHI, numbers are presented as mean (standard deviation) [range].

All models were implemented using gradient-boosted decision trees (LightGBM [48]), chosen for their suitability for heterogeneous tabular features and efficient training. For arousal and respiratory events, the 1-second epoch-wise probabilities output by the classifier were merged using task-specific post-processing procedures to reconstruct contiguous events.

The following subsections describe the task-specific feature sets, temporal context structures, classification models, and post-processing steps. Throughout all tasks, airflow features refer specifically to nasal pressure airflow derived from a cannula sensor, which was the airflow modality used in this study.

Sleep Stage Classification

Sleep Stage Feature Extraction

Data were segmented into consecutive, non-overlapping 30-second epochs. For each epoch, a total of 1450 features were computed from EEG, EOG, and EMG channels.

For the EEG signals (six channels), power spectral density (PSD) features were computed using the Welch method [49], which estimates the power spectrum by averaging modified periodograms of overlapping signal segments to reduce variance. Each 30-second epoch was divided into six 5-second subsegments, and the log-power

Table 4. PSG datasets for the consensus studies

Characteristic	Healthy	SDB
Recording period	2022/05/04 –2022/10/29	2023/08/01 –2023/08/16
N participants	27	5
Sex, F/M [%]	29.6/70.4	0.0/100.0
Age [years]	40.1 (11.5) [23, 59]	56.8 (13.6) [39, 73]
BMI [kg/m^2]	22.1 (2.0) [18.8,26.0]	24.9 (4.7) [20.3, 32.5]
TST [h]	6.7 (1.2) [4.8, 9.5]	5.5 (0.5) [5.1, 6.1]
ArI [1/h]	13.3 (5.3) [2.6, 26.2]	19.7 (7.6) [12.9, 29.4]
AHI [1/h]	-	32.1 (16.7) [13.4, 51.2]

BMI: body mass index; TST: total sleep time; ArI: arousal index; AHI: apnea-hypopnea index.

The Healthy group refers to participants without suspected sleep disorders, while the SDB group includes patients referred for evaluation of suspected sleep-disordered breathing.

For age, BMI, TST, ArI, and AHI, numbers are presented as mean (standard deviation) [range].

spectral density was estimated for the following frequency bands: slow delta (0.5–2 Hz), fast delta (2–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and spindle (11–16 Hz). Band-power ratios were also derived. Identical PSD and ratio features were computed from the PSD estimated over the full 30-second epoch. Time-domain features included the mean, root mean square (RMS), skewness, kurtosis, and 0.75 quantile for each channel.

For the EOG signals (two channels), PSD features were extracted at 1-Hz resolution from 0–35 Hz using the same six-segment structure as for EEG. Time-domain statistics (mean, RMS, skewness, kurtosis, 0.75 quantile) were also computed.

For the EMG signal (one channel), RMS values were computed for each 1-second segment within the epoch, and the sequence of RMS values served as EMG features.

Sleep Stage Classification Model and Temporal Context

To incorporate temporal information, features from the target epoch and the two preceding and two following epochs were concatenated, forming a five-epoch context window (7250 features per input sample). LightGBM was used to classify each epoch into Wake, N1, N2, N3, or REM.

Arousal Detection

Arousal Feature Extraction

For arousal detection, data were segmented into 1-second epochs, and labels were generated by converting event-based arousal annotations into 1-second epoch-wise labels.

For each epoch, 34 features were computed from EEG, EOG, EMG, nasal pressure airflow, and SpO_2 signals, in addition to the estimated sleep stage.

The sleep stage estimated by the sleep stage model was included as a categorical variable and converted into dummy variables for N1, N2, N3, and REM.

For the EEG signals, PSD features were computed from each 1-second segment using the same spectral bands defined for sleep stage classification. From these, only beta power, spindle power, and the theta/alpha ratio were retained, based on feature-importance analysis indicating their relevance for arousal discrimination. RMS was also computed for each EEG channel.

For the EOG and EMG channels, the RMS of each 1-second segment was extracted.

For the nasal pressure airflow signal, respiratory volume per time (RVT) was computed using a Hilbert-based method that estimates instantaneous respiratory amplitude and period [50] and the RVT values were then averaged within each 1-second segment.

For the SpO_2 signal, the RMS and mean of each 1-second segment were used.

Arousal Classification Model and Temporal Context

Features from the target epoch and the 30 preceding and 30 following epochs were concatenated to form a 61-epoch context window (2074 features per sample). LightGBM was used to classify each epoch as arousal or non-arousal.

Arousal Post-Processing

Epochs with predicted arousal probability above 0.3 were marked as arousal candidates. Gaps of 5 seconds or less between candidates were bridged to form continuous events. Events shorter than 3 seconds were removed. Thresholds were selected by minimizing the mean absolute error between predicted and annotated Arousal Index.

Respiratory Event Detection

Respiratory Event Feature Extraction

For respiratory event detection, data were segmented into 1-second epochs, and respiratory annotations were converted into epoch-wise labels following the same procedure as for arousals. For each epoch, seven features were computed from the estimated sleep stage, estimated arousal labels, nasal pressure airflow, and SpO_2 signals.

The estimated sleep stage was included as dummy variables (N1, N2, N3, REM). The arousal detection model provided a binary arousal indicator (1 for predicted arousal, 0 otherwise).

For the nasal pressure airflow signal, RVT was computed for each 1-second segment as described in the Arousal Feature Extraction section. For the SpO_2 signal, the mean of each 1-second segment was included.

Respiratory Event Classification Model and Temporal Context

A temporal context window of 61 seconds was constructed by concatenating features from the target epoch with the 30 preceding and 30 following epochs (427 features per input sample). LightGBM was used to classify each epoch as either a respiratory event or a non-event.

Respiratory Event Post-Processing

Epochs with predicted probability above 0.7 were marked as respiratory event candidates. Gaps of 2 seconds or less between candidate epochs were bridged to form continuous events. Events shorter than 10 seconds were removed. Thresholds were optimized by minimizing the mean absolute error in apnea-hypopnea index (AHI) compared with manual annotations.

Model Training and Evaluation

Cross-Validation

All models were trained and evaluated using a participant-level five-fold cross-validation (CV) framework. Because the dataset comprised healthy individuals and participants with suspected SDB, the two groups were handled separately when constructing the folds. Within each group, participants were randomly divided into five approximately equal subsets. Five final folds were then created by pairing the corresponding healthy and SDB subsets, ensuring comparable proportions of healthy and SDB recordings in each fold.

In each CV iteration, three folds were used for training, one fold served as the validation set for early stopping, and the remaining fold was held out for testing. For arousal detection and respiratory event detection, the same participant-level fold assignments established for sleep stage classification were retained. Each participant thus had a consistent fold index (folds 1-5) across all tasks, while the actual recordings used for arousal or respiratory event detection corresponded to the task-specific subsets. This ensured consistent participant grouping across tasks and prevented information leakage when sleep stage or arousal predictions were used as input features in downstream models.

Unless otherwise specified, all performance metrics reported in the Results section were computed on the held-out test fold in each CV iteration and averaged across the five folds.

Sleep Stage Evaluation

To evaluate sleep stage classification performance, true and predicted sleep stages were compared for all 30-second epochs. From these epoch-wise labels, metrics (accuracy, Cohen's κ coefficient, macro-averaged F_1-score, and stage-wise F_1-score for Wake, N1, N2, N3, and REM) were computed for each test fold and then averaged across the five folds.

At the recording level, total sleep time (TST) was calculated separately from the reference annotations and from the model outputs by summing the durations of all epochs labeled as N1, N2, N3, or REM for each recording. For each recording, the residual was defined as predicted TST minus true TST. Agreement between true and

predicted TST was then summarized using residual plots and Bland-Altman limits of agreement (LoA) [51].

Event-Based Metrics

For arousal and respiratory event detection, event-based performance was evaluated by treating each true event (reference annotation) and each predicted event (model output) as a time interval defined by its onset and duration. For every pair of true and predicted events, the intersection over union (IoU) metric was computed as the ratio of the duration of their temporal overlap to the duration of their union [52]. A predicted event was considered a match if its IoU with a true event was at least 0.3. Matched pairs were counted as true positives (TP), true events without a matching prediction (IoU ≤ 0.3) were counted as false negatives (FN), and predicted events without a corresponding true event were counted as false positives (FP). The IoU threshold of 0.3 was chosen to align with previously proposed evaluation protocols for respiratory event detection, allowing direct comparison with prior human-AI benchmarking studies [46].

From TP, FP, and FN, precision, recall, and F_1-score were calculated for each test fold and then averaged across the five folds.

At the recording level, the arousal index (ArI) and apnea-hypopnea index (AHI) were computed from both reference annotations and model outputs by dividing the total number of detected events by the TST in hours. For each recording, the residual was defined as predicted index minus true index.

Consensus Annotations and Human-Model Comparison

Consensus subsets were defined a priori for inter-scorer and human-model comparison analyses. For sleep stage classification and arousal detection, 27 PSG recordings from healthy participants were annotated independently by all four scorers (Table 4). For respiratory event detection, five recordings from patients with suspected SDB were annotated by all four scorers.

For sleep stage classification, each scorer's annotations were compared against consensus annotations derived from other scorers' labels [53]. The model outputs were compared against consensus annotations created from the three highest-accuracy scorers out of four [45].

For arousal and respiratory event detection, event-based annotations on the consensus subset were converted to 1-second epoch-wise binary labels (arousal vs. non-arousal) for each scorer. For each epoch, a majority vote among the three selected scorers (at least two arousal labels) was required to label the epoch as arousal, and consecutive arousal-labeled epochs were merged into continuous events, yielding consensus arousal annotations [34]. Each scorer's annotations were compared against consensus annotations derived from the other scorers' labels. The model outputs were compared against consensus annotations created from the three highest-accuracy scorers out of four.

For the consensus analyses, all metrics were computed based on each recording in the consensus subset and then averaged across recordings.

Results

Performance Metrics

Sleep Stage Classification

Table 5 summarizes the normalized confusion matrix across all folds.

Table 5. Normalized confusion matrix for sleep stage classification, expressed as row-wise percentages. True labels (reference annotations) are shown in rows, and predicted labels (model outputs) are shown in columns.

True \ Predicted	Wake	N1	N2	N3	REM
Wake	89.3	7.6	0.9	0.1	2.0
N1	6.3	68.6	19.6	0.1	5.4
N2	0.4	6.4	86.9	4.7	1.5
N3	0.1	0.1	14.4	85.4	0.0
REM	1.6	6.1	3.6	0.1	88.7

Across the test folds, the model achieved the following performance, reported as mean (standard deviation; SD): accuracy of 0.840 (0.006), Cohen's κ of 0.791 (0.007), and an overall F_1-score of 0.841 (0.005). Stage-wise F_1-scores for Wake, N1, N2, N3, and REM were 0.890 (0.012), 0.711 (0.004), 0.848 (0.005), 0.865 (0.009), and 0.892 (0.006), respectively. These results indicate strong discrimination for Wake, N2, N3, and REM, whereas the lower performance observed for N1 is consistent with previous PSG-based classification studies.

Figure 1 shows the residual distribution for TST estimates. TSTs were calculated for each recording by summing the durations of N1, N2, N3, and REM stages. Residuals were symmetrically distributed around zero without a systematic bias, and the LoA were approximately ± 0.5 hours.

Arousal Detection

Event-based performance was evaluated using the IoU-based matching criterion defined in the Model Training and Evaluation section. The model achieved mean (SD) recall, precision, and F_1-scores of 0.725 (0.010), 0.742 (0.018), and 0.733 (0.011), respectively. These results indicate balanced sensitivity and specificity, with stable performance across folds.

Figure 2 shows the residual plot for the ArI. The LoA were approximately ± 15 events/h. Variability in the residuals remained consistent over the range of ArI values, and only a small number of outliers were observed.

Respiratory Event Detection

Respiratory event detection employed the same IoU-based matching and performance metrics as arousal detection. The model achieved mean (SD) recall, precision, and F_1-scores of 0.829 (0.010), 0.807 (0.016), and 0.818 (0.012), respectively.

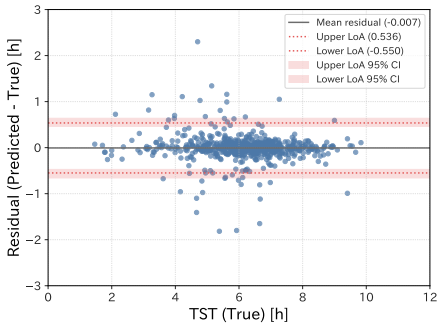

Fig. 1: Residual plot for total sleep time (TST). The horizontal axis represents the true TST (from reference annotations) in hours, and the vertical axis represents the residual, defined as predicted TST (from model output) minus true TST. Each point corresponds to an individual recording. The solid horizontal line indicates the mean residual. The upper and lower dashed lines represent the limits of agreement (LoA), calculated as the mean residual ± 1.96 SD of the residuals. The shaded regions indicate the 95% confidence intervals for the mean and LoA lines.

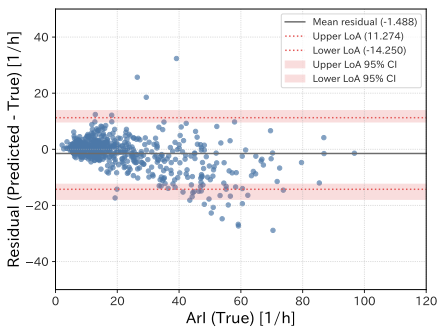

Fig. 2: Residual plot for the arousal index (ArI). The horizontal axis represents the true ArI (from reference annotations), and the vertical axis represents the residual, defined as predicted ArI (from model output) minus true ArI. Each point corresponds to an individual recording. The solid horizontal line indicates the mean residual. The upper and lower dashed lines represent the limits of agreement (LoA), calculated as the mean residual ± 1.96 SD of the residuals. The shaded regions indicate the 95% confidence intervals.

Figure 3 shows the residual plot for the AHI. Residuals were centered around zero with no systematic bias across the range of AHI values. LoAs were approximately ±15 events/h, and only a few outliers were observed in the high-AHI range.

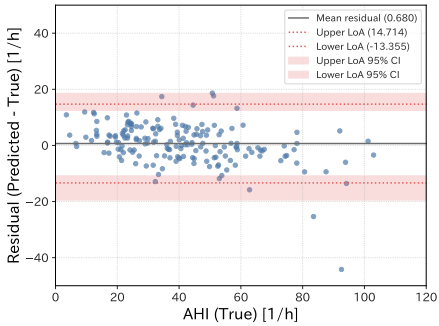

Fig. 3: Residual plot for the apnea-hypopnea index (AHI). The horizontal axis represents the true AHI (from reference annotations), and the vertical axis represents the residual, defined as predicted AHI (from model output) minus true AHI. Each point corresponds to an individual recording. The solid horizontal line indicates the mean residual. The dashed lines indicate the limits of agreement (LoA), calculated as the mean residual ± 1.96 SD of the residuals. The shaded regions indicate the 95% confidence intervals.

Consensus Study

Sleep Stage Classification Consensus

Consensus analysis for the sleep stage classification task was performed on the 27 recording subset described in the Consensus Annotations and Human-Model Comparison section. Metrics were obtained for each recording, and means and SD were calculated across the recordings. Tables 6 and 7 summarize the overall and stage-wise performance metrics for individual scorers and the model against the consensus labels. Figure 4 shows the distribution of stage-wise F_1-score across recordings.

Table 6. Performance metrics for sleep stage classification of individual scorers and the model.

Scorer	Accuracy	Cohen's κ	F_1-score
Scorer 1	0.915 (0.030)	0.878 (0.042)	0.916 (0.029)
Scorer 2	0.916 (0.031)	0.880 (0.045)	0.917 (0.031)
Scorer 3	0.902 (0.032)	0.859 (0.046)	0.902 (0.033)
Scorer 4	0.912 (0.022)	0.874 (0.029)	0.911 (0.021)
Model	0.907 (0.040)	0.864 (0.061)	0.905 (0.042)

Values are presented as mean (standard deviation) calculated from 27 recordings.

The model demonstrated comparable performance to that of human scorers across all sleep stages. The score for the N1 stage was lower than for other stages but still within the range of human scorer performance.

Table 7. Stage-wise F_1-scores for sleep stage classification of individual scorers and the model.

Scorer	Wake	N1	N2	N3	REM
Scorer 1	0.947 (0.037)	0.789 (0.061)	0.912 (0.038)	0.817 (0.199)	0.961 (0.026)
Scorer 2	0.928 (0.053)	0.751 (0.075)	0.921 (0.027)	0.890 (0.111)	0.960 (0.030)
Scorer 3	0.925 (0.052)	0.736 (0.069)	0.905 (0.036)	0.842 (0.167)	0.946 (0.048)
Scorer 4	0.942 (0.040)	0.783 (0.055)	0.911 (0.034)	0.838 (0.104)	0.961 (0.024)
Model	0.897 (0.103)	0.713 (0.078)	0.920 (0.035)	0.874 (0.127)	0.931 (0.060)

Values are presented as mean (standard deviation) calculated from 27 recordings.

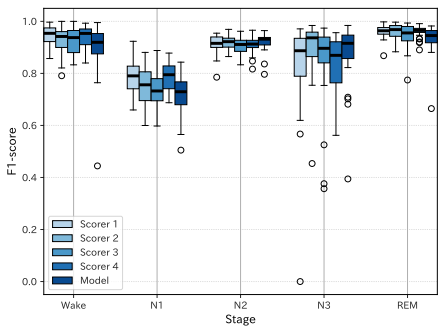

Fig. 4: F_1-score for each sleep stage of individual scorers and the model against consensus annotations. Boxplots showing the distribution of F_1-score across the 27 recordings for each sleep stage (Wake, N1, N2, N3, REM) are presented for each scorer and the model. For each group, the box represents the interquartile range (IQR), with the thick horizontal line indicating the median. Whiskers extend to 1.5 × IQR or to the most extreme data points within that range, and individual points beyond the whiskers are plotted as outliers.

Compared to the results reported in [45], although both models achieved human-level performance, this model showed better results, particularly for the N1 stage, where [45] reported a mean F_1-score of 0.56 (0.14) for the healthy dataset.

Arousal Detection Consensus

Consensus analysis for the arousal detection task was performed on the 27 recording subset described in the Consensus Annotations and Human-Model Comparison section. Event-based performance of each scorer and the model was evaluated against the consensus arousal annotations using the IoU-based matching criterion and event-level metrics defined in the Model Training and Evaluation section. Table 8 summarizes recall, precision, and F_1-score for individual scorers and the model, and Figure 5 shows recall and precision for each against the consensus annotations.

Table 8. Performance metrics for arousal detection of individual scorers and the model.

Scorer	Recall	Precision	F_1-score
Scorer 1	0.846 (0.062)	0.814 (0.069)	0.827 (0.047)
Scorer 2	0.746 (0.090)	0.810 (0.090)	0.770 (0.062)
Scorer 3	0.856 (0.055)	0.724 (0.091)	0.780 (0.061)
Scorer 4	0.768 (0.101)	0.819 (0.076)	0.786 (0.067)
Model	0.845 (0.083)	0.780 (0.078)	0.807 (0.056)

Values are presented as mean (standard deviation) calculated from 27 recordings.

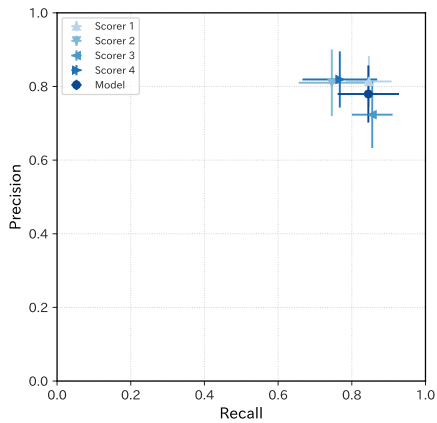

Fig. 5: Recall and precision for arousal detection by individual scorers and by the model, evaluated against consensus annotations. Mean recall and precision across recordings are shown, with error bars indicating the corresponding standard deviations (SD).

The model demonstrated comparable performance to that of human scorers. The F_1-score is higher than that reported in [35] (0.76), although direct comparison is limited due to differences in datasets and annotation protocols.

Respiratory Event Detection Consensus

For respiratory event detection, consensus analysis was conducted on five recordings from patients with suspected SDB, as defined in the Consensus Annotations and Human-Model Comparison section. Consensus respiratory event annotations were compared with each scorer's annotations and with the model outputs using the same IoU-based matching criterion and event-based metrics as for arousals. Table 9 summarizes recall, precision, and F_1-score for individual scorers and the model, and Figure 6 illustrates recall and precision for each against the consensus annotations.

The model demonstrated slightly lower overall performance than human scorers and a characteristic pattern of higher recall coupled with lower precision. While this

Table 9. Performance metrics for respiratory event detection of individual scorers and the model.

Scorer	Recall	Precision	F_1-score
Scorer 1	0.861 (0.045)	0.916 (0.020)	0.888 (0.033)
Scorer 2	0.907 (0.034)	0.835 (0.131)	0.865 (0.078)
Scorer 3	0.821 (0.118)	0.870 (0.039)	0.842 (0.073)
Scorer 4	0.895 (0.029)	0.864 (0.037)	0.878 (0.011)
Model	0.944 (0.039)	0.648 (0.071)	0.767 (0.054)

Values are presented as mean (standard deviation) calculated from 5 recordings.

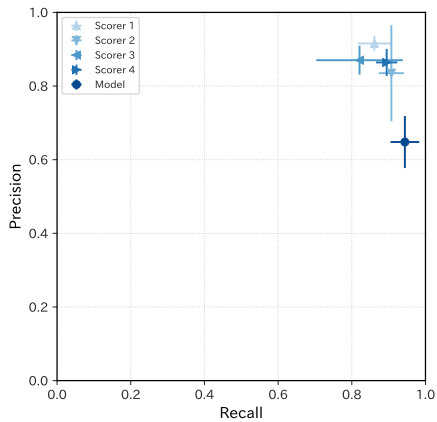

Fig. 6: Recall and precision for respiratory event detection by individual scorers and by the model, evaluated against consensus annotations. Mean recall and precision across recordings are shown, with error bars indicating the corresponding standard deviations (SD).

imbalance warrants careful interpretation, it provides a useful point of comparison with prior work.

A previous study [46] reported human-level performance in a similar evaluation setting; however, both human and model F_1-score were substantially lower in that work, with the model achieving an F_1-score of 0.57 (0.23). This lower performance corresponded with the relatively limited inter-scorer agreement among the annotators in that study. In contrast, the present study employed annotations with substantially higher consistency, which was reflected in the higher performance achieved by both human scorers and the model. Although the model does not yet reach human-level accuracy, these observations suggest that improvements in annotation quality and scorer alignment may directly contribute to the development of more accurate automated systems. Accordingly, the current results should be viewed as indicative of progress rather than a definitive performance limit.

A closer examination of the consensus subset revealed a recall-precision imbalance suggestive of over-detection. Notably, this pattern differs from the balanced precision and recall observed in the full-dataset evaluation described in Section Respiratory Event Detection, where model tuning ensured close alignment with reference AHI values. The discrepancy therefore appears to be specific to the consensus subset rather than reflective of a systematic model tendency.

Several methodological considerations may account for this subset-specific deviation. First, the consensus subset comprises only five recordings, rendering its performance metrics particularly sensitive to sampling variability. Second, these recordings were annotated at an early stage of the project, whereas subsequent annotations were produced over a longer period. Although scorers underwent structured calibration, gradual shifts in annotation tendencies cannot be entirely excluded, and such drift may have contributed to the observed imbalance. Taken together, these factors indicate that the subset's results should be interpreted with caution and should not be generalized beyond their methodological context.

Discussion

Overview of Model Performances

This study evaluated three models designed for distinct sleep-related tasks: sleep stage classification, arousal detection, and respiratory event detection from PSG data. These models were designed to operate sequentially in a pipeline configuration, where the arousal detection model utilizes the outputs of the sleep stage classification, and the respiratory event detection model further incorporates both sleep and arousal results. This hierarchical design reflects the dependency structure of physiological events during sleep and allows each model to benefit from contextual information provided by the preceding stages.

Each model was developed using decision-tree-based architectures rather than deep learning frameworks, demonstrating that accurate predictions can be achieved even with relatively simple and interpretable models.

As previously reported [29], non-deep-learning methods such as gradient-boosted decision trees can also achieve strong performance in sleep stage classification tasks, supporting the notion that model simplicity and interpretability can coexist with high predictive accuracy. In the present work, this concept was extended to event-based tasks—arousal and respiratory event detection—by combining decision-tree classifiers with tailored post-processing algorithms capable of estimating both the onset and duration of each event with high temporal precision.

Annotation Quality and Consensus Study

In this study, particular attention was paid to the quality and consistency of the reference annotations. Before scoring, all scorers participated in calibration sessions to align their interpretation criteria and reduce subjective variability. As a result, the inter-scorer agreement achieved in this dataset was notably high, especially for the N1 stage of sleep and for respiratory event annotations, which have often been sources of disagreement in previous studies.

When compared with human scorers, the proposed models demonstrated performance closely matching human-level accuracy across all sleep-related tasks. In sleep stage classification, the model achieved F_1-score comparable to those of individual scorers for all stages. Notably, the model's F_1-score for the N1 stage was higher than that reported by [45], reflecting the improved consistency among scorers in the present dataset, which in turn allowed the model to achieve better performance.

For arousal detection, the model achieved high precision and recall values, both within the range observed among human scorers, demonstrating consistent event-level accuracy. In respiratory event detection, the model exhibited slightly higher recall and slightly lower precision than human scorers. Nevertheless, its overall performance substantially exceeded that reported by [46]. This difference can be largely attributed to the higher inter-scorer agreement in the present study, which provided more consistent reference labels for training and evaluation.

These findings demonstrate that even relatively simple, decision-tree-based models can reach human-level accuracy. Importantly, the achievable performance ceiling of any model is inherently constrained by the consistency of the human annotations on which it is trained. If annotation quality is unknown or inconsistent, it becomes difficult to determine the true upper bound of model accuracy, and efforts to improve performance by increasing model complexity may yield diminishing returns. Therefore, assessing and ensuring annotation quality is essential for the development of reliable sleep analysis models. Higher annotation quality not only improves the validity of model evaluation but also enables the creation of more accurate and generalizable models.

Limitations

This study has several limitations that should be considered. The dataset used in this work excluded recordings that contained extensive signal loss, severe noise, or missing channels. While this ensured reliable evaluation, real clinical environments often include recordings with such imperfections, and diagnostic interpretation must still be performed under these conditions. For practical use, it will therefore be necessary to establish procedures that can automatically detect poor-quality data and refer them for manual review, or alternatively, to develop models that remain robust even when portions of the input data are missing or corrupted.

High-quality annotations were obtained through calibration sessions among four expert scorers to ensure consistency before labeling. Although this approach effectively reduced inter-scorer variability, it also introduced dataset-specific characteristics influenced by scorer discussions and interpretations that go beyond the strict AASM definitions. Annotation practices can vary across countries, institutions, and individual clinicians, and thus the present model may not necessarily achieve the same level of agreement when applied to independently annotated datasets, even those produced with similarly high inter-scorer reliability.

The study population consisted exclusively of Japanese participants. Physiological and demographic differences, including variations in respiratory patterns and arousal thresholds, may affect the generalizability of the model. Extending the evaluation

to larger and more diverse cohorts will be essential to confirm the robustness of the findings across populations.

In addition, although the results demonstrated that decision-tree-based models can achieve near-human performance under well-controlled conditions, their robustness to real-world variability remains to be verified. Differences in hardware configuration, signal acquisition protocols, and scoring conventions could affect performance when deployed in other clinical environments. Future studies should investigate adaptive learning and cross-dataset calibration strategies to enhance the model's generalizability and resilience.

The current respiratory event detection model treats all events as a single category and does not distinguish between apnea and hypopnea, or between obstructive and central apnea. Expanding the model to detect and classify these subtypes will be an important direction for future development.

Clinical Implications and Future Directions

It is important to emphasize that the proposed models are not intended to replace human scorers, but to complement their work and enhance scoring efficiency. In routine clinical practice, PSG recordings often contain artifacts, missing data, or ambiguous signal patterns that require visual inspection and expert judgment. Manual review therefore remains essential, both for verifying signal integrity and for ensuring that the final interpretation is clinically appropriate. Within this workflow, automated systems have the potential to serve as stable and reliable assistants that provide consistent, reproducible, and fatigue-free preliminary assessments.

Whereas human scorers inevitably show intra- and inter-individual variability in their decisions, automated models produce the same output for identical input, thereby offering a reproducible reference against which human scoring can be calibrated. Such reproducibility may support scorer training, promote harmonization across institutions, and contribute to the development of more standardized scoring practices. Integrating model predictions into educational contexts may further enable constructive feedback loops: experts refine the model using discrepant or challenging cases, and the improved model, in turn, helps new scorers acquire more consistent decision patterns.

Beyond scorer variability, several structural issues in current PSG practice limit the achievable level of standardization, regardless of whether scoring is performed manually or automatically. Although the AASM scoring manual provides standardized rules for sleep staging and event detection, a number of criteria remain inherently ambiguous or allow scorer discretion. This flexibility can lead to systematic differences across laboratories, highlighting the need for clearer operational definitions and more objective thresholds. Automated systems may also help clarify where guideline refinement is most urgently needed by making inconsistencies more visible.

An additional challenge lies in the considerable heterogeneity that characterizes current PSG acquisition systems. Across devices, sensor technologies, embedded filtering approaches, amplifier behavior, synchronization mechanisms, and overall hardware stability can vary in meaningful ways, yet explicit performance specifications—such as calibration tolerances, amplitude fidelity, minimum sampling

requirements, or standards for sensor-placement precision—are often defined only in broad terms and do not fully constrain device-dependent variability. Even when calibration procedures are performed prior to recording, the lack of clearly articulated accuracy criteria means that two systems deemed "AASM-compliant" may nonetheless produce physiologically different signal representations. Such variability places fundamental limits on standardization: even perfectly harmonized scoring guidelines cannot ensure uniform outcomes when the underlying physiological inputs themselves are not measured in a consistent and reproducible manner.

More broadly, PSG acquisition lacks many of the engineering rigor and quality-control practices common to other physiological measurement domains. Sensor placement techniques vary considerably between technicians; synchronization between EEG, EOG, EMG, airflow, and respiratory effort channels is not routinely quantified; device-level noise characteristics and dropout rates differ widely; and movement- or position-induced artifacts remain insufficiently addressed. Establishing clearer technical standards and minimum performance specifications for PSG devices—along with standardized calibration protocols and evaluation metrics—would substantially enhance the reliability of both human and automated scoring and form the foundation for truly interoperable datasets.

Looking ahead, a synergistic interaction between human expertise, improved acquisition standards, refined guidelines, and automated analysis may facilitate continuous improvement in both scoring reliability and clinical workflow. As methodological and engineering foundations become more consistent, automated systems may increasingly support tasks such as preliminary triage, quality control, and large-scale research curation, while final diagnostic responsibility appropriately remains with trained clinicians.

Acknowledgements

The authors thank Ms. Sayaka Chino, Mr. Kazuyoshi Namba, and Ms. Masako Sato for their invaluable contributions to the annotation of the polysomnography recordings. The authors appreciate Mr. Noboru Oki of Y.K. Norupro Light Systems for his technical support during the data analysis process. The authors further acknowledge Dr. Koji Ode, Dr. Akifumi Kishi, and Dr. Hiroki Ueda of the University of Tokyo for their constructive discussions and perspectives that helped shape the conceptual foundation of this research.

Natural language processing tools were used to assist with language editing.

Disclosure statement

Financial Disclosure: This research was designed, conducted, and funded by ACCELStars, Inc. All authors are financially affiliated with ACCELStars, Inc. in relation to this work.
Non-financial Disclosure: None.

References

1. Juan ID, Carmenates LIP. Obstructive sleep apnea syndrome. *Journal of Otolaryngology-ENT Research*. 2023. doi.org/10.15406/joentr.2023.15.00533.
2. Lavalle S, Masiello E, Iannella G, et al. Unraveling the Complexities of Oxidative Stress and Inflammation Biomarkers in Obstructive Sleep Apnea Syndrome: A Comprehensive Review. *Life*. 2024;14. doi.org/10.3390/life14040425.
3. Ong J, Crawford M. Insomnia and Obstructive Sleep Apnea. *Sleep medicine clinics*. 2013;8 3:389-98. doi.org/10.1016/j.jsmc.2013.04.004.
4. Fava C, Montagnana M, Favaloro E, Guidi G, Lippi G. Obstructive Sleep Apnea Syndrome and Cardiovascular Diseases. *Seminars in Thrombosis and Hemostasis*. 2011;37(03):280–297. doi.org/10.1055/s-0031-1273092.
5. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MW, Morrell MJ. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*. 2019;7(8):687-98. doi.org/10.1016/S2213-2600(19)30198-5.
6. Chang JL, Goldberg AN, Alt JA, et al. International Consensus Statement on Obstructive Sleep Apnea. *International Forum of Allergy & Rhinology*. 2023;13(7):1061-482. doi.org/10.1002/alr.23079.
7. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. *Journal of Clinical Sleep Medicine*. 2017;13(03):479-504. doi.org/10.5664/jcsm.6506.
8. Penzel T, Conradt R. Computer based sleep recording and analysis. *Sleep Medicine Reviews*. 2000;4(2):131-48. doi.org/10.1053/smrv.1999.0087.
9. 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. *Journal of Clinical Sleep Medicine*. 2022;18(5):1419-25. doi.org/10.5664/jcsm.9952.
10. 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. *Journal of Clinical Sleep Medicine*. 2021;17(6):1287-93. doi.org/10.5664/jcsm.9240.
11. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *Journal of Sleep Research*. 2004;13(1):63-9. doi.org/10.1046/j.1365-2869.2003.00375.x.
12. DANKER-HOPFE H, ANDERER P, ZEITLHOFER J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*. 2009;18(1):74-84. doi.org/10.1111/j.1365-2869.2008.00700.x.
13. Rosenberg RS, Hout SV. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine*. 2013;09(01):81-7. doi.org/10.5664/jcsm.2350.

14. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. *Journal of Clinical Sleep Medicine*. 2022;18(1):193-202. doi.org/10.5664/jcsm.9538.
15. Cesari M, Stefani A, Penzel T, et al. Interrater sleep stage scoring reliability between manual scoring from two European sleep centers and automatic scoring performed by the artificial intelligence–based Stanford-STAGES algorithm. *Journal of Clinical Sleep Medicine*. 2021;17(6):1237–1247. doi.org/10.5664/jcsm.9174.
16. Koley B, Dey D. An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine*. 2012;42(12):1186-95. doi.org/10.1016/j.compbiomed.2012.09.012.
17. Khalighi S, Sousa T, Pires G, Nunes U. Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels. *Expert Systems with Applications*. 2013;40(17):7046-59. doi.org/10.1016/j.eswa.2013.06.023.
18. Zhang Y, Zhang X, Liu W, et al. Automatic Sleep Staging using Multi-dimensional Feature Extraction and Multi-kernel Fuzzy Support Vector Machine. *Journal of Healthcare Engineering*. 2014;5(4):410705. doi.org/10.1260/2040-2295.5.4.505.
19. Lajnef T, Chaibi S, Ruby P, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of Neuroscience Methods*. 2015;250:94-105. Cutting-edge EEG Methods. doi.org/10.1016/j.jneumeth.2015.01.022.
20. Giannakeas N. EEG-Based Automatic Sleep Stage Classification. *Biomedical Journal of Scientific & Technical Research*. 2018;7(4). doi.org/10.26717/bjstr.2018.07.001535.
21. Gunnarsdottir KM, Gamaldo CE, Salas RME, Ewen JB, Allen RP, Sarma SV. A Novel Sleep Stage Scoring System: Combining Expert-Based Rules with a Decision Tree Classifier. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC);2018. p. 3240-3. doi.org/10.1109/EMBC.2018.8513039.
22. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2017;25(11):1998-2008. doi.org/10.1109/TNSRE.2017.2721116.
23. Perslev M, Jensen MH, Darkner S, Jennum PJ, Igel C. In: U-Time: a fully convolutional network for time series segmentation applied to sleep staging. Red Hook, NY, USA: Curran Associates Inc.;2019. .
24. Eldele E, Chen Z, Liu C, et al. An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2021;29:809-18. doi.org/10.1109/TNSRE.2021.3076234.
25. Li C, Qi Y, Ding X, Zhao J, Sang T, Lee M. A Deep Learning Method Approach for Sleep Stage Classification with EEG Spectrogram. *International Journal of Environmental Research and Public Health*. 2022;19(10):6322. doi.org/10.3390/ijerph19106322.

26. Phan H, Chén OY, Tran MC, Koch P, Mertins A, De Vos M. XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(9):5903-15. doi.org/10.1109/TPAMI.2021.3070057.
27. Pradeepkumar J, Anandakumar M, Kugathasan V, et al. Toward Interpretable Sleep Stage Classification Using Cross-Modal Transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024;32:2893-904. doi.org/10.1109/TNSRE.2024.3438610.
28. Lee H, Choi YR, Lee HK, et al. Explainable vision transformer for automatic visual sleep staging on multimodal PSG signals. *npj Digital Medicine*. 2025;8(1). doi.org/10.1038/s41746-024-01378-0.
29. Van Der Donckt J, Van Der Donckt J, Deprost E, et al. Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring. *Biomedical Signal Processing and Control*. 2023;81:104429. doi.org/10.1016/j.bspc.2022.104429.
30. Espiritu H, Metsis V. Automated Detection of Sleep Disorder-Related Events from Polysomnographic Data. In: 2015 International Conference on Healthcare Informatics;2015. p. 562-9. doi.org/10.1109/ICHI.2015.105.
31. Ugur TK, Erdamar A. An efficient automatic arousals detection algorithm in single channel EEG. *Computer Methods and Programs in Biomedicine*. 2019;173:131-8. doi.org/10.1016/j.cmpb.2019.03.013.
32. Liu Y, Liu H, Yang B. Automatic Sleep Arousals Detection From Polysomnography Using Multi-Convolution Neural Network and Random Forest. *IEEE Access*. 2020;8:176343-50. doi.org/10.1109/ACCESS.2020.3026814.
33. Rohan R, Kumari LVR. Classification of Sleep Apneas using Decision Tree Classifier. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS);2021. p. 1310-6. doi.org/10.1109/ICICCS51141.2021.9432197.
34. Chambon S, Thorey V, Arnal PJ, Mignot E, Gramfort A. DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal. *Journal of Neuroscience Methods*. 2019;321:64-78. doi.org/10.1016/j.jneumeth.2019.03.017.
35. Brink-Kjaer A, Olesen AN, Peppard PE, et al. Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness. *Clinical Neurophysiology*. 2020;131(6):1187-203. doi.org/10.1016/j.clinph.2020.02.027.
36. Li H, Guan Y. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Communications Biology*. 2021;4(1). doi.org/10.1038/s42003-020-01542-8.
37. 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 Transactions on Biomedical Engineering*. 2022;69(6):2094-104. doi.org/10.1109/TBME.2021.3136753.
38. Zahid AN, Jennum P, Mignot E, Sorensen HBD. MSED: A Multi-Modal Sleep Event Detection Model for Clinical Sleep Analysis. *IEEE Transactions on Biomedical Engineering*. 2023;70(9):2508-18. doi.org/10.1109/TBME.2023.3252368.

39. Nasiri S, Ganglberger W, Nassi T, et al. CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics. *Sleep*. 2025;48(8):zsaf134. doi.org/10.1093/sleep/zsaf134.
40. Rechtschaffen A, Kales A. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Public Health Service. Washington DC: US Government Printing Office;1968.
41. Imtiaz SA, Rodriguez-Villegas E. Recommendations for performance assessment of automatic sleep staging algorithms. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society;2014. p. 5044-7. doi.org/10.1109/EMBC.2014.6944758.
42. Stephansen JB, Olesen AN, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*. 2018;9(1). doi.org/10.1038/s41467-018-07229-3.
43. Bakker JP, Ross M, Cerny A, et al. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. *Sleep*. 2022;46(2):zsac154. doi.org/10.1093/sleep/zsac154.
44. Anderer P, Ross M, Cerny A, Vasko R, Shaw E, Fonseca P. Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing. *Frontiers in Sleep*. 2023;Volume 2 - 2023. doi.org/10.3389/frsle.2023.1163477.
45. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*. 2021;4(1). doi.org/10.1038/s41746-021-00440-5.
46. Thorey V, Hernandez AB, Arnal PJ, During EH. AI vs Humans for the diagnosis of sleep apnea. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC);2019. p. 1596-600. doi.org/10.1109/EMBC.2019.8856877.
47. of Sleep Medicine AA. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.6 ed. Darien, IL: American Academy of Sleep Medicine;2020.
48. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.;2017. p. 3149–3157. doi.org/10.5555/3294996.3295074.
49. Welch P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*. 1967;15(2):70-3. doi.org/10.1109/TAU.1967.1161901.
50. Harrison SJ, Bianchi S, Heinzle J, Stephan KE, Iglesias S, Kasper L. A Hilbert-based method for processing respiratory timeseries. *NeuroImage*. 2021;230:117787. doi.org/10.1016/j.neuroimage.2021.117787.
51. Martin Bland J, Altman D. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet*. 1986;327(8476):307-10. Originally published as Volume 1, Issue 8476. doi.org/10.1016/S0140-6736(86)90837-8.

52. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR);2016. p. 779-88. doi.org/10.1109/CVPR.2016.91.
53. Guillot A, Sauvet F, During EH, Thorey V. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2020;28(9):1955-65. doi.org/10.1109/TNSRE.2020.3011181.