HHS Public Access

Author manuscript

J Clin Sleep Med. Author manuscript; available in PMC 2026 January 21.

Published in final edited form as:

J Clin Sleep Med. 2025 November 01; 21(11): 1821–1829. doi:10.5664/jcsm.11848.

Automated Analysis of the AASM Inter-Scorer Reliability Gold Standard Polysomnogram Dataset

Ayush Tripathi, PhD[1,2], **Samaneh Nasiri, PhD**[1,2,3,4], **Wolfgang Ganglberger, PhD**[1,2], **Thijs Nassi, PhD**[1,2,5], **Erik-Jan Meulenbrugge, MSc**[1,2], **Haoqi Sun, PhD**[1,2,3], **Katie L. Stone, PhD**[6,7], **Emmanuel Mignot, MD,PhD**[8], **Dennis Hwang, MD**[9], **Lynn Marie Trotti, MD,MSc**[10], **Matthew A. Reyna, PhD**[4], **Gari D. Clifford, PhD**[4,11], **Umakanth Katwa, MD**[2,12], **Robert J Thomas, MD**[†,2,13], **M Brandon Westover, MD,PhD**[†,1,2]

[1]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA

[2]Harvard Medical School, Boston, MA, USA

[3]Department of Neurology, Massachusetts General Hospital, 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

[6]Department of Epidemiology and Biostatistics, University of California, San Francisco, USA

[7]California Pacific Medical Center Research Institute, San Francisco, CA, USA

[8]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]Boston Children's Hospital, Boston, MA, USA

[13]Department of Medicine, Division of Pulmonary Critical Care & Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

Abstract

Study Objectives: To compare the performance of a comprehensive automated polysomnogram (PSG) analysis algorithm—CAISR (Complete Artificial Intelligence Sleep Report)—to a multi-expert gold standard panel, crowdsourced scorers, and experienced technicians for sleep staging and detecting arousals, respiratory events, and limb movements.

[*]**Corresponding author**: Michael Brandon Westover, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA 02215 (mwestove@bidmc.harvard.edu).

[†]co senior authors

Institution where work was performed: Beth Israel Deaconess Medical Center.

Author Approval: All authors have seen and approved the final manuscript.

Clinical Trial: This manuscript does not report the results of a clinical trial.

Methods: A benchmark dataset of 57 PSG records (Inter-Scorer Reliability dataset) with 200 30-second epochs scored per AASM guidelines was used. Annotations were obtained from (1) the AASM multi-expert gold standard panel, (2) AASM Inter-Scorer Reliability (ISR) platform users ("crowd," averaging 6,818 raters per epoch), (3) three experienced technicians, and (4) CAISR. Agreement was assessed via Cohen's Kappa (κ) and percent agreement.

Results: Across tasks, CAISR achieved performance comparable to experienced technicians but did not match consensus-level agreement between the multi-expert gold standard and the crowd. For sleep staging, CAISR's agreement with multi-expert gold standard was 82.1% (κ = 0.70), comparable to experienced technicians but below the crowd (κ = 0.88). Arousal detection showed 87.81% agreement (κ = 0.45), respiratory event detection 83.18% (κ = 0.34), and limb movement detection 94.89% (κ = 0.11), each aligning with performance equivalent to experienced technicians but trailing crowd agreement (κ = 0.83, 0.78 and 0.86 for detection of arousal, respiratory events and limb movements respectively).

Conclusions: CAISR achieves experienced technician-level accuracy for PSG scoring tasks but does not surpass the consensus-level agreement of a multi-expert gold standard or the crowd. These findings highlight the potential of automated scoring to match experienced technician-level performance while emphasizing the value of multi-rater consensus.

Brief Summary:

A state-of-the-art automated artificial intelligence approach can achieve performance comparable to that of individual experienced technicians across all standard PSG tasks. This result underscores the potential of AI to enhance scoring efficiency, standardize PSG analysis, and address resource challenges in sleep medicine. However, CAISR does not yet match the multi-expert gold standard or large-scale crowd annotations, highlighting opportunities to improve future models through enhanced training datasets.

Keywords

Polysomnography; sleep staging; artificial intelligence; inter-rater reliability; arousal detection; respiratory events; limb movement

Introduction

Sleep is a vital physiological process essential for cognitive function, immune regulation, metabolic homeostasis, and emotional well-being. Insufficient or disordered sleep is linked to a wide array of adverse health outcomes, including cardiovascular disease, neurodegenerative disorders, impaired glucose metabolism, and increased mortality risk[1-7]. Accurate and consistent assessment of sleep quality and architecture is thus central to both clinical diagnosis and population-level sleep research.

Polysomnography (PSG) is the gold standard for evaluating sleep architecture and identifying sleep-related disorders, including obstructive sleep apnea (OSA), periodic limb movement disorder (PLMD), and insomnia[8,9]. Standard PSG assessments involve the recording of multiple physiological signals—including electroencephalogram (EEG) (brain activity), electrocardiogram (ECG) (heart activity), respiratory signals, electromyography

(EMG) (muscle activity), and oxygen saturation—followed by manual scoring by trained technologists in 30-second epochs in accordance with American Academy of Sleep Medicine (AASM) guidelines[10,11]. While reliable when performed by experts, manual PSG scoring is labor-intensive, time-consuming, and prone to substantial inter- and intra-rater variability[12-14]. This variability has been well documented for sleep staging[15,16], arousal detection[17] and respiratory event classification[18], contributing to inconsistencies in clinical diagnoses and treatment decisions.

Over the past decade, the rapid advancement of artificial intelligence (AI) and deep learning has catalyzed the development of automated PSG analysis tools aimed at reducing human workload and improving scoring consistency. Several deep learning models have shown promising results in automated sleep staging using single or multi-channel EEG data[18-23], achieving performance comparable to that of expert scorers on various benchmark datasets. Notably, models like U-Sleep[24], SeqSleepNet[25], and SleepTransformer[26] have introduced robust, generalizable architectures capable of handling noisy signals and variable sleep patterns. In parallel, other efforts have targeted specific PSG tasks such as arousal detection[27,28], periodic limb movement classification[29,30], and respiratory event recognition[31-33], often using supervised learning on hand-scored datasets. However, most existing systems are specialized for individual tasks, lack cross-task generalization, or are trained on datasets with limited ground-truth annotations and uncertain scoring consistency.

A key limitation in the field has been the scarcity of gold-standard datasets that combine expert consensus scoring with large-scale inter-rater comparisons. The AASM Inter-Scorer Reliability (ISR) dataset represents a unique and underutilized resource for addressing this gap. It contains 57 PSG studies with 200 pre-selected 30-second epochs per study, each scored by a five-member gold standard expert panel that includes contributors to the AASM manual itself. In addition, each epoch has been scored by thousands of registered ISR users —primarily certified technologists—yielding robust consensus labels based on real-world variability. While the ISR dataset is modest in size, its rigorous expert consensus and breadth of crowd-based annotations offer an unprecedented opportunity to benchmark both human and algorithmic performance under standardized conditions.

In this study, we evaluated the “Complete Artificial Intelligence Sleep Report” (CAISR), a unified AI system designed to perform comprehensive PSG analysis across four core domains: sleep staging, arousal detection, respiratory event detection, and limb movement identification. Unlike many task-specific algorithms, CAISR is engineered for end-to-end sleep report generation, integrating multiple deep learning modules optimized for different signal modalities and event types. Although CAISR has previously demonstrated strong generalization across large-scale clinical datasets (see Methods), its performance against a multi-expert gold standard and high-density crowd ratings has not been previously reported.

Here, we leverage the ISR dataset to assess how CAISR compares to individual experienced technicians, aggregated ISR user consensus, and the multi-expert Gold Standard panel across the four key PSG tasks. By analyzing inter-rater agreement using Cohen’s Kappa and percent agreement, we aim to evaluate CAISR’s alignment with expert practice, quantify

the limitations of current automated methods, and propose future directions for AI-based standardization in sleep medicine.

Methods

Automated Sleep Scoring: The CAISR System[34]

The Complete Artificial Intelligence Sleep Report (CAISR) is a comprehensive software system designed to automate conventional polysomnographic (PSG) scoring, including sleep staging, arousal detection, respiratory event detection, and limb movement identification.

CAISR processes a subset of PSG signals commonly used for manual scoring, including EEG (C3-M2, C4-M1), EOG (E1-M2, E2-M1), ECG, chin EMG, thoracic and abdominal effort belts, airflow, oxygen saturation, and limb EMGs. Signal selection varies by task: EEG, EOG, ECG, chin EMG, and respiratory effort signals are used for sleep staging and arousal detection. For respiratory event detection, oronasal flow/pressure, respiratory rate, respiratory effort, oxygen saturation signals are utilized, while leg EMG signals are used for limb movement analysis. Signals are bandpass filtered, notch filtered at 60 Hz, and downsampled based on individual tasks for efficient training. For sleep staging, 30-second non-overlapping epochs are used; for the other tasks, segment-wise scoring is performed at 1-second resolution.

CAISR comprises four independent models, each optimized for a specific PSG task:

- Sleep staging is performed using ProductGraphSleepNet[19], which combines several components such as: 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.
- Arousal detection uses a U-Sleep[24] inspired architecture designed to detect short-duration arousals during sleep.
- Respiratory event detection is based on a signal processing-based model that classifies each second of data into one of multiple event types (e.g., obstructive apnea, hypopnea, central apnea, mixed apnea, and respiratory effort related arousal).
- The CAISR limb movement automatic detector was adapted from the Ferri model[30] which was developed based on AASM rules. Unlike the original Ferri model, CAISR utilizes Variable Amplitude Thresholding (VAT) to address signal quality issues.

The model utilized in this study was trained with a large (N = 21,764 subjects) multicenter dataset that included PSGs from one clinical sleep center (MGH) and three large epidemiological cohorts (SHHS, MESA, MrOs), see Table 1. These data were recorded between 125 and 512 Hz and resampled to 200 Hz for analysis. Each model was trained independently while employing task-specific loss functions to handle class imbalance, and early stopping was used to prevent overfitting.

ISR Dataset and Study Design

This study utilized PSG data from the American Academy of Sleep Medicine's Sleep Inter-Scorer Reliability program[35], a widely adopted resource for monitoring and improving the scoring accuracy of sleep technologists. The ISR platform, used by thousands of individual scorers and over 2,300 sleep centers, supplies monthly sets of 200 pre-selected 30-second epochs from 57 PSG recordings. The demographic details of the AASM ISR dataset are presented in Table 2. These epochs are purposefully chosen to represent a variety of sleep stages, arousals, respiratory events, and limb movements, thus reflecting clinically relevant scenarios. A distinguished panel of experts—who helped develop the AASM guidelines—provided a rigorous multi-expert gold standard for comparison, enabling participants to gauge and refine their scoring consistency. Beyond serving as a practice tool, the ISR program is also used to fulfill accreditation requirements, deliver continuing education credits, and support quality assurance efforts across sleep centers worldwide.

Annotations for each epoch were provided in CSV files (answers.csv), detailing scoring outcomes from the ISR multi-expert gold standard panel and ISR users. The following PSG features were evaluated:

- Sleep Stage – Wake (W), N1, N2, N3, REM (R)
- Respiratory Events – Obstructive Apnea, Central Apnea, Mixed Apnea, Hypopnea
- Limb Movements – presence or absence within the epoch
- Arousals –presence or absence within the epoch

Each CSV file reported the multi-expert gold standard labels and the most common answer given by ISR users for each epoch. Scores from individual members of the ISR panel or from individual members of the "crowd" were not available.

The study was conducted under a central IRB at BIDMC (IRB # 2024P000804, 2022P000417) and a data use agreement with the ISR scoring committee, which allowed access to the dataset for analysis.

Participants and Scoring

The primary objective of this study was to evaluate CAISR's performance on the ISR benchmark dataset by comparing its outputs to three sources of annotations:

1. **Multi-expert Gold Standard** – Each epoch was scored by five highly experienced PSG scorers, the ISR Gold Standard Panel assembled by the AASM.
1. **Crowd** – Thousands of ISR users scored each epoch (averaging 6,818.27 ± 553 raters, consisting primarily of sleep technicians and trainees), and the most frequently chosen label became the "crowd" score.
2. **Independent Experienced Technicians** – Three additional experienced sleep scoring technicians from an R1 university-affiliated sleep lab scored each epoch,

enabling in-depth comparisons of inter-rater agreement and alignment with CAISR's predictions.

PSG Tasks and Labeling

We evaluated CAISR's performance across four PSG analysis tasks. Sleep staging and respiratory event detection were multi-class classification tasks, while arousal and limb movement detection were converted to binary classification to simplify analysis and mitigate class imbalance—a condition where one type of event (e.g., no arousal or no limb movement) occurs much more frequently than the other, which can bias automated learning systems and distort evaluation metrics.

Sleep Staging –: Each 30-second epoch was labeled as one of five sleep stages based on AASM scoring guidelines: Wake (W), N1 (Stage 1), N2 (Stage 2), N3 (Stage 3, deep sleep), or REM (R). The distribution of epochs across sleep stages was as follows: N1 – 716, N2 – 7002, N3 – 573, REM – 2600, Wake – 509.

Arousal Detection –: Annotations for arousals were coded into binary format. Epochs with one or more arousals were labeled 1, while epochs without arousals were labeled 0. The dataset contained 952 arousal epochs and 10,448 non-arousal epochs.

Respiratory Event Detection –: Respiratory events were categorized as Obstructive Apnea (OA), Central Apnea (CA), Mixed Apnea (MA), or Hypopnea (H). Epochs with no respiratory events were labeled as None. The distribution of epochs for respiratory events included: Central Apnea – 84, Hypopnea – 1024, Mixed Apnea – 15, Obstructive Apnea – 96, and None – 10,181.

Limb Movement Detection –: Limb movement annotations were similarly converted to binary format. Epochs with one or more limb movements were labeled 1, while those without limb movements were labeled 0. The dataset comprised 422 limb movement epochs and 10,978 non-limb movement epochs.

Evaluation Metrics and Statistical Analysis

Model performance and inter-rater reliability were assessed using percent agreement and Cohen's Kappa (κ). Percent agreement is simply the percentage of epochs where scores from two different sources match. Cohen's Kappa measures agreement after discounting agreement that could occur by chance[36].

Pairwise Cohen's Kappa (κ) values were calculated by pooling all scores across all recordings to generate a single Kappa value for each pair of scoring sources. This was done for all combinations of scoring sources across each of the four PSG tasks—sleep staging, arousal detection, respiratory event detection, and limb movement detection. Heatmaps were utilized to visualize the results. To further quantify performance, bar plots were created comparing CAISR, crowd ratings, and the three experienced technicians against the multi-expert gold standard. Confidence intervals on these bar plots were generated by applying bootstrapping with 1,000 iterations.

Results

Figure 1 presents heatmaps illustrating pairwise inter-rater reliability for each PSG task, and Figure 2 provides bar plots comparing each scorer's performance against the multi-expert gold standard. The four primary tasks analyzed were sleep staging, arousal detection, respiratory event detection, and limb movement detection. Both percent agreement and Cohen's Kappa (κ) were used to quantify performance, with Cohen's Kappa accounting for chance agreement. While high percent agreement indicates consistency in scoring, low Kappa values can reflect challenge with class imbalance—a scenario where some categories (e.g., presence of an arousal) are far less common than others (e.g., absence of arousal)—as well as the difficulty of reliably identifying rare or borderline events. These factors can disproportionately affect agreement metrics and mask true performance.

Sleep staging

For sleep staging, the highest agreement was observed between the Gold Standard and the crowd, with 93.54% agreement and a Cohen's Kappa of 0.88. Among the experienced technicians, Tech 1 demonstrated the strongest alignment with the Gold Standard (88.75%, $\kappa = 0.80$), followed by Tech 3 (82.25%, $\kappa = 0.69$), while Tech 2 showed lower alignment (75.64%, $\kappa = 0.57$). Inter-technician agreement revealed an average percent agreement of 82.21% (range: 75.64%–88.75%) and an average Kappa of 0.69 (range: 0.57–0.80). The strongest inter-technician agreement was between technicians 1 and 3, with 82.56% agreement and $\kappa = 0.70$, while technician 2's scores diverged more significantly from the other raters.

CAISR achieved 82.07% agreement ($\kappa = 0.70$) with the multi-expert gold standard, closely aligning with the average performance of the individual experienced technicians. Its agreement with individual technicians ranged from 72.32% ($\kappa = 0.54$) with Tech 2 to 83.04% ($\kappa = 0.72$) with Tech 3, reflecting variability in alignment with different technicians. These results demonstrate that CAISR performs at a level comparable to individual experienced technicians in sleep staging.

Arousal detection

Arousal detection showed strong agreement between the multi-expert gold standard and the crowd, with 97.47% agreement and $\kappa = 0.83$. Among the experienced technicians, Tech 1 was most closely aligned with the multi-expert gold standard (90.08%, $\kappa = 0.54$), while Tech 3 showed comparable percent agreement (90.13%) but a lower Kappa value ($\kappa = 0.43$), indicating slightly less consistency after accounting for chance. Tech2 achieved the highest percent agreement with multi-expert gold standard rating (92.17%) but exhibited a moderate Kappa value of $\kappa = 0.50$, reflecting variability in arousal detection.

Inter-technician agreement for arousal detection averaged 90.79% (range: 90.08%–92.17%), with an average Kappa of 0.49 (range: 0.43–0.54). CAISR achieved 87.81% agreement ($\kappa = 0.45$) with the multi-expert gold standard, comparable to the range of technician performance. When comparing CAISR to individual experienced technicians, the highest alignment was observed with Tech1 (89.29%, $\kappa = 0.61$), while the agreement with Tech 2

was lower (85.78%, $\kappa = 0.37$). These results highlight that the performance of CAISR is as reliable as an individual experienced technician for the arousal detection task.

Respiratory event detection

Respiratory event detection exhibited lower agreement overall compared to sleep staging and arousal detection. The highest agreement was observed between the multi-expert gold standard and the crowd (96.11%, $\kappa = 0.78$). Among the individual experienced technicians, Tech 1 showed the strongest alignment with the Gold Standard (90.39%, $\kappa = 0.49$), while Tech 2 demonstrated the lowest agreement (86.61%, $\kappa = 0.31$). Inter-technician agreement averaged 89.13% (range: 86.61%–90.39%) with a mean Kappa of 0.40 (range: 0.31–0.49), reflecting substantial variability among the three experienced technicians.

CAISR achieved 83.18% agreement ($\kappa = 0.34$) with the multi-expert gold standard, placing it within the range of experienced technician-level performance. Its agreement with individual technicians ranged from 80.94% ($\kappa = 0.25$) with Tech 2 to 83.22% ($\kappa = 0.33$) with Tech 1. Despite the relatively high percent agreement values, the lower Kappa values indicate challenges associated with quantifying reliability of scoring rare respiratory events, such as mixed apneas.

Limb movement detection

Limb movement detection demonstrated high percent agreement across scorers, but Kappa values were lower, reflecting the substantial class imbalance, where non-movements dominate the dataset. The highest agreement was observed between the multi-expert gold standard and the crowd (99.04%, $\kappa = 0.86$). Among the individual experienced technicians, Tech 3 exhibited the strongest alignment with the Gold Standard (97.33%, $\kappa = 0.57$), while Tech 2 showed lower agreement (95.50%, $\kappa = 0.12$). Inter-technician agreement averaged 96.06% (range: 95.50%–97.33%) with a mean Kappa of 0.33 (range: 0.12–0.57).

CAISR achieved 94.89% agreement ($\kappa = 0.11$) with the multi-expert gold standard, comparable to the lower range of individual experienced technician performance. Its agreement with individual technicians ranged from 95.45% ($\kappa = 0.04$) with Tech 3 to 97.07% ($\kappa = 0.10$) with Tech 1. The low Kappa values, despite high percent agreement, highlight the limitations of binary classification for limb movement detection, where chance agreement contributes disproportionately to the observed concordance.

Discussion

In this study, we evaluated CAISR (Complete Artificial Intelligence Sleep Report) on the Sleep ISR dataset, comparing its performance to the ISR multi-expert gold standard panel, large-scale crowd consensus, and three experienced technician scores. Our primary finding is that CAISR achieved accuracy comparable to individual experienced technicians for sleep staging, arousal detection, respiratory event detection, and limb movement detection, thereby validating its potential as an automated tool for PSG analysis. However, neither CAISR nor any single experienced technician matched the higher consensus-level agreement observed between the multi-expert gold standard and the crowd. These results underscore both the promise and the limitations of current automated methods: while CAISR can alleviate the

manual burden and reduce scoring variability, additional refinements are needed to meet the more rigorous multi-expert standard. The results also highlighted differences between highly experienced technicians and the role of CAISR and similar systems to provide learning feedback.

Inter-Rater Variability and the Value of Crowd Consensus

One of the most notable observations is the consistently high agreement between the ISR multi-expert Gold Standard and the crowd labels. Sleep staging and arousal detection both exceeded 90% concordance, and even the more challenging tasks—respiratory and limb movement detection—demonstrated strong consensus. These results highlight the value of large-scale crowd scoring in achieving multi-expert gold standard-level agreement—a phenomenon that individual scorers, including both experienced technicians and CAISR, generally fail to match. This outcome is consistent with the well-known statistical principle of the “wisdom of the crowd,” where aggregating responses from a diverse pool of raters can converge toward a more accurate consensus than any single rater. While our analysis relied on majority-vote labels, scores from individual ISR users and individual members of the Gold Standard panel were not available. As a result, we were unable to explore within-group disagreement or uncertainty. Future work may benefit from modeling full score distributions or using probabilistic consensus methods that go beyond simple majority voting to better capture scoring variability and latent ambiguities.

Automated Performance and Limitations

CAISR nonetheless aligns closely with individual experienced technicians across all tasks. Its performance in sleep staging and arousal detection is particularly robust, suggesting that these tasks, which rely heavily on EEG patterns, may be especially amenable to automation at scale. However, assessment of performance on respiratory event and limb movement detection posed greater challenges. The relatively lower Cohen’s Kappa values point to two key issues. First, respiratory events can be subtle: differentiating obstructive, central, and mixed apneas is challenging, leading to relatively high variability and low Kappa values for both CAISR and for individual experienced technicians. Scoring rules also changed during the course of the data collected, introducing some additional variance (e.g., degree of desaturation linked to hypopneas). Respiratory Effort Related arousals were not scored and have a somewhat fluid boundary with arousal/3% desaturation scoring tag. Second, for both limb movements and respiratory events, non-events far outnumber events. This imbalance leads to low kappa values even in the presence of high percent agreement. Nevertheless, these issues also apply to the performance of individual experts.

There are at least two sets of periodic limb movement scoring systems: the AASM rule secludes PLMS associated with respiratory events, while that of the International Restless Legs Group (IRLSG) scores and categorizes them. Our experienced technicians normally follow the IRLSG guidelines. Though they were asked to use the AASM rules for this project, some pattern bias was inevitable.

These results do not address an important practical issue - data quality. The ISR segments are relatively pristine/high-quality signals, while clinical PSG data can have

variable quality, complex pathologies, signal dropout, artifacts and movement-related signal distortion. Automated systems need to be evaluated against such studies too, to enhance generalizability.

Additionally, future evaluations of automated scoring systems should incorporate complementary performance metrics such as precision, recall, and F1-score, especially for event-driven tasks like respiratory event and limb movement detection. These metrics can offer a more nuanced understanding of model performance on imbalanced tasks where agreement-based measures like Cohen's Kappa may underrepresent detection quality for rare but clinically important events.

Clinical Relevance and Integration

Despite these limitations, CAISR holds significant clinical promise. By automating routine PSG analysis, the model can streamline workflows, mitigate scorer fatigue, and help standardize results across facilities. This is especially relevant as demand for sleep studies continues to rise. For tasks where CAISR's performance matches individual experienced technicians—such as sleep staging and arousal detection—automated scoring could function as a reliable first pass or adjunct, freeing humans to focus on more complex cases or confirm borderline events.

Nonetheless, caution is warranted before fully relying on CAISR in critical diagnostic workflows. Given its lower Kappa values for respiratory and limb movement detection, human oversight remains essential. In practice, a blended approach—where CAISR pre-scores epochs and humans review flagged or ambiguous events—may maximize both efficiency and accuracy.

Future Directions

The findings of this study demonstrate that CAISR performs at the level of individual experienced technicians across a variety of PSG analysis tasks. While this is a significant achievement, advancing CAISR to meet or exceed the performance of the multi-expert gold standard will require careful expansion of training and testing datasets.

One key step is the development of a large-scale, multi-expert gold standard dataset similar to the ISR dataset but encompassing a much larger and more diverse patient population. Unlike current single-scored datasets, which now exceed 100,000 patients, such a multi-expert gold standard dataset would necessarily be smaller due to the complexity of multi-expert scoring. However, it could be designed to maximize the diversity of patient characteristics, including age, sex, comorbidities, and sleep disorders. This diversity would enable CAISR to better generalize to real-world clinical populations. In addition, current datasets—including the AASM ISR—typically use discrete 30-second epochs for scoring, which simplifies the temporal resolution of events like arousals or limb movements. Future efforts could incorporate continuous or high-resolution event annotations, which would allow for more granular evaluation of inter-rater agreement. In such cases, alternative metrics such as Conger's Kappa for continuous behaviors[37] may offer a more accurate assessment of scoring consistency across time.

Currently, CAISR employs separate, task-specific models for sleep staging, arousal detection, respiratory event classification, and limb movement identification. While this modular approach simplifies development and evaluation, it does not account for the known physiological interdependencies between tasks. For example, arousals are often associated with transitions from deeper to lighter sleep or to wake. Future work may explore joint or multi-task learning architectures that can model these dependencies explicitly, potentially improving scoring accuracy and internal consistency across tasks.

To further enhance CAISR's ability to detect rare events, advanced techniques such as few-shot learning could be employed[38]. Fine-tuning using a carefully curated diverse multi-expert gold standard dataset would allow the model to adapt its predictions to match the multi-expert consensus. Additionally, incorporating data from diverse clinical sites would improve the robustness of the algorithm, ensuring that CAISR performs consistently across different healthcare environments.

Finally, implementing CAISR in real-world clinical workflows and refining the model using annotated feedback loops could accelerate its improvement. By learning from diverse patient populations in clinical practice, CAISR can continue to adapt and improve its accuracy, especially in detecting complex and subtle PSG patterns. By focusing on these strategies, CAISR has the potential to progress beyond expert-level performance, achieving the consistency and reliability required to match a true multi-expert Gold Standard.

Conclusion

In summary, CAISR effectively matches individual experienced technicians on key PSG tasks, demonstrating its potential to standardize and expedite sleep study analysis. At the same time, the enhanced performance of crowd-based and multi-expert gold standard reveals that achieving 'superhuman' levels of consistency remains a challenge for both automated tools and individual experienced technicians. Future work should focus on expanding the scale and diversity of gold standard training data and leveraging advanced modeling techniques to fine-tune CAISR to exceed its current performance. With these refinements, automated sleep scoring stands poised to become a cornerstone in sleep medicine, offering an efficient, objective, and scalable solution for today's growing diagnostic demands.

Declarations:

• Financial Support:

This work was funded by grants from the NIH (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119).

• Disclosures:

- Dr. Westover is a co-founder, scientific advisor, and consultant to, and has a personal equity interest in Beacon Biosignals.
- Dr. Clifford has received research funding from the NSF, NIH, and LifeBell AI, and unrestricted donations from AliveCor Inc, Amazon Research, the Center for Discovery, the Gates Foundation, Google, the Gordon and Betty Moore Foundation, MathWorks, Microsoft Research, Nextsense Inc, One Mind Foundation, and the Rett Research Foundation. Dr Clifford has advisory roles and financial interests in AliveCor Inc and Nextsense Inc. He is also the CTO of MindChild Medical with significant stock. These relationships are unconnected to the current work.

- Thijs Nassi is co-inventor of a signal analysis patent for estimating respiratory self-similarity for detection of high loop gain sleep apnea.
- Dr. Thomas is co-inventor of: 1) Cardiopulmonary sleep spectrogram to assess sleep stability/quality and sleep apnea, licensed by the Beth Israel Deaconess Medical Center to MyCardio, LLC; 2) Patent for Enhanced Expiratory Rebreathing Space to treat high loop gain sleep apnea; 3) Patent for estimating respiratory self-similarity for detection of high loop gain sleep apnea. 4) General sleep medicine consulting: GLG Councils, Guidepoint, Beacon Biosignals, Jazz Pharmaceuticals.
- Dr. Trotti is a member of the Board of Directors of the American Academy of Sleep Medicine; any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the American Academy of Sleep Medicine.
- Dr. Stone reports grant funding from Eli Lilly and is consultant for Axsome Therapeutics.
- All other authors report no disclosures.

Data Availability Statement

The data used in the study are not publicly available. Access to the data may be made available on request by contacting the AASM ISR committee (https://isr.aasm.org/).

List of Abbreviations

AASM	American Academy of Sleep Medicine
AI	Artificial Intelligence
AGC	Attentive Graph Convolutional Network
BiGRU	Bidirectional Gated Recurrent Unit
CAISR	Complete Artificial Intelligence Sleep Report
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyography
EOG	Electrooculogram
GAT	Graph-wise Attention
ISR	Inter-Scorer Reliability
Kappa (κ)	Cohen’s Kappa
MA	Mixed Apnea
MGH	Massachusetts General Hospital
MESA	Multi-Ethnic Study of Atherosclerosis
MrOS	Osteoporotic Fractures in Men Study
NREM	Non-Rapid Eye Movement
OA	Obstructive Apnea

PLMD	Periodic Limb Movement Disorder
PSG	Polysomnography
REM	Rapid Eye Movement
RERA	Respiratory Effort Related Arousal
SHHS	Sleep Heart Health Study
SpAtt	Spatial Attention
VAT	Variable Amplitude Thresholding

References

1. Zucconi M, Ferri R, Allen R, et al. The official World Association of Sleep Medicine (WASM) standards for recording and scoring periodic leg movements in sleep (PLMS) and wakefulness (PLMW) developed in collaboration with a task force from the International Restless Legs Syndrome Study Group (IRLSSG). Sleep Med.. 2006;7(2):175–84. [PubMed: 16459136]
2. Zinchuk A, Srivali N, Qin L, et al. Association of Periodic Limb Movements and Obstructive Sleep Apnea With Risk of Cardiovascular Disease and Mortality. J Am Heart Assoc.. 2024;13(3):e031630. [PubMed: 38240208]
3. Zhu J, Sanford LD, Ren R, et al. Multiple machine learning methods reveal key biomarkers of obstructive sleep apnea and continuous positive airway pressure treatment. Front Genet.. 2022;13:927545. [PubMed: 35910196]
4. Zhao Y, Lin X, Zhang Z, et al. STDP-based adaptive graph convolutional networks for automatic sleep staging. Front Neurosci.. 2023;17:1158246. [PubMed: 37152593]
5. Phan H, Andreotti F, Cooray N, et al. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng.. 2019;27(3):400–10. [PubMed: 30716040]
6. Zhang H, Wang X, Li H, et al. Auto-annotating sleep stages based on polysomnographic data. Patterns.. 2022;3(1).
7. Zhang GQ, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc.. 2018;25(10):1351–8. [PubMed: 29860441]
8. 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(2):963–8. [PubMed: 22078131]
9. Ye P, Qin H, Zhan X, et al. Diagnosis of obstructive sleep apnea in children based on the XGBoost algorithm using nocturnal heart rate and blood oxygen feature. Am J Otolaryngol.. 2023;44(2):103714. [PubMed: 36738700]
10. Berry RB, Budhiraja R, Gottlieb DJ, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. J Clin Sleep Med.. 2012;8(5):597–619. [PubMed: 23066376]
11. Berry RB, Abreu AR, Krishnan V, et al. 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(5):1419–25. [PubMed: 35197190]
12. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med.. 2002;3(1):43–7. [PubMed: 14592253]
13. 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(1):63–9. [PubMed: 14996037]

14. Lee YJ, Lee JY, Cho JH, et al. Interrater reliability of sleep stage scoring: a meta-analysis. J Clin Sleep Med.. 2022;18(1):193–202. [PubMed: 34310277]
15. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. J Clin Sleep Med.. 2016;12(6):885–94. [PubMed: 27070243]
16. Muto V, Berthomier C, Schmidt C, et al. 0315 inter-and intra-expert variability in sleep scoring: comparison between visual and automatic analysis. Sleep.. 2018;41(suppl_1):A121-
17. Loredo JS, Clausen JL, Ancoli-Israel S, et al. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep.. 1999;22(7):916–20. [PubMed: 10566909]
18. Magalang UJ, Chen NH, Cistulli PA, et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep.. 2013;36(4):591–6. [PubMed: 23565005]
19. Einizade A, Nasiri S, Sardouie SH, Clifford GD. ProductGraphSleepNet: Sleep staging using product spatio-temporal graph learning with attentive temporal aggregation. Neural Networks.. 2023;164:667–80. [PubMed: 37245479]
20. Tsinalis O, Matthews PM, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Ann Biomed Eng.. 2016;44:1587–97. [PubMed: 26464268]
21. Jirakittayakorn N, Wongsawat Y, Mitrirattanakul S. ZleepAnlystNet: a novel deep learning model for automatic sleep stage scoring based on single-channel raw EEG data using separating training. Sci Rep.. 2024;14(1):9859. [PubMed: 38684765]
22. Mousavi S, Afghah F, Acharya UR. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. PLoS One.. 2019;14(5):e0216456. [PubMed: 31063501]
23. Supratak A, Dong H, Wu C, et al. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng.. 2017;25(11):1998–2008. [PubMed: 28678710]
24. Perslev M, Darkner S, Kempfner L, et al. U-Sleep: resilient high-frequency sleep staging. NPJ Digit Med.. 2021;4(1):72. [PubMed: 33859353]
25. Phan H, Andreotti F, Cooray N, et al. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng.. 2019;27(3):400–10. [PubMed: 30716040]
26. Phan H, Mikkelsen K, Chén OY, et al. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Trans Biomed Eng.. 2022;69(8):2456–67. [PubMed: 35100107]
27. Zan H, Yildiz A. Multi-task learning for arousal and sleep stage detection using fully convolutional networks. J Neural Eng.. 2023;20(5):056034.
28. 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(6):1187–203. [PubMed: 32299002]
29. Urtnasan E, Park JU, Lee JH, et al. Deep learning for automatic detection of periodic limb movement disorder based on electrocardiogram signals. Diagnostics.. 2022;12(9):2149. [PubMed: 36140550]
30. Ferri R, Koo BB, Picchietti DL, Fulda S. Periodic leg movements during sleep: phenotype, neurophysiology, and clinical significance. Sleep Med.. 2017;31:29–38. [PubMed: 28341521]
31. Thomas RJ. Scoring of Sleep-Related Breathing Events. In: Sleep Disorders Medicine: Basic Science, Technical Considerations and Clinical Aspects. 2017:431–47.
32. Thomas RJ. Arousals in sleep-disordered breathing: patterns and implications. Sleep.. 2003;26(8):1042–7. [PubMed: 14746388]
33. Nassi TE, Ganglberger W, Sun H, et al. Automated respiratory event detection using deep neural networks. arXiv preprint arXiv:2101.04635. 2021 Jan 12.
34. Nasiri S, Ganglberger W, Nassi T, et al. CAISR: Achieving human-level performance in automated sleep analysis across all clinical sleep metrics. Sleep.. Forthcoming 2025. Accepted for publication.
35. American Academy of Sleep Medicine. Inter-scorer Reliability Program (ISR) [Internet]. Available from: https://isr.aasm.org/. Accessed 2025 Jan 20.
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas.. 1960;20(1):37–46.

37. Conger AJ. Kappa reliabilities for continuous behaviors and events. Educ Psychol Meas. 1985;45(4):861–868.

38. Ganglberger W, Nasiri S, Sun H, et al. Refining sleep staging accuracy: transfer learning coupled with scorability models. Sleep.. 2024;47(11):zsae202. [PubMed: 39215679]

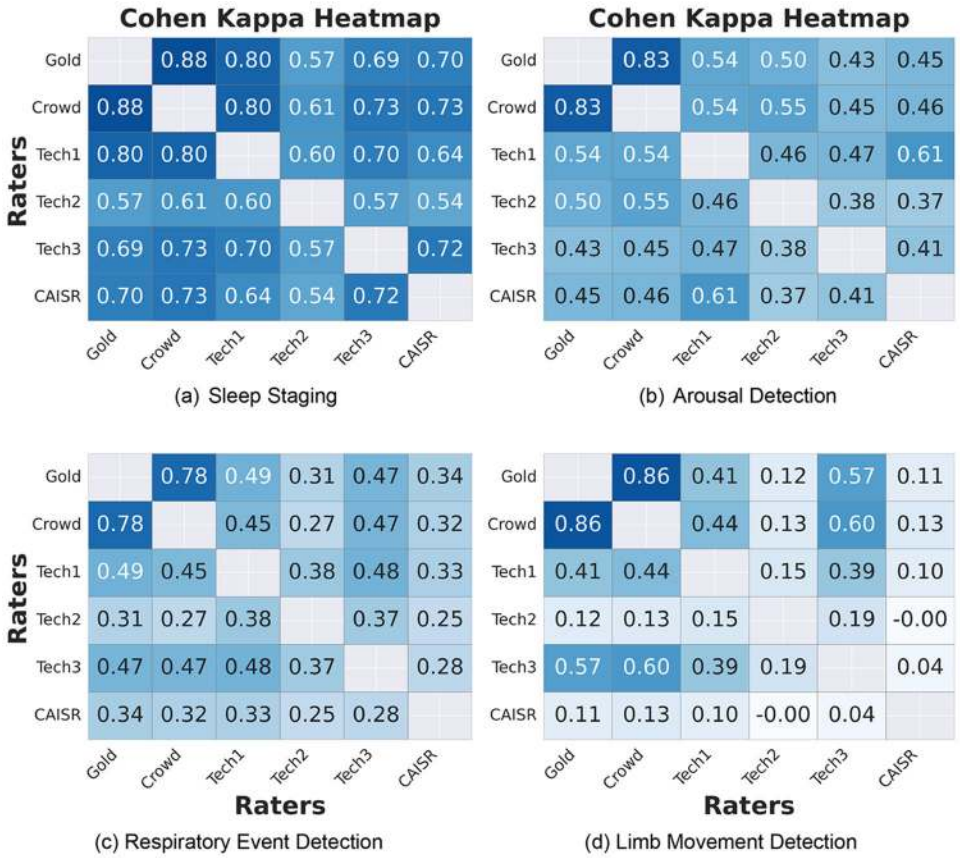

Figure 1:

Pairwise Cohen's Kappa heatmaps for CAISR and human scorers across the four PSG tasks: sleep staging, arousal detection, respiratory event detection, and limb movement detection. Note that the diagonals are "1.00" (perfect agreement) as these measure agreement of scores with themselves.

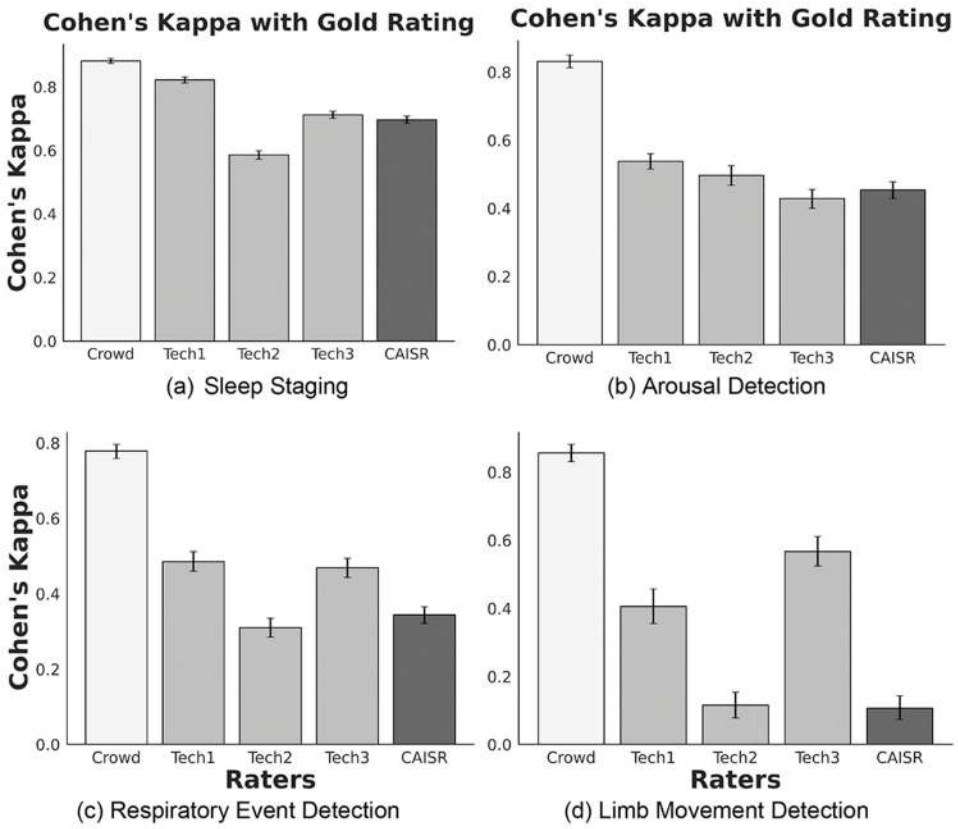

Figure 2:

Bar plots comparing CAISR, crowd ratings, and experience technician scores to the multi-expert gold standard across the four PSG tasks. Error bars represent 95% confidence intervals obtained via bootstrapping.

Table 1:

Demographics of the datasets used for training CAISR

	MGH	SHHS	MESA	MrOS
N subjects	14859	5793	2055	2898
Age (Mean, IQR)	52 (41, 65)	63 (55, 72)	68 (62, 76)	76 (72, 80)
N (%) Age 0-20	507 (3)	0 (0)	0 (0)	0 (0)
N (%) Age 20-40	2991 (20)	6 (0)	0 (0)	0 (0)
N (%) Age 40-60	6147 (41)	2291 (40)	331 (16)	0 (0)
N (%) Age 60-80	4775 (32)	3098 (54)	1399 (68)	2085 (72)
N (%) Age 80-100	438 (3)	311 (5)	274 (13)	683 (24)
N (%) Sex female	6396 (43)	3033 (52)	1102 (54)	0 (0)
Race/Ethnicity *				
N (%) Asian	486 (3)	N/A	250 (12)	85 (3)
N (%) Black	913 (6)	514 (9)	571 (28)	99 (3)
N (%) White	11390 (75)	4899 (85)	743 (36)	2638 (91)
N (%) Hispanic	N/A	N/A	491 (24)	N/A
N (%) Other	1845 (12)	380 (7)	0 (0)	76 (3)
Type cohort	Sleep laboratory, attended PSG	Community-dwelling, home-recording	Community-dwelling, home-recording	Community-dwelling, home-recording

* MESA collected race “Chinese”, not “Asian”. MESA collected “Hispanic” as part of race questionnaire. All KoGES participants lived in South Korea and identified as “northeast Asian”.

Table 2:

Demographics of the AASM ISR Dataset

Dataset	N subjects	Age (Mean, IQR)	N (%) Age 0-20	N (%) Age 20-40	N (%) Age 40-60	N (%) Age 60-80	N (%) Age 80-100	N (%) Sex female
AASM ISR	57	44.9 (37, 53)	0 (0.0%)	23 (40.4%)	27 (47.4%)	7 (12.3%)	0 (0.0%)	20 (35.1%)