SLEEP, 2026, **00**, 1–16
doi:10.1093/sleep/zsag064
Advance Access publication 10 March 2026
Original Article

Original Article

Pediatric SleepNet: a deep learning network for reliable pediatric sleep staging across developmental stages

Ayush Tripathi[1,2], Arnav Gupta[1,3], Wolfgang Ganglberger[1,2], Samuel Waters[1,2], Haoqi Sun[1,2], Samaneh Nasiri[1,2,4], Ayan Mitra[1,2], Katie L. Stone[5,6], Emmanuel Mignot[7], Dennis Hwang[8], Matthew A. Reyna[4], Lynn Marie Trotti[4,9], Gari D. Clifford[4,10], Kiran Maski[2,11,†], Umakanth Katwa[2,11,†], Robert J. Thomas[2,12,†] and M. Brandon Westover[1,2,12,†,*]

[1]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, United States, [2]Harvard Medical School, Boston, MA, United States, [3]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, United States, [4]Emory University School of Medicine, Atlanta, GA, United States, [5]Department of Epidemiology and Biostatistics, University of California, San Francisco, United States, [6]California Pacific Medical Center Research, San Francisco, CA, United States, [7]Stanford University, Palo Alto, CA, United States, [8]Kaiser Permanente, San Bernardino County Sleep Disorders Center, San Bernardino, CA, United States, [9]Department of Neurology and Emory Sleep Center, Emory University School of Medicine, Atlanta, GA, United States, [10]Georgia Institute of Technology, Atlanta, GA, United States, [11]Sleep Center, Boston Children's Hospital, Boston, MA, United States and [12]Department of Medicine, Division of Pulmonary Critical Care & Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

[*]Corresponding author. M. Brandon Westover, Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA.
Email: bwestove@bidmc.harvard.edu.
[†]Co-senior authors.

Abstract

Study Objectives: Manual sleep staging in pediatric populations is challenging due to developmental variability and limited scoring consistency, especially in infants and toddlers. We developed a multimodal deep learning model for pediatric sleep staging and evaluated its performance across a broad age range and diverse clinical subgroups.

Methods: We trained a U-Net-inspired encoder-decoder model (pediatric SleepNet) using 9-channel input signals: electroencephalography (EEG), electrooculography (EOG), and chin electromyography (EMG) using 35-epoch segments from clinical pediatric polysomnograms (PSGs). Models were trained separately on three age groups (<6 months, 6-12 months, >1 year) using 9150 PSGs, with 2455 PSGs reserved for validation. Evaluation was conducted on 3804 held-out test recordings. Performance was compared with U-Sleep and the Complete Artificial Intelligence Sleep Report (CAISR), and stratified analyses were performed across ages, sexes, and seven ICD-10-based disease categories. External validation was conducted on two independent datasets, CHAT and PATS.

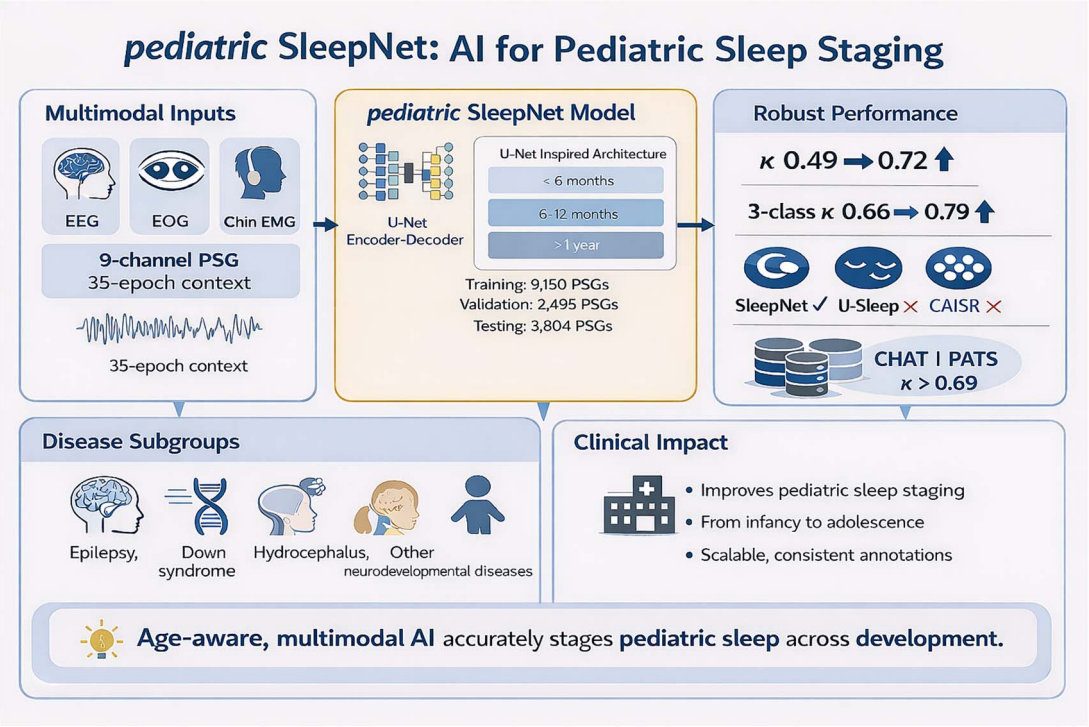
Results: pediatric SleepNet achieved robust performance across all age groups, with mean Cohen's Kappa increasing from 0.49 (0–6 months) to 0.72 (>12 years). It significantly outperformed U-Sleep and CAISR across early developmental stages. Three-class staging yielded mean Cohen's Kappa increasing from 0.66 (0–6 months) to 0.79 (>12 years). Sex-based differences were negligible. However, significant reductions in performance were observed in children with epilepsy, Down syndrome, hydrocephalus, and other neurodevelopmental conditions. External validation yielded Kappa values >0.69 comparable to the internal test set.

Conclusions: pediatric SleepNet demonstrates reliable sleep staging across pediatric development. Its robust performance across age, disease, and external datasets supports its potential for clinical and research use in pediatric sleep medicine.

Key words: pediatric sleep; sleep staging; deep learning; polysomnography; automated scoring

Submitted: 13 December, 2025; **Revised:** 18 February, 2026; **Accepted:** 23 February, 2026

Graphical Abstract



Statement of Significance

Automated sleep staging in pediatric populations remains a critical unmet need due to the complexity of developmental electroencephalography (EEG) patterns and limited generalizability of adult-trained models. We present pediatric SleepNet, a deep learning model that provides reliable sleep stage classification across the pediatric age span, from infants to adolescents. Our model significantly outperforms existing approaches in early childhood and maintains high agreement with expert scorers throughout development. By leveraging age-specific training and multimodal input (EEG, EOG, chin electromyography), it captures physiologically meaningful patterns and improves consistency in pediatric sleep scoring. These findings highlight the potential of pediatric SleepNet to enhance pediatric sleep research and support clinical decision-making by providing consistent, scalable, and accurate sleep stage annotations from infancy through adolescence.

Introduction

Sleep is a fundamental physiological process that underpins health and well-being by facilitating crucial restorative functions such as memory consolidation, immune regulation, hormonal balance, emotional stability, and cognitive performance [1–3]. Beyond its restorative role, sleep serves as a sensitive indicator of physiological and neurological conditions. Alterations in sleep architecture and quality are frequently associated with a wide array of medical conditions, ranging from primary sleep disorders (e.g. insomnia, sleep apnea, narcolepsy) to complex neurological, psychiatric, and developmental disorders [4, 5]. Untreated sleep disturbances are known to exacerbate cognitive decline, metabolic syndrome, cardiovascular disease, and mental health conditions, emphasizing the critical role of accurate and timely sleep assessment in clinical care. Moreover, sleep disorders often present alongside chronic illnesses, compounding their burden and significantly impairing patients' quality of life [6, 7].

Sleep staging forms the cornerstone of pediatric polysomnogram (PSG)-based sleep evaluation. The American Academy of Sleep Medicine (AASM) divides sleep into Wake, REM (rapid eye movement), and three non-REM (N1, N2, N3) stages, each 30-s characterized by distinct electroencephalography (EEG) features such as sleep spindles, K-complexes, slow-wave activity, and muscle atonia [8]. However, manual scoring is labor-intensive, time-consuming, and subject to substantial inter-scorer variability, especially in pediatric populations where EEG signatures are developmentally dynamic and often ambiguous [9, 10].

These challenges have driven the development of automated sleep staging systems using machine learning and deep learning methods [11–13]. State-of-the-art deep neural networks, including Convolutional Neural Networks, Recurrent Neural Networks, and transformer-based models, have demonstrated high agreement with human scorers in adult datasets (F1 = 0.79 and kappa = 0.7–0.8) [11, 12]. However, pediatric sleep staging presents additional challenges. Infants and children exhibit distinct neurophysiological characteristics and sleep patterns that evolve with age. EEG waveforms, event morphology, spectral content, and stage transitions vary dramatically from neonates to adolescents, leading to reduced performance of models trained on adult data when applied to pediatric cohorts. Despite the growing use of AI in sleep medicine, only a handful of studies have specifically addressed the pediatric domain [14–17], and most existing models fail to

generalize across developmental stages due to inadequate age diversity and sample size. Furthermore, inter-scorer agreement is especially low in infants and toddlers, where stage boundaries are less clearly defined, and sleep architecture is immature. This variability further complicates model training and evaluation. To develop clinically useful pediatric sleep staging models, large-scale datasets with high-quality annotations across a broad age range are essential. However, publicly available pediatric PSG datasets are relatively scarce, typically small, and often restricted to narrow age bands or specific disease populations [18–25]. These limitations hinder robust model development and benchmarking.

Previous work has shown that automatic sleep scoring is feasible and effective in adult populations. In this article, we first evaluated the performance of two models trained for the adult population on pediatric patients and then demonstrate that a model trained specifically on pediatric data achieves substantially better performance. We introduce pediatric SleepNet, a deep learning-based model for pediatric sleep staging, trained and evaluated on the Boston Children's Hospital (BCH) Sleep Corpus, the largest annotated pediatric sleep dataset to date, comprising 15 695 PSGs recorded at BCH. The dataset spans a wide developmental range from birth to 18+ years and includes detailed age, sex, and clinical diagnosis metadata. The pediatric SleepNet model leverages a U-Net-inspired encoder–decoder architecture, trained on multimodal PSG signals (EEG, electrooculography [EOG], electromyography [EMG]) and optimized across age-specific subgroups. We comprehensively evaluated model performance using Cohen's Kappa and confusion matrices across nine age bins, investigated sex-related differences, and assessed generalizability to external pediatric datasets. We also explored a simplified three-class staging scheme and analyze the impact of clinical diagnoses on staging reliability.

Materials and Methods

Ethical approval

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

Dataset

This study utilized the BCH Sleep Corpus, a large-scale pediatric PSG dataset comprising overnight recordings from patients aged from birth to over 18 years. Each PSG includes comprehensive multi-channel physiological recordings acquired in a clinical setting and scored by trained sleep technicians according to AASM guidelines. The dataset is described in detail in [26] and includes demographic metadata (age, sex), ICD-10 disease codes, and manually annotated 30-second sleep stage labels for WAKE, N1, N2, N3, and REM stages. The 0–6 months old subjects includes infants that are 0–2 months old. In the BCH corpus, annotations are distributed as 30-s epoch labels in the 5-class set {W, N1, N2, N3, REM} across all ages, we therefore used those labels directly and did not apply any post-hoc mapping for infants.

Recordings with missing annotations or significant signal corruption were excluded. From the remaining PSGs, we created fixed training, validation, and test splits, ensuring no subject overlap across subsets. The final dataset consisted of 9150 PSGs from 7604 subjects for training, 2455 from 2000 subjects for validation, and 3804 from 3000 subjects for testing. The test set was carefully curated to ensure broad representation across developmental stages and clinical subgroups. Table 1 summarizes the distribution of PSGs by age (postnatal) and sex (when available) across the three splits. Age was treated both as a continuous variable and also stratified into nine developmental bins: 0–6 months, 6–12 months, 1–2 years, 2–3 years, 3–4 years, 4–5 years, 5–6 years, 6–12 years, and >12 years. The final bin ">12 years" includes all subjects aged ≥12 years and therefore includes a small subset of recordings from individuals >18 years present in the BCH corpus.

Signal preprocessing

For each PSG recording, we extracted a standardized set of nine physiological signals: six EEG channels (C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, O2-M1), two EOG channels (E1-M2, E2-M1), and one chin EMG channel (chin1-chin2). PSG signals were accessed via BDSP in a standardized EDF format and loaded using MNE-Python for signal extraction and preprocessing. Preprocessing was performed uniformly across training, validation, and test datasets to ensure consistency. Signals were processed using the MNE-Python library version 1.10.2 [27].

A notch filter was applied at 60 Hz to suppress power line interference. Subsequently, EEG and EOG signals were bandpass filtered from 0.3 to 35 Hz, while the EMG channel was filtered from 10 to 100 Hz to isolate muscle activity. After filtering, all channels were resampled to 128 Hz to reduce computational load and standardize input dimensions. To emphasize transient EMG bursts, the analytic envelope of the EMG signal was computed using the Hilbert transform.

All signals were then normalized on a per-channel basis using a robust scaling approach. Specifically, each channel was median-centered and scaled by its interquartile range, with the resulting values clipped to ±20 to suppress extreme outliers. This normalization strategy was applied independently to each PSG recording to reduce inter-subject variability while preserving physiologically relevant dynamics.

Each recording was segmented into non-overlapping 30-s epochs (equivalent to 3840 samples at 128 Hz). Corresponding sleep stage labels were obtained from the original technician annotation. Epochs with invalid or missing annotations were excluded from training and evaluation. To enable contextual learning, we constructed sequences of 35 consecutive epochs, resulting in fixed-length input segments of shape (134 400 × 9). We fixed the sequence length to 35 epochs (17.5 min) to match prior U-Sleep work that provides sufficient temporal context while remaining computationally tractable. For edge segments with fewer than 35 valid epochs, the final valid epoch was repeated as needed to ensure full segment coverage. Padding occurs only at the end of a recording and padded epochs are excluded from reported evaluation metrics. This formulation enables the model to capture both local signal characteristics and longer temporal transitions across sleep stages.

Model training

We developed a deep learning framework for pediatric sleep staging based on a modified U-Net architecture [28], following the U-Sleep fully convolutional sequence-to-sequence sleep staging model, we employ an encoder–decoder structure with skip connections. The model takes as input a fixed-length segment of 35 epochs (30 s per epoch; 17.5 minutes of context) and outputs a corresponding 35-epoch sequence of sleep stage probabilities. We fixed 35 to match the established setup in prior U-Sleep work, where window length is treated as a design parameter that provides sufficient temporal context for capturing stage transitions. Each input segment has shape (134 400 × 9), where 134 400 represents 35 epochs × 3840 samples (at 128 Hz), and 9

Table 1. Distribution of PSGs in training, validation, and test sets by age group and sex

Age group	N train (%female)	N validation (%female)	N test (%female)
0–6 months	613 (39.3)	194 (44.3)	381 (43.0)
6–12 months	280 (36.1)	76 (43.4)	284 (43.0)
1–2 years	691 (36.9)	186 (38.7)	261 (39.5)
2–3 years	641 (40.2)	165 (32.7)	230 (41.3)
3–4 years	717 (43.9)	187 (35.8)	240 (38.3)
4–5 years	679 (40.8)	188 (43.6)	217 (40.1)
5–6 years	579 (42.5)	163 (44.8)	262 (42.4)
6–12 years	2602 (41.2)	685 (42.0)	937 (43.4)
>12 years	2348 (45.8)	611 (40.8)	991 (46.1)
Total	9150 (42.0)	2455 (40.9)	3804 (43.1)

corresponds to the number of input channels (six EEG, two EOG, and one EMG).

The model comprises 12 encoder blocks with progressively increasing filter sizes, followed by a bottleneck and 12 decoder blocks. Each encoder block consists of a one-dimensional convolutional layer (kernel size = 9), ELU activation, batch normalization, optional dropout, and max pooling. Decoder blocks use upsampling followed by convolution, ELU activation, and skip connections from the corresponding encoder layers. The final output is a 35×5 sequence of softmax probabilities, representing five sleep stages: WAKE, N1, N2, N3, and REM. The full model contains approximately 4.8 million trainable parameters.

To account for developmental variability, we adopted a two-stage age-stratified training protocol. First, the model was trained from scratch using recordings from subjects older than one year (>365 days). This base model was then fine-tuned separately for two younger age bins: 6–12 months and <6 months. Fine-tuning was performed by initializing the weights from the >1 year model and continuing training on the age-specific subsets. We adopted this approach to account for sleep EEG morphology and stage characteristics that change substantially across development, and a single unified model can suffer negative transfer when fitting different age-dependent distributions.

The model was implemented in TensorFlow and trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 64. The loss function was categorical cross-entropy, with class weights computed from the empirical label distribution for each age group. Early stopping was applied based on validation accuracy with a patience of 50 epochs. Final models were saved after convergence and used for inference without any additional post-processing.

Internal validation and evaluation framework

Internal validation was performed on the held-out test set consisting of 3804 PSG recordings. Each recording underwent the same preprocessing steps described previously, including filtering, resampling, normalization, and segmentation into non-overlapping 30-s epochs. These epochs were then grouped into fixed-length sequences of 35 consecutive epochs, resulting in input segments of shape $134\,400 \times 9$. In cases where a segment contained fewer than 35 valid epochs, the final valid epoch was repeated as many times as necessary to complete the segment. These segments were passed through the appropriate trained model, which returned a sequence of softmax probabilities representing the likelihood of each of the five sleep stages for each epoch. The final predicted label for each epoch was determined by selecting the class with the highest predicted probability. Padded epochs were excluded from final evaluation to ensure that only original data contributed to performance metrics.

Each recording in the test set was evaluated using the model appropriate for the subject's age group. For subjects older than one year (>365 days), we used the base model trained on the >1 year training set. For infants aged between 6 and 12 months, we used the model fine-tuned on that specific age bin. For subjects younger than 6 months, the model fine-tuned on the <6 month subset was used. This age-specific inference strategy ensured that each test sample benefited from training data matched in the developmental stage.

Model performance was assessed using multiple quantitative metrics. The primary metric used was Cohen's Kappa [29], which accounts for chance agreement and is more appropriate than raw accuracy for multi-class classification tasks with imbalanced class distributions. Overall accuracy was also reported for completeness. To further assess per-class performance, we computed confusion matrices between predicted and ground truth labels. Evaluation was performed using the original 5-class sleep stage formulation, including WAKE, N1, N2, N3, and REM. In addition, we investigated a simplified 3-class formulation by merging N1, N2, and N3 into a single "NREM" class, retaining WAKE and REM as separate categories. This formulation was motivated by the observation that scoring agreement for different non-REM stages is particularly low in pediatric recordings, especially in infants and toddlers, and that a coarse-grained staging system may be more appropriate for such populations.

Age-wise and sex-wise performance characterization

To characterize performance as a function of developmental age, we conducted two complementary forms of age-wise analysis. In the first approach, recordings were stratified into nine predefined age bins: 0–6 months, 6–12 months, 1–2 years, 2–3 years, 3–4 years, 4–5 years, 5–6 years, 6–12 years, and >12 years. Within each bin, we computed the mean and standard deviation of subject-level Cohen's Kappa values. In the second approach, we treated age as a continuous variable and plotted individual subject Kappa scores against age in years. The age values, originally recorded in days, were converted to years and divided into 0.5-year intervals. For each interval, we computed the median Kappa score and the associated 95% confidence interval estimated using bootstrapping. A univariate smoothing spline was then fit to the sequence of median values and the confidence intervals. The resulting curve provided a continuous estimate of performance across development and was plotted along with the raw Kappa values and confidence interval bands. This analysis revealed a steady increase in model reliability with age and identified regions of

higher variability in the early years of life, consistent with known challenges in sleep scoring in infants and toddlers.

We also examined the potential influence of biological sex on model performance. Each test subject was labeled as male or female according to demographic records, and Cohen's Kappa scores were compared across sexes within each age bin. Although absolute differences between male and female performance were generally small, we assessed statistical significance for each age bin using Welch's *t*-test. Results were reported with 95% confidence intervals, and differences with $p < .05$ were noted.

To evaluate the robustness of the pediatric SleepNet model under a simplified scoring framework, we performed 3-class sleep staging by merging N1, N2, and N3 into a single NREM category, alongside WAKE and REM. This approach is often used in pediatric settings to reduce ambiguity in stage boundaries and improve inter-scorer agreement.

We compared the performance of our model (pediatric SleepNet) to two state-of-the-art sleep staging models: U-Sleep [12] and Complete Artificial Intelligence Sleep Report (CAISR) [11]. These models, originally trained on large-scale datasets dominated by adult PSGs, represent current benchmarks in automated sleep staging. We evaluated U-Sleep and CAISR using the same set of input channels provided to Pediatric SleepNet (six EEG, two EOG, one EMG). Both U-Sleep and CAISR were run using their own published preprocessing and built-in handling of channel mapping/mismatch where applicable. However, the evaluation pipeline used was the same as SleepNet. Cohen's Kappa was computed for each subject and averaged within each age bin along wide confidence intervals for comparison of the three models. Moreover, To evaluate the generalizability of our pediatric SleepNet model beyond the BCH dataset, we tested its performance on two external pediatric cohorts: CHAT and PATS. These datasets, distinct from the training distribution, contain a wide age range of pediatric PSG recordings scored by independent annotators, providing a robust benchmark for assessing cross-cohort reliability.

Finally, to evaluate the impact of clinical conditions on model performance, we leveraged diagnostic information available in the BCH dataset to stratify recordings based on the presence or absence of specific ICD9/10 categories. Seven diagnostic groups were selected based on clinical relevance and sample size: epilepsy, Down's syndrome, pervasive developmental disorders (including autism spectrum conditions), hyperkinetic disorders (e.g. Attention Deficit Hyperactivity Disorder), hydrocephalus, asthma, and other congenital malformations. For each diagnostic group, Kappa scores were compared between subjects with and without the condition, both across all ages and within the nine developmental age bins. These comparisons were visualized using grouped bar plots, and statistical significance was assessed using Welch's *t*-test. Comparisons with $p < .05$ were indicated with asterisks on the bar plots.

External validation

To assess the generalizability of the proposed sleep staging model beyond the BCH dataset, we conducted external validation on two independent pediatric cohorts: the Childhood Adenotonsillectomy Trial (CHAT) [21, 30] and the Pediatric Adenotonsillectomy Trial for Snoring (PATS) [24, 30, 31]. These datasets differ from BCH in terms of participant demographics, clinical conditions, and recording protocols, and thus offer a meaningful evaluation of model robustness across diverse pediatric populations.

The CHAT dataset includes overnight PSG recordings from children aged 5-9.9 years who were enrolled in a multi-site clinical trial evaluating the impact of adenotonsillectomy on sleep, cognition, and behavior. The PATS dataset comprises recordings from children aged 3–12 years with snoring and mild sleep-disordered breathing. Inclusion criteria for PATS included snoring on at least three nights per week, an apnea–hypopnea index (AHI) ≤ 3/h, no oxygen desaturations below 90%, and evidence of tonsillar hypertrophy. All recordings in both datasets were scored according to AASM guidelines and included the standard montage of EEG, EOG, and chin EMG channels required for staging.

The same preprocessing pipeline used for the BCH data (including notch filtering, bandpass filtering, resampling, normalization, segmentation into 30-s epochs, and grouping into 35-epoch segments) was applied to 1629 PSGs from 448 unique subjects from CHAT and 717 PSGs from 492 unique subjects from PATS. Inference was performed using the appropriate age-specific version of the pediatric SleepNet model, selected based on the subject's age at the time of recording. Model performance on the external datasets was accessed using confusion matrices and Cohen's Kappa, following the same evaluation framework used for the BCH test set.

Results

Sleep staging performance on the BCH test set

We first evaluated the performance of the proposed pediatric SleepNet model on the held-out BCH test set, which included 3804 pediatric PSG recordings spanning a broad developmental age range. Each recording was assigned to the appropriate age-specific model—<6 months, 6–12 months, or >1 year—based on subject age at the time of the study. Epoch-level predictions were generated and compared to technician-provided reference labels, and performance was quantified using mean Cohen's Kappa. Across the entire test set, the model achieved an overall mean Cohen's Kappa of 0.68, indicating substantial agreement between model predictions and expert annotations.

Table 2 summarizes mean Cohen's Kappa values across nine predefined age bins: 0–6 months, 6–12 months, 1–2 years, 2–3 years, 3–4 years, 4–5 years, 5–6 years, 6–12 years, and >12 years. A clear developmental trajectory was observed, with performance increasing progressively across age bins. The lowest agreement was observed in infants under 6 months of age, where the mean Cohen's Kappa value was 0.49 indicating challenges in EEG interpretation and immature sleep patterns are well known to reduce sleep staging reliability. Performance improved sharply through toddlerhood and early childhood, stabilizing in late childhood and adolescence, where mean Kappa values reached 0.72.

To explore this trend in more detail, we treated age as a continuous variable and plotted individual subject-level Kappa scores against age in years as presented in Figure 1. Median values were computed within 0.5-year bins, and a smoothing spline was fit to estimate the performance trajectory across development. The resulting curve confirmed the steady increase in staging reliability with age, particularly during the first 5 years of life. The spread of Cohen's Kappa as depicted in Figure 1, was widest during infancy, reflecting greater variability in model performance during this period.

The confusion matrices for each of the nine age bins are shown in Figure 2. These matrices provide a detailed view of model performance across developmental stages. The most prominent source of error throughout all age groups was confusion between N1 and N2, consistent with known scoring challenges in early non-REM stages. This confusion was most pronounced in infants and gradually decreased with age. The model consistently

Table 2. Mean Cohen's Kappa for 5-class sleep staging across different age bins (0–6 months to >12 years)

Age group	0–6 months	6–12 months	1–2 years	2–3 years	3–4 years	4–5 years	5–6 years	6–12 years	>12 years
Cohen's κ	0.48	0.63	0.65	0.64	0.69	0.69	0.68	0.72	0.72

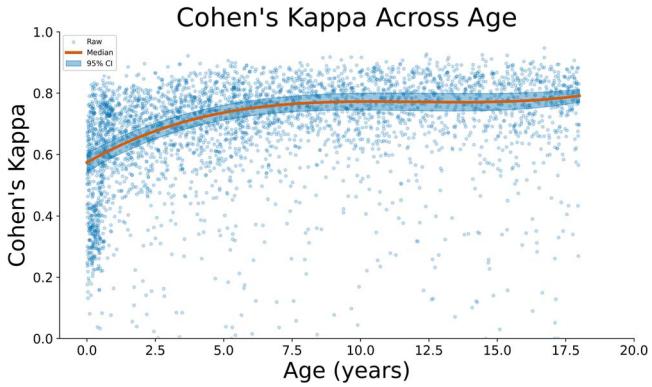


Figure 1. Scatter plot of individual subject Cohen's Kappa versus age in years, overlaid with a smoothing spline and 95% confidence interval computed using bootstrapping.

achieved high true positive rates for REM and WAKE stages, with minimal misclassification across all age bins. Agreement for all the sleep stages improved steadily with age, reaching over 60% for N1 and over 80% for all other sleep stages in children older than twelve years.

Comparison with existing sleep staging models

We compared the performance of our proposed pediatric SleepNet model against two existing sleep staging models: U-Sleep and CAISR. Both these models were developed and validated primarily on adult or mixed-age populations, and their performance on pediatric cohorts has not been extensively characterized. To ensure a fair comparison, all models were evaluated on the same held-out BCH test set. Mean Cohen's Kappa computed separately for each age bin and used as the primary performance metric.

As shown in Figure 3, pediatric SleepNet consistently outperformed both U-Sleep and CAISR across all age groups. The largest performance gains were observed in younger children. For instance, among infants aged 0-6 months, pediatric SleepNet achieved a mean Cohen's Kappa of 0.49, substantially higher than 0.30 for U-Sleep and 0.34 for CAISR. Similarly, in the 6-12 month group, pediatric SleepNet reached a mean Kappa of 0.64, while U-Sleep and CAISR achieved only 0.41 and 0.37, respectively. These differences were statistically significant ($p < 10^{-50}$) and likely reflect the pediatric-specific training and fine-tuning of the proposed model.

The performance advantage of pediatric SleepNet was sustained throughout childhood. For example, at age 2–3 years, the mean Kappa was 0.65 for pediatric SleepNet, compared to 0.52 for U-Sleep and 0.48 for CAISR. By ages 10–11 years, when sleep architecture became more stable, pediatric SleepNet still maintained an edge, achieving a mean Kappa of 0.74, versus 0.71 for U-Sleep and 0.68 for CAISR. However, at later stages of development-including adolescence and early adulthood-the performance gap between models narrowed considerably. Additionally, for a small subset of subjects older than 18 years, pediatric SleepNet achieved a mean Kappa of 0.68, with U-Sleep and CAISR reaching 0.66 and 0.65, respectively.

Across the full BCH test set, the overall mean Cohen's Kappa was 0.69 for pediatric SleepNet, compared to 0.62 for U-Sleep and 0.58 for CAISR. These results demonstrate the importance of pediatric-specific modeling and highlight the superior accuracy and reliability of pediatric SleepNet for automated sleep staging in children, particularly in early developmental stages.

Sex-based performance differences

We examined whether sleep staging performance varied by biological sex across the pediatric test population. Figure 4 shows mean Cohen's Kappa scores for male and female subjects within each of the nine developmental age bins, as well as for the entire test set. Across all age groups, performance was largely comparable between sexes, with no statistically significant differences observed (all p-values $>.05$).

In infancy (0–6 months), the model achieved mean Cohen's Kappa values of 0.48 for males ($n = 216$) and 0.49 for females ($n = 164$), with a p-value of .67. This trend of similar performance continued consistently across development. For example, in the 6–12 years group, male and female subjects achieved Kappa values of 0.73 ($n = 530$) and 0.72 ($n = 407$), respectively ($p = .18$). In the oldest bin (>12 years), performance remained equivalent with mean Kappa values of 0.73 for males and 0.72 for females ($p = .62$).

On the entire test set, the mean Kappa was 0.678 for males ($n = 2164$) and 0.680 for females ($n = 1637$), with no statistically significant difference ($p = .65$). These findings indicate that the pediatric SleepNet model provides equitable performance across sexes, with staging accuracy driven predominantly by developmental stage rather than sex-related differences.

Impact of clinical diagnoses on sleep staging performance

To assess the influence of neurological and developmental conditions on model performance, we compared sleep staging performance in children with specific ICD-based clinical diagnoses against age-matched children without those diagnoses. Figure 5 presents the mean Cohen's Kappa values across age groups for seven commonly observed diagnostic categories: epilepsy (G40), other congenital malformations of the nervous system (Q07), hydrocephalus (G91), asthma (J45), Down's syndrome (Q90), pervasive developmental disorders (F84), and hyperkinetic disorders (F90). The figure includes 95% confidence intervals and statistical significance ($p < .05$) indicators from unpaired t-tests.

Epilepsy had the most prominent and consistent impact on model performance across all age groups. Kappa values were lower in children with epilepsy compared to their non-epileptic peers across all bins, with statistically significant differences observed in eight out of nine age groups. The performance gap existed throughout the developmental ages: for instance, in the 0–6 month group, mean Kappa was 0.44 in children with epilepsy versus 0.49 in others ($p = .085$), and in the 6–12 month group, the gap widened to 0.56 vs. 0.65 ($p = .0011$). Even in older children (>12 years), the difference remained statistically significant (Kappa: 0.64 vs. 0.73, $p = .0003$), underscoring the challenges posed by epilepsy-associated EEG abnormalities.

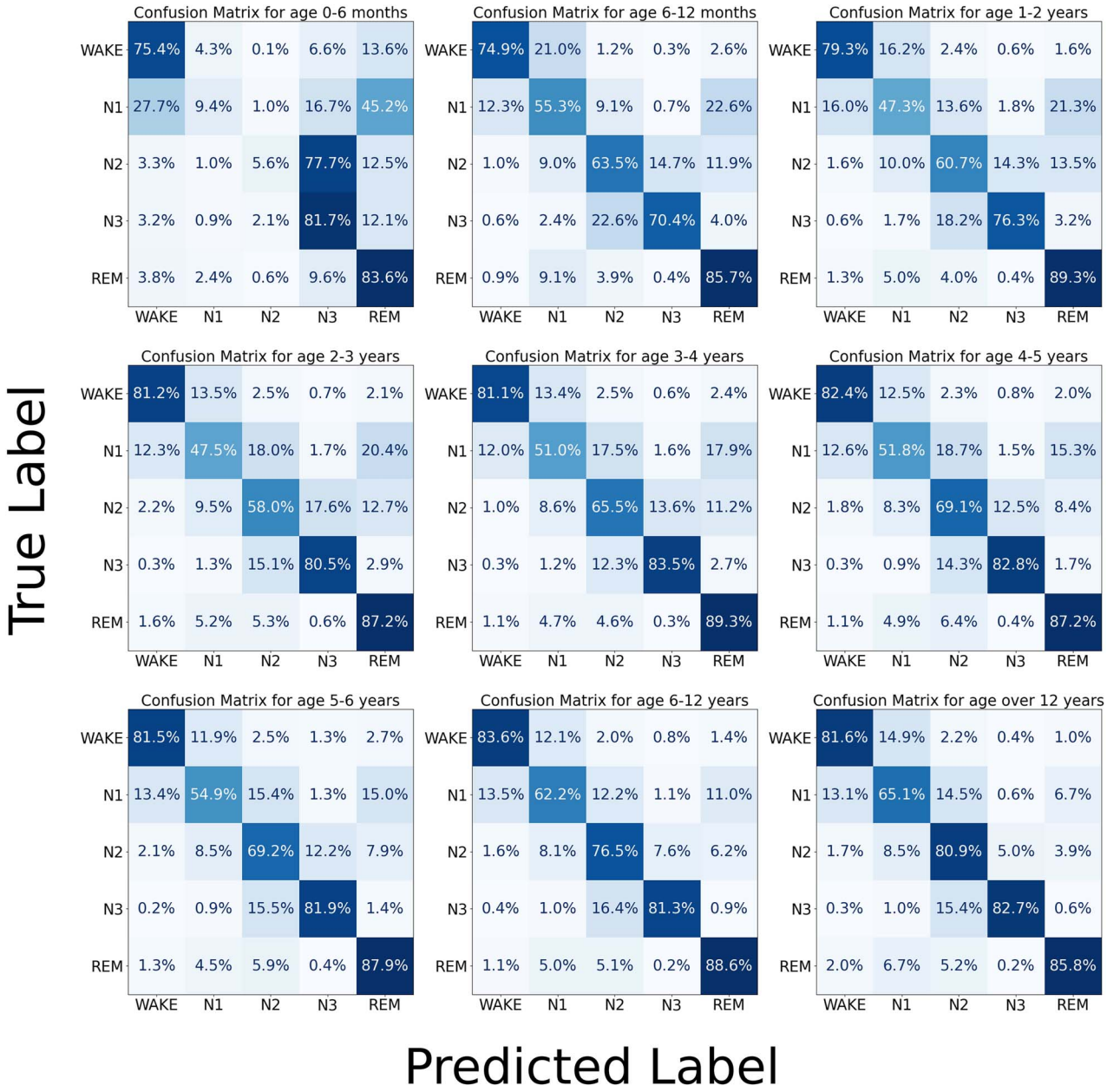


Figure 2. Confusion matrices for 5-class sleep staging (WAKE, N1, N2, N3, REM) across nine pediatric age groups. Each matrix shows the distribution of model predictions relative to expert annotations.

Other conditions, such as hydrocephalus, also showed significantly reduced staging performance, particularly in mid-childhood (e.g. Kappa: 0.62 vs. 0.69 at age 1–2 years, $p = .27$; 0.44 vs. 0.70 at age 3–4 years, $p = .040$). Similar trends were observed for Down's syndrome, pervasive developmental disorders, and hyperkinetic disorders, where staging reliability was generally lower in the presence of the diagnosis, especially in younger age groups. For example, in children aged 0–6 months with hyperkinetic disorders, mean Kappa dropped to 0.36 compared to 0.49 in non-diagnosed children ($p = .0059$). In pervasive developmental disorders, differences were significant at multiple ages including 0–6 months (Kappa: 0.41 vs. 0.49, $p = .036$) and 6–12 months (0.67 vs. 0.63, $p = .044$). In contrast, conditions such as asthma and congenital malformations of the nervous system had more modest or inconsistent effects. In most bins, differences were not statistically significant, though asthma showed a lower mean Kappa in the 0-6 month bin (0.43 vs. 0.50, $p = .0087$).

Visualization of model predictions

Figure 6 presents examples of model predictions, displaying averaged frontal, central, and occipital EEG spectrograms alongside hypnograms from the reference technician scoring (Tech), pediatric SleepNet, U-Sleep, and CAISR. The corresponding Cohen's Kappa values quantify agreement between each model and the expert scoring. Averaging spectrograms within frontal, central, and occipital derivations highlights regional EEG dynamics while minimizing channel-specific variability, providing a clearer representation of the dominant spectral patterns associated with different sleep states.

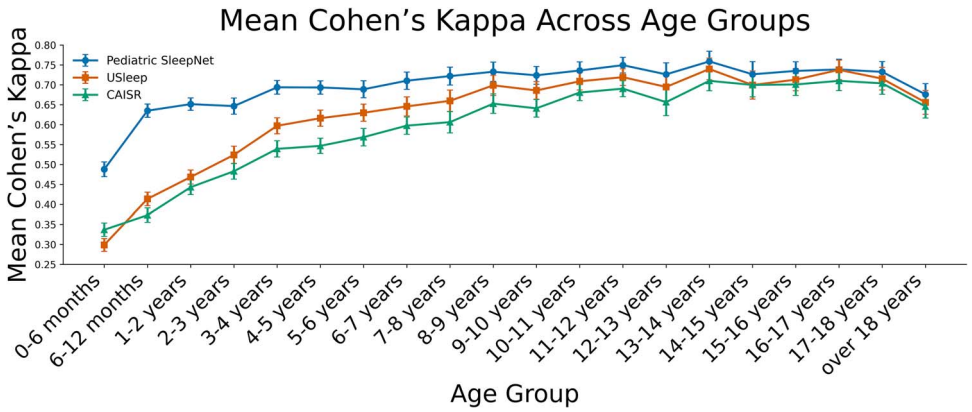


Figure 3. Comparison of mean Cohen's Kappa across pediatric age groups for three sleep staging models: pediatric SleepNet (blue), U-Sleep (orange), and CAISR (green). Error bars denote 95% confidence intervals. The proposed pediatric SleepNet model consistently outperforms both models across all age groups, with the largest performance gains in early childhood. Model performance converges in adolescence and early adulthood.

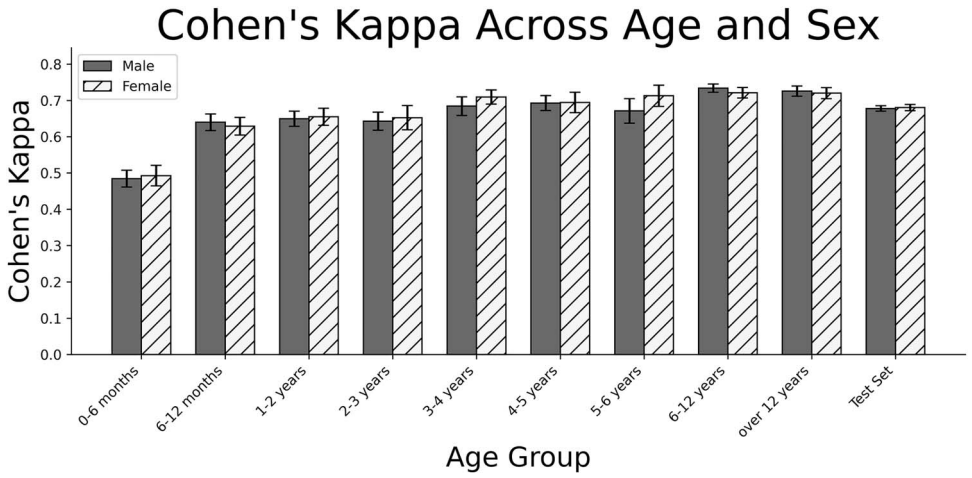


Figure 4. Mean Cohen's Kappa across age groups stratified by biological sex. Bars show the performance of the pediatric SleepNet model separately for male and female subjects within each age group. Error bars indicate 95% confidence intervals. Overall, staging reliability was comparable between sexes across all age bins, with no statistically significant differences observed ($p > .05$ for all comparisons).

In younger infants, presented in Figure 6A and B, EEG features are in the developing phase. U-Sleep and CAISR tend to exhibit greater variability in classification during transitional periods, often confusing REM and non-REM segments, which is consistent with their training on predominantly older populations. In Figure 6C, sample spectrograms and predictions from a 1.83-year-old child are presented. The plots reveal a gradual transition in spectral morphology, with spindles and slow waves becoming more prominent and REM periods more clearly differentiated. In this case, the pediatric SleepNet model demonstrates superior stability in sustained NREM and REM segments compared to both USleep and CAISR. Subsequently, in Figure 6D, we present sample plots from a PSG of a child aged 3.41 years. In this case, the EEG demonstrates well-defined NREM–REM cycling, with robust spindles, slow-wave activity, and distinct REM-associated low-amplitude mixed-frequency patterns. The pediatric SleepNet model shows near-perfect alignment with the technician, while U-Sleep and CAISR also achieve competitive performance but occasionally misclassify isolated epochs at stage transitions.

Performance under 3-class sleep staging

The mean Cohen's Kappa across different age bins is presented in Table 3. For 3-class staging, mean Cohen's Kappa was 0.66 for infants aged 0–6 months, improving to 0.73 for the 6–12 month group, and further increasing to 0.74 in the 1–2 year group. Between 2 and 6 years of age, performance stabilized with Kappa values ranging from 0.74 to 0.78. The highest agreement was observed in the 6–12 year age bin, where the model achieved a Kappa of 0.81. In adolescents (>12 years), the agreement remained high at 0.78.

The confusion matrices in Figure 7 provide a breakdown of prediction accuracy across the three classes. NREM sleep was reliably classified across all ages, with accuracy increasing from 78.4% in the 0–6 month group to 94.3% in the >12 year group. REM classification also remained robust, with correct prediction rates exceeding 83% in all the age groups. WAKE was the most challenging stage to classify accurately in infants, with a precision of 75.4% in the youngest group and frequent misclassification as NREM. However, accuracy improved with age, reaching 83.6% in children aged 6–12 years.

Generalization to external pediatric cohorts

On the CHAT dataset, pediatric SleepNet achieved a mean Cohen's Kappa of 0.76, indicating strong agreement with expert annotations. The confusion matrix for CHAT presented in Figure 8A shows that the model performed consistently well across all sleep stages, with particularly high agreement for N3 (93.6%) and REM (93.6%) stages. Moderate confusion was observed between

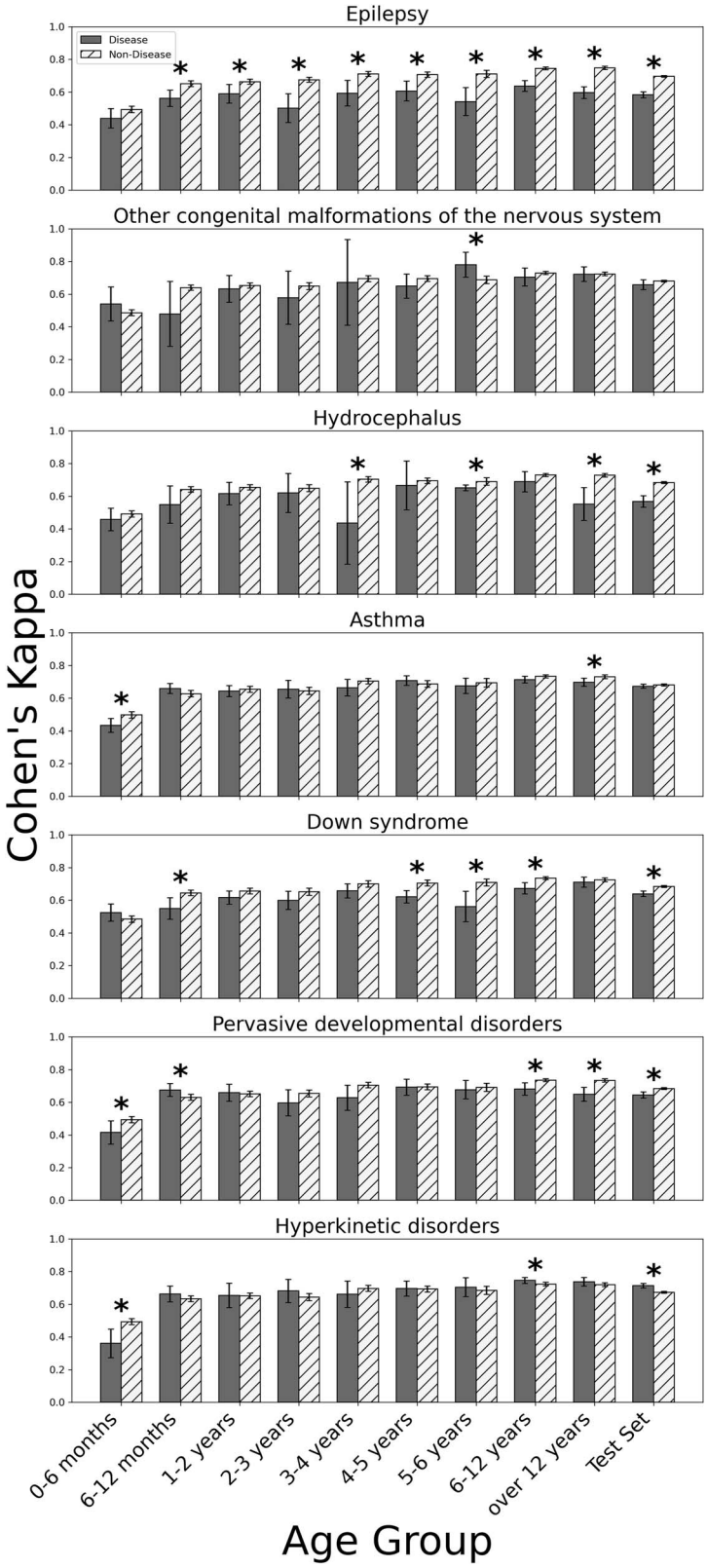


Figure 5. Impact of clinical diagnoses on sleep staging performance across pediatric development. Mean Cohen's Kappa values are shown for children with (solid bars) and without (hatched bars) seven common clinical conditions, across different developmental age bins. Error bars indicate 95% confidence intervals. Asterisks (∗) denote statistically significant differences ($p < .05$) between diseased and non-diseased groups for that age bin.

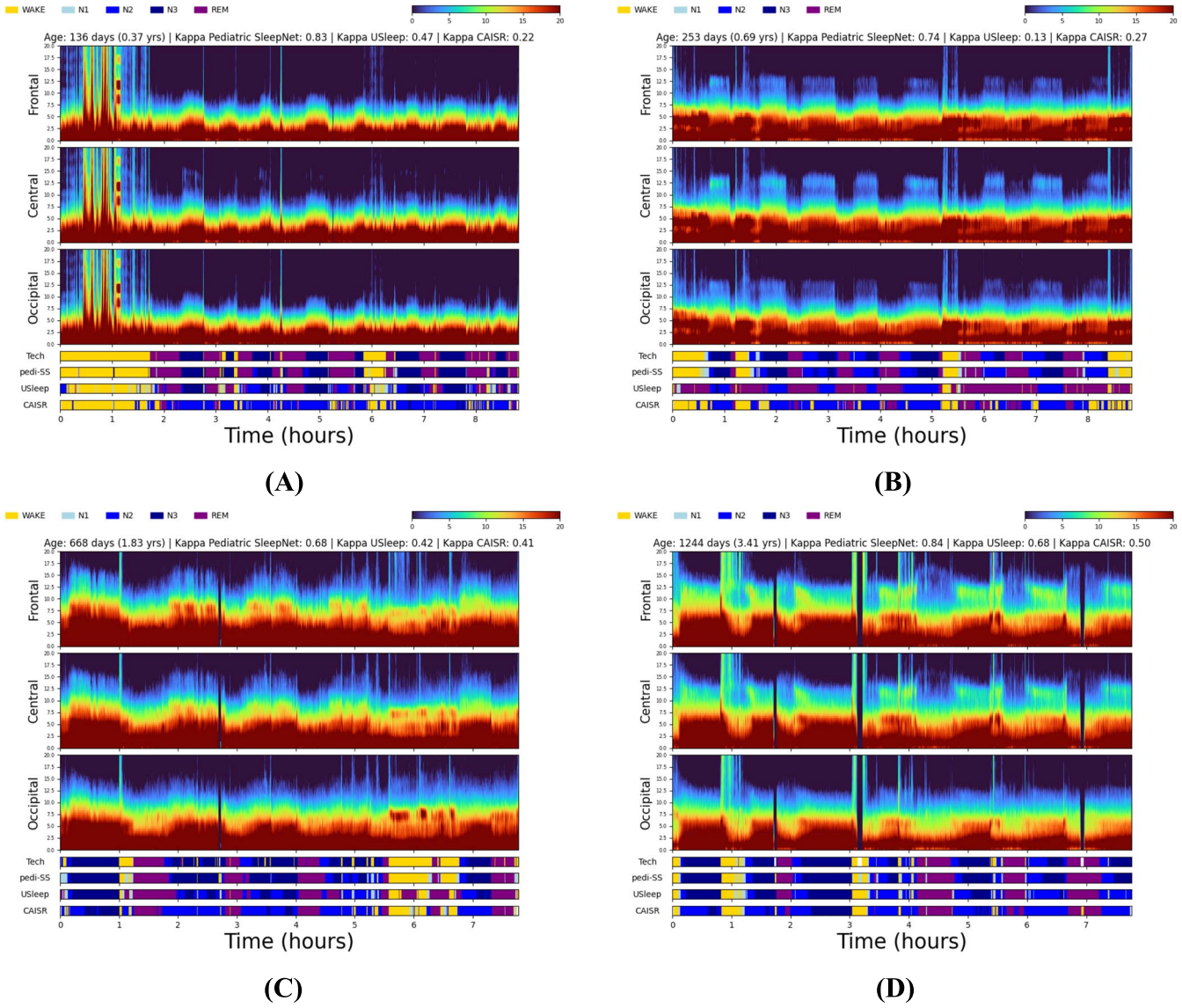


Figure 6. Representative examples of pediatric sleep staging showing averaged frontal, central, and occipital EEG spectrograms, with hypnograms from technician scoring (Tech), pediatric SleepNet, U-Sleep, and CAISR. Cohen's Kappa values indicate agreement with the technician assigned sleep staging labels.

Table 3. Mean Cohen's Kappa for 3-class sleep staging (WAKE, N, REM) across the nine pediatric age groups

Age group	0–6 months	6–12 months	1–2 years	2–3 years	3–4 years	4–5 years	5–6 years	6–12 years	>12 years
Cohen's κ	0.66	0.73	0.74	0.74	0.77	0.78	0.78	0.81	0.78

N1 and N2 (22.5%), between N1 and REM (13.7%), and between N2 and N3 (19.4%), reflecting known challenges in scoring these transitions in young children. In the case of the PATS dataset, the model obtained a slightly lower mean Cohen's Kappa of 0.6921. As seen in the corresponding confusion matrix in Figure 8B, agreement was highest for N3 (96.1%) and REM (94.1%), while N1 and N2 stages exhibited greater confusion. Specifically, N2 was misclassified as N3 in 26.7% of epochs, and N1 misclassified as REM in 20.6% of epochs, and as N2 in 19.4% of epochs-both higher than in the BCH or CHAT datasets. This suggests that cohort-specific differences in sleep architecture may impact model performance.

Discussion

This study presents a comprehensive evaluation of a deep learning network, pediatric SleepNet, for automated sleep staging in pediatric populations. Through rigorous internal and external validation, we demonstrate the model's ability to reliably classify sleep stages across a wide developmental spectrum, outperform existing models, generalize across diverse clinical diagnoses, and exhibit robust performance on external datasets. The findings demonstrate that sleep staging performance is sensitive to specific clinical conditions, particularly those associated with abnormal neural development or disrupted EEG signals.

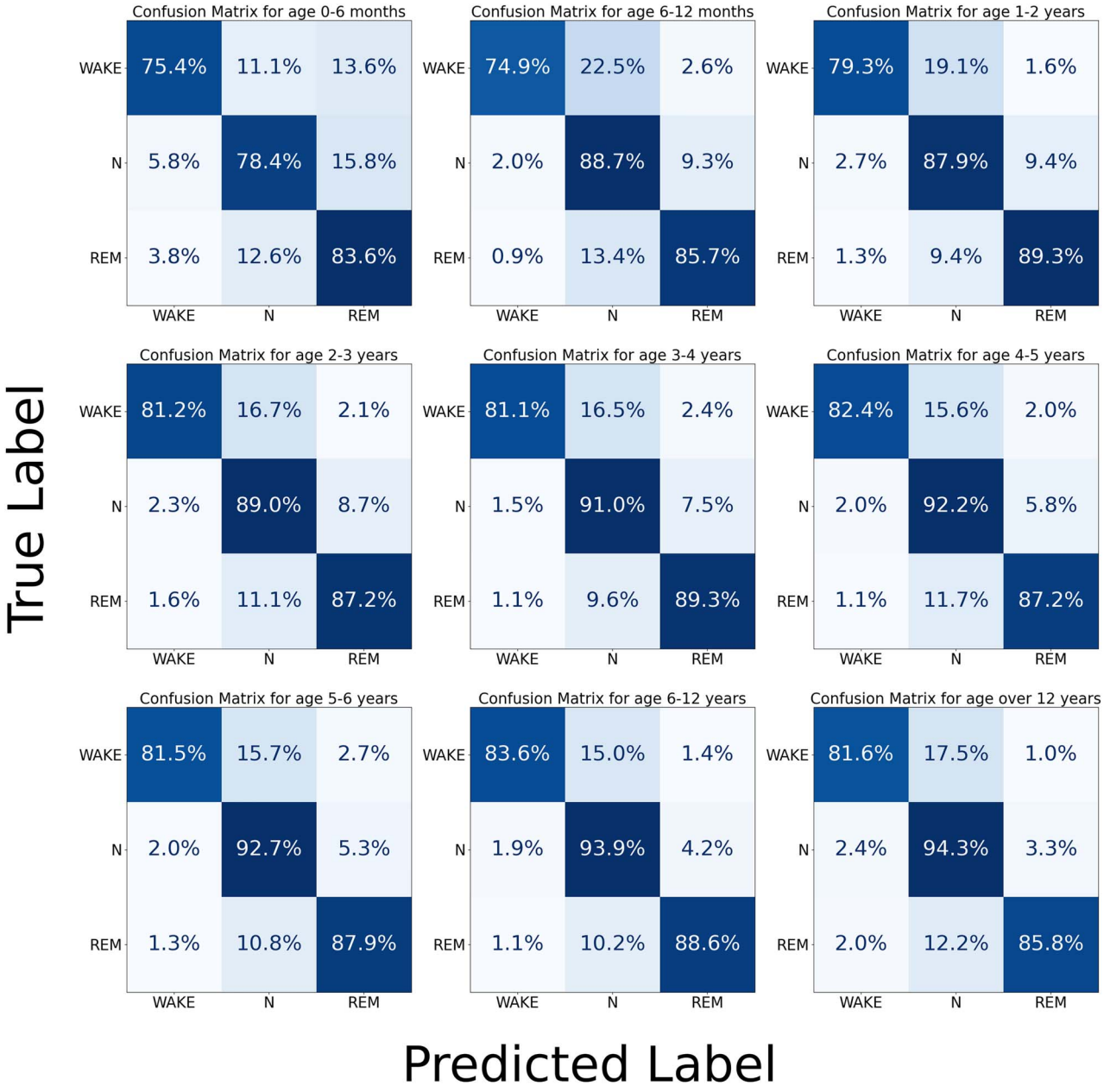


Figure 7. Confusion matrices for 3-class sleep staging (WAKE, NREM, REM) across nine pediatric age groups. Each matrix shows the distribution of model predictions relative to expert annotations. N includes all non-REM sleep stages (N1, N2, N3).

Notably, the performance gap between diseased and non-diseased groups tended to narrow with age, and for most conditions, model performance converged in later childhood. This suggests that developmental maturation may mitigate the impact of certain neurological diagnoses on automated sleep staging.

The findings underscore the feasibility and clinical utility of AI-assisted pediatric sleep staging, especially in light of the limitations of manual scoring in younger populations.

Our results demonstrate a strong age-dependent trajectory in model performance, with staging accuracy improving markedly from infancy to adolescence, mirroring neurodevelopmental maturation. The model achieves high agreement in older children and adolescents and exhibits stable performance across sexes and disease categories. Notably, the 3-class staging variant offers a practical alternative in settings where fine-grained staging is unreliable. Finally, pediatric SleepNet generalizes well to external pediatric datasets, highlighting its robustness and potential for real-world deployment.

Overall, these results demonstrate that pediatric SleepNet generalizes well to external pediatric populations, maintaining high accuracy across most sleep stages and strong agreement with expert labels. While slight performance drops were observed relative to the BCH test set, especially for light non-REM stages, the model's robustness across independently curated datasets underscores its clinical applicability in diverse pediatric settings.

Developmental trends in sleep staging performance

We observe in the results that there exists strong age-dependency of sleep staging in the pediatric age groups. In particular, we observed a progressive increase in Cohen's Kappa from infancy through adolescence, with the lowest agreement in infants under six months ($\kappa \approx 0.49$) and highest in children

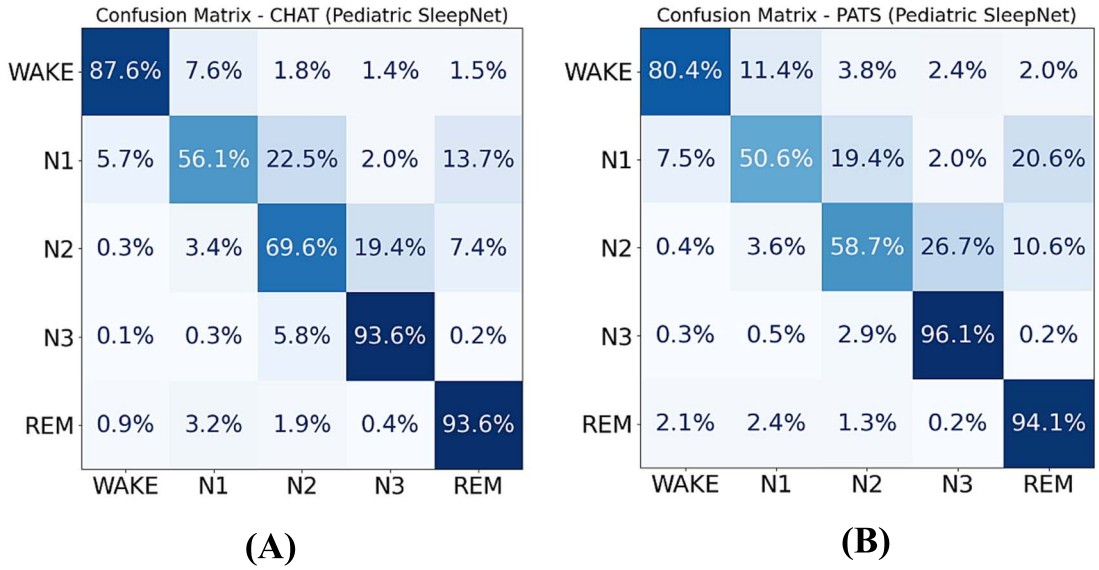


Figure 8. Confusion matrices for sleep staging performance of the pediatric SleepNet model on two external pediatric cohorts: CHAT (A) and PATS (B).

older than twelve years ($\kappa > 0.72$). This pattern aligns with well-established neurodevelopmental changes in sleep physiology and EEG morphology. In early infancy, sleep architecture is evolving and immature, with poorly defined stage boundaries, frequent state transitions, absence of consolidated sleep cycles, and incomplete expression of canonical EEG features such as spindles, K-complexes, and delta activity [32–34]. This increases ambiguity in the reference labels and likely contributes to the lower model performance observed in the <6-month age group. Additionally, the older age bins include more recordings than early infancy. Therefore, differences in sample size may also contribute to lower performance in younger subjects.

The high prevalence of indeterminate or transitional sleep patterns in infants presents significant challenges to both manual and automated scoring [35]. For example, the model has particular difficulty distinguishing N1 from REM sleep under 6 months, which can be explained by the similarly slower EEG backgrounds relative to older age groups. As the brain matures, particularly across the first few years of life, sleep stages become more distinct, circadian and homeostatic regulation stabilizes, and EEG features such as slow-wave activity and spindles become more reliable and stage-specific [36–38]. These maturational changes enhance the discriminability of sleep stages, thereby improving both human scoring reliability and deep learning model performance.

A decrease in delta power and a redistribution of sleep intensity throughout the night are associated with pubertal remodeling and synaptic pruning in adolescence. These changes are indicative of decreased cortical synchronization and thalamocortical circuit maturation. Improved scoring agreement and the relative ease with which models identify stages in later childhood and adolescence can be explained by these developmental alterations that can decrease stage borders (e.g., lighter NREM, changing spindle profiles). Our findings reinforce the necessity of age-aware modeling strategies in pediatric sleep research and suggest that incorporating developmental priors may further enhance accuracy in the youngest age groups. We found that models trained specifically for very young age groups performed better than a unified pediatric model. These findings may reflect meaningful developmental differences in underlying sleep physiology and point to the importance of age-aware model development and use in pediatric populations.

Comparison with existing models

In pediatric populations, EEG features differ markedly from adults in frequency content, amplitude, and event morphology [39]. Our proposed model, pediatric SleepNet, consistently outperformed two widely used automated sleep staging models: U-Sleep and CAISR across all pediatric age groups. Both U-Sleep and CAISR models were trained on over 18 000 adult PSGs across multiple cohorts, yet we observed reduced performance in infants and toddlers when applied to our test set. This difference is likely related to the fact that pediatric SleepNet was specifically designed and trained for pediatric populations using age-stratified modeling, particularly improving performance in infants. In infants under six months of age, pediatric SleepNet achieved a mean Cohen's Kappa of 0.49, significantly higher than U-Sleep ($\kappa = 0.30$) and CAISR ($\kappa = 0.34$). These gains persisted through toddlerhood and early childhood, with the largest improvements observed in the 6–12 month and 1–2 year bins ($\Delta\kappa \approx 0.22$–$0.24$ vs. U-Sleep and $\approx$ 0.26–0.28 vs. CAISR). Statistical comparisons using paired t-tests confirmed the significance of these differences ($p <$ 1e-20 across most early bins). Even in late childhood and adolescence-where all models performed better-pediatric SleepNet maintained a consistent performance advantage, although the margin narrowed (e.g. $\kappa = 0.76$ for pediatric SleepNet vs. 0.74 for U-Sleep in the 13–14 year bin).

Sex-based performance stability

We observed no statistically significant differences in sleep staging performance between male and female subjects across any of the age bins, with *p*-values well above .05 in all comparisons. The overall mean Cohen's Kappa was virtually identical—0.678 for males and 0.680 for females—demonstrating the robustness of our pediatric SleepNet model to sex-related biological variability. Even when stratified into narrower developmental age bins, such as 6–12 months or 6–12 years, performance remained stable between sexes, with overlapping confidence intervals and consistent accuracy.

This lack of sex-based discrepancy is encouraging, particularly in the context of pediatric populations, where sex-related differences in EEG maturation, sleep architecture, or arousal thresholds have been reported in some studies. For instance, Markovic et al. [40] found that in early adolescence, females show significantly greater sleep spindle activity and overall brain connectivity during sleep than males, suggesting stronger thalamocortical circuitry and functional connectivity. However, these physiological differences appear not to translate into measurable disparities in the automated sleep staging performance. This suggests that while there may be sex-linked neurophysiological differences in sleep, they do not meaningfully affect the reliability or accuracy of automatic sleep staging when models are properly trained. This supports the generalizability of pediatric staging models across male and female children without requiring sex-specific calibration.

Effect of clinical diagnoses

The presence of neurological and developmental disorders significantly impacted the sleep staging performance of our deep learning model. Across a range of ICD-10 diagnoses, we observed consistently lower Cohen's Kappa values in patients with diagnosed conditions compared to their age-matched non-diseased counterparts. This effect was particularly pronounced in cohorts with epilepsy, Down syndrome, pervasive developmental disorders, and hyperkinetic disorders, where statistically significant differences in model performance were observed across nearly all age groups as presented in Figure 5.

Among these diagnoses, epilepsy exhibited the most consistent degradation in model performance. This aligns with prior studies noting that epilepsy is associated with abnormal EEG patterns, such as interictal epileptiform discharges, focal slowing, and atypical sleep architecture, which can confound both human and automated scorers [40–43]. Similarly, children with Down syndrome often present with craniofacial abnormalities, obstructive sleep apnea, and altered EEG rhythms during sleep, which may contribute to reduced staging accuracy [44–46]. In the case of pervasive developmental disorders, which include autism spectrum disorders, sleep disturbances are common, often accompanied by atypical sleep–wake regulation, reduced REM sleep, and abnormalities in circadian rhythms [47–49]. These deviations from typical sleep physiology may make it harder for the model to generalize standard EEG features to such populations. Prior literature has highlighted altered spectral EEG features in children with autism during sleep, including increased theta and reduced sigma activity, which could interfere with the detection of spindles or stage-specific transitions [50]. We also noted degraded performance in children with hyperkinetic disorders, such as attention-deficit/hyperactivity disorder, especially in younger age groups. This population is known to have increased sleep onset latency, frequent arousals, and reduced sleep efficiency, often accompanied by increased beta activity during NREM sleep, which could challenge the model's ability to detect canonical sleep patterns [51, 52].

In children with hydrocephalus or congenital malformations of the nervous system, the effect on model performance was less consistent, likely due to smaller sample sizes and the heterogeneity of underlying pathophysiology. Some age groups even showed parity between diseased and non-diseased groups, suggesting that certain diagnoses may impact sleep staging reliability more than others, depending on the severity and nature of the EEG abnormalities. Taken together, these results underscore the importance of accounting for clinical diagnoses when deploying automated sleep staging models in pediatric populations. Models trained on general pediatric data may not generalize well to specific clinical subgroups without targeted adaptation or augmentation. Future work should explore diagnosis-aware training strategies and consider integrating clinical metadata as auxiliary inputs to improve robustness and interpretability.

Limitations and future scope

Several limitations merit further consideration in future work. Although our model achieved strong agreement with technician-provided reference labels, these labels themselves may be subject to inter-rater variability, especially in younger age groups where sleep stage boundaries are less distinct [53]. While technician annotations remain the clinical gold standard, future work should explore multi-expert consensus labels to further enhance training fidelity and to enable comparison of algorithm-expert reliability with expert-expert reliability.

We tested generalization to two external pediatric cohorts (CHAT and PATS), but both datasets were derived from American clinical populations. The extent to which our model generalizes to global populations with different sleep environments, cultural practices, or recording protocols remains to be validated. Furthermore, although we showed robustness across a wide range of ICD-10 diagnoses, our analysis was limited to categories with sufficient sample size. Rare disorders or highly comorbid conditions may require specialized modeling approaches or task-specific fine-tuning.

Our current approach treats sleep staging as a standalone task. In reality, sleep architecture is often influenced by and must be interpreted in the context of respiratory events, arousals, and limb movements. Joint modeling of sleep stages and other events may yield richer representations and improved performance.

Data on sleep staging in neonates, particularly pre-term versus full-term newborns, are still scarce and inconsistent, and comparability is hampered by variations in scoring standards, sensor configurations, and clinical settings. This gap limits the construction and validation of models in the earliest developmental window.

Our model operates on EEG, EOG, and EMG channels commonly recorded in clinical PSGs, performance in resource-constrained or ambulatory settings with fewer channels remains an open question. Future work should investigate the trade-off between modality reduction and staging accuracy to enable broader deployment in home-based or wearable systems.

Conclusion

This study presents a state of the art deep learning model for pediatric sleep staging, evaluated on the largest clinically annotated pediatric PSG corpus to date. Our model demonstrates robust performance across a wide age spectrum, is largely agnostic to sex, with clear improvements in staging accuracy from infancy through adolescence. By outperforming existing models trained on adults, particularly in younger children, pediatric SleepNet addresses longstanding challenges in pediatric sleep research-namely, the variability in EEG maturation and the scarcity of reliable automated tools tailored for infants and children. Disorders of the brain likely to impact EEG morphology expectedly degraded performance.

Acknowledgments

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129,

UL1RR024134, UL1RR024989). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24HL114473, 75N92019R002). The Pediatric Adenotonsillectomy Trial for Snoring (PATS) study was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (U01HL125307, U01HL125295). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24HL114473, 75N92019R002).

Funding

This work was supported by grants from the NIH (R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119) and AWS.

Disclosure statement

Financial disclosure: Dr. Westover is a co-founder, scientific advisor, and consultant to Beacon Biosignals and holds personal equity in the company. Dr. Clifford has received research funding from the NSF, NIH, and LifeBell AI, and has received 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, the Rett Research Foundation, and AliveCor Inc. Dr. Clifford also holds advisory roles and financial interests in AliveCor Inc and NextSense Inc, and serves as Chief Technology Officer of MindChild Medical with significant stock ownership. Dr. Thomas is a co-inventor of intellectual property licensed by Beth Israel Deaconess Medical Center to MyCardio, LLC (cardiopulmonary sleep spectrogram) and has submitted patents (respiratory self-similarity, Enhanced Expiratory Rebreathing Space) related to treatment and detection of high loop gain sleep apnea; he also provides paid consulting services to GLG Councils, Guidepoint, Beacon Biosignals, and Jazz Pharmaceuticals. Dr. Stone reports grant funding from Eli Lilly and consults for Axsome Therapeutics. Dr. Stone also receives a stipend from Sleep Research Society as deputy editor for *SLEEP*. Dr. Maski consults for Alkermes, Avadel, Harmony Biosciences, Jazz Pharmaceuticals, and Takeda Pharmaceuticals; has received grant funding from Harmony Biosciences and Jazz Pharmaceuticals; serves as Data and Safety Monitoring Board chair for Idorsia; and is a collaborator on clinical trials sponsored by Alkermes and Takeda. Dr. Sun reports grant funding via a Strategic Research Award from the American Academy of Sleep Medicine.

Non-financial disclosure: Dr. Trotti is a member of the Board of Directors of the American Academy of Sleep Medicine; the opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the American Academy of Sleep Medicine. All other authors report no non-financial conflicts of interest relevant to this work.

Data availability

The BCH Sleep Study Corpus is hosted on the Brain Data Science Platform (BDSP) and can be accessed at https://bdsp.io/content/l8c86mgywuneuy74ae71/1.0.1/. Researchers can request access and download the dataset through AWS using the instructions provided on the website. The code is available at https://github.com/bdsp-core/BCH-SleepStaging.

References

1. Worley SL. The extraordinary importance of sleep: the detrimental effects of inadequate sleep on health and public safety drive an explosion of sleep research. *Pharm Ther*. 2018;**43**:758.
2. Vyazovskiy VV. Sleep, recovery, and metaregulation: explaining the benefits of sleep. *Nat Sci Sleep*. 2015;**7**:171–184. https://doi.org/10.2147/NSS.S54036
3. Scott AJ, Webb TL, Martyn-St James M, Rowse G, Weich S. Improving sleep quality leads to better mental health: a meta-analysis of randomised controlled trials. *Sleep Med Rev*. 2021;**60**:101556. https://doi.org/10.1016/j.smrv.2021.101556
4. Lumeng JC, Chervin RD. Epidemiology of pediatric obstructive sleep apnea. *Proc Am Thorac Soc*. 2008;**5**(2):242–252. https://doi.org/10.1513/pats.200708-135MG
5. Beebe DW, Rausch J, Byars K, *et al*. Neuropsychological effects of pediatric obstructive sleep apnea. *J Int Neuropsychol Soc*. 2004;**10**(7):962–975. https://doi.org/10.1017/S135561770410708X
6. Malow BA, McGrew SG. Sleep disturbances and autism. *Sleep Med Clin*. 2008;**3**(3):479–488. https://doi.org/10.1016/j.jsmc.2008.04.004
7. Bruni O, Novelli L. Sleep disorders in children. *BMJ Clin Evid*. 2010;**2010**:2304.
8. Berry R, Quan S, Abreu A. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, Version 2.6. American Academy of Sleep Medicine: Darien, IL; 2020.
9. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med*. 2013;**9**(1):81–87. https://doi.org/10.5664/jcsm.2350
10. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med*. 2016;**12**(6):885–894. https://doi.org/10.5664/jcsm.5894
11. 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). https://doi.org/10.1093/sleep/zsaf134
12. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-sleep: resilient high-frequency sleep staging. *NPJ Digit Med*. 2021;**4**(1):72. https://doi.org/10.1038/s41746-021-00440-5
13. Einizade A, Nasiri S, Sardouie SH, Clifford GD. ProductGraph-SleepNet: sleep staging using product spatio-temporal graph learning with attentive temporal aggregation. *Neural Netw*. 2023;**164**:667–680. https://doi.org/10.1016/j.neunet.2023.05.016
14. Bruni O. Artificial intelligence in pediatric sleep staging: a new era or a complementary tool? *Sleep*. 2025;**48**(7). https://doi.org/10.1093/sleep/zsaf067
15. Moeller AL, Perslev M, Paulsrud C, *et al*. Artificial intelligence or sleep experts: comparing polysomnographic sleep staging in children and adolescents. *Sleep*. 2025;**48**(7). https://doi.org/10.1093/sleep/zsaf053
16. Somaskandhan P, Leppänen T, Terrill PI, *et al*. Deep learning-based algorithm accurately classifies sleep stages in preadolescent children with sleep-disordered breathing symptoms and age-matched controls. *Front Neurol*. 2023;**14**:1162998. https://doi.org/10.3389/fneur.2023.1162998
17. Vaquerizo-Villar F, Alvarez D, Kraemer JF, *et al*. Automatic sleep staging in children with sleep apnea using photoplethysmography and convolutional neural networks. *Annu Int Conf IEEE Eng Med Biol Soc*. 2021;**2021**:216–219. https://doi.org/10.1109/EMBC46164.2021.9629995

18. Bernardi G, Betta M, Ricciardi E, Pietrini P, Tononi G, Siclari F. Regional delta waves in human rapid eye movement sleep. *J Neurosci*. 2019;**39**(14):2686–2697. https://doi.org/10.1523/JNEUROSCI.2298-18.2019
19. Lo JC, Ong JL, Leong RL, Gooley JJ, Chee MW. Cognitive performance, sleepiness, and mood in partially sleep-deprived adolescents: the need for sleep study. *Sleep*. 2016;**39**(3):687–698. https://doi.org/10.5665/sleep.5552
20. Yu X, Quante M, Rueschman M, *et al*. Emergence of racial/ethnic and socioeconomic differences in objectively measured sleep–wake patterns in early infancy: results of the Rise & SHINE study. *Sleep*. 2021;**44**(3). https://doi.org/10.1093/sleep/zsaa193
21. Marcus CL, Moore RH, Rosen CL, *et al*. A randomized trial of adenotonsillectomy for childhood sleep apnea. *N Engl J Med*. 2013;**368**(25):2366–2376. https://doi.org/10.1056/NEJMoa1215881
22. Rosen CL, Larkin EK, Kirchner HL, *et al*. Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: association with race and prematurity. *J Pediatr*. 2003;**142**(4):383–389. https://doi.org/10.1067/mpd.2003.28
23. Hunter SJ, Gozal D, Smith DL, Philby MF, Kaylegian J, Kheirandish-Gozal L. Effect of sleep-disordered breathing severity on cognitive performance measures in a large community cohort of young school-aged children. *Am J Respir Crit Care Med*. 2016;**194**(6):739–747. https://doi.org/10.1164/rccm.201510-2099OC
24. Wang R, Bakker JP, Chervin RD, *et al*. Pediatric Adenotonsillectomy Trial for Snoring (PATS): protocol for a randomised controlled trial to evaluate the effect of adenotonsillectomy in treating mild obstructive sleep-disordered breathing. *BMJ Open*. 2020;**10**(3):e033889. https://doi.org/10.1136/bmjopen-2019-033889
25. Lee H, Li B, DeForte S, *et al*. A large collection of real-world pediatric sleep studies. *Sci Data*. 2022;**9**(1):421. https://doi.org/10.1038/s41597-022-01545-6
26. Tripathi A, Ganglberger W, Sun H, *et al*. The Boston Children's Hospital Sleep Corpus: a collection of 15,695 annotated pediatric polysomnograms. *Sleep*. 2025:zsaf273. https://doi.org/10.1093/sleep/zsaf273
27. Gramfort A, Luessi M, Larson E, *et al*. MEG and EEG data analysis with MNE-Python. *Front Neurosci*. 2013;**7**:267. https://doi.org/10.3389/fnins.2013.00267
28. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer International Publishing; 2015:234-241, https://doi.org/10.1007/978-3-319-24574-4_28
29. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;**20**(1):37–46. https://doi.org/10.1177/001316446002000104
30. 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):1351–1358. https://doi.org/10.1093/jamia/ocy064
31. Redline S, Cook K, Chervin RD, *et al*. Adenotonsillectomy for snoring and mild sleep apnea in children: a randomized clinical trial. *JAMA*. 2023;**330**(21):2084–2095. https://doi.org/10.1001/jama.2023.22114
32. Grigg-Damberger MM. The visual scoring of sleep in infants 0 to 2 months of age. *J Clin Sleep Med*. 2016;**12**(3):429–445. https://doi.org/10.5664/jcsm.5600
33. Dereymaeker A, Pillay K, Vervisch J, *et al*. Review of sleep-EEG in preterm and term neonates. *Early Hum Dev*. 2017;**113**:87–103. https://doi.org/10.1016/j.earlhumdev.2017.07.003
34. McClain IJ, Lustenberger C, Achermann P, Lassonde JM, Kurth S, LeBourgeois MK. Developmental changes in sleep spindle characteristics and sigma power across early childhood. *Neural Plast*. 2016;**2016**:1–9. https://doi.org/10.1155/2016/3670951
35. Satomaa AL, Saarenpää-Heikkilä O, Paavonen EJ, Himanen SL. The adapted American Academy of Sleep Medicine sleep scoring criteria in one-month-old infants: a means to improve comparability? *Clin Neurophysiol*. 2016;**127**(2):1410–1418. https://doi.org/10.1016/j.clinph.2015.08.013
36. Kurth S, Jenni OG, Riedner BA, Tononi G, Carskadon MA, Huber R. Characteristics of sleep slow waves in children and adolescents. *Sleep*. 2010;**33**(4):475–480. https://doi.org/10.1093/sleep/33.4.475
37. Feinberg I, Campbell IG. Longitudinal sleep EEG trajectories indicate complex patterns of adolescent brain maturation. *Am J Physiol Regul Integr Comp Physiol*. 2013;**304**(4):R296–R303. https://doi.org/10.1152/ajpregu.00422.2012
38. D'Atri A, Novelli L, Ferrara M, Bruni O, De Gennaro L. Different maturational changes of fast and slow sleep spindles in the first four years of life. *Sleep Med*. 2018;**42**:73–82. https://doi.org/10.1016/j.sleep.2017.11.1138
39. Baumert M, Hartmann S, Phan H. Automatic sleep staging for the young and the old–evaluating age bias in deep learning. *Sleep Med*. 2023;**107**:18–25. https://doi.org/10.1016/j.sleep.2023.04.002
40. Markovic A, Kaess M, Tarokh L. Gender differences in adolescent sleep neurophysiology: a high-density sleep EEG study. *Sci Rep*. 2020;**10**(1):15935. https://doi.org/10.1038/s41598-020-72802-0
41. Proost R, Heremans E, Lagae L, Van Paesschen W, De Vos M, Jansen K. Automated sleep staging on reduced channels in children with epilepsy. *Front Neurol*. 2024;**15**:1390465. https://doi.org/10.3389/fneur.2024.1390465
42. Conrad EC, Revell AY, Greenblatt AS, *et al*. Spike patterns surrounding sleep and seizures localize the seizure-onset zone in focal epilepsy. *Epilepsia*. 2023;**64**(3):754–768. https://doi.org/10.1111/epi.17482
43. Wang YQ, Zhang MQ, Li R, Qu WM, Huang ZL. The mutual interaction between sleep and epilepsy on the neurobiological basis and therapy. *Curr Neuropharmacol*. 2018;**16**(1):5–16. https://doi.org/10.2174/1570159X15666170509101237
44. Shott SR, Amin R, Chini B, Heubi C, Hotze S, Akers R. Obstructive sleep apnea: should all children with Down syndrome be tested? *Arch Otolaryngol Head Neck Surg*. 2006;**132**(4):432–436. https://doi.org/10.1001/archotol.132.4.432
45. Horne RS, Wijayaratne P, Nixon GM, Walter LM. Sleep and sleep disordered breathing in children with Down syndrome: effects on behaviour, neurocognition and the cardiovascular system. *Sleep Med Rev*. 2019;**44**:1–11. https://doi.org/10.1016/j.smrv.2018.11.002
46. Breslin J, Spanò G, Bootzin R, Anand P, Nadel L, Edgin J. Obstructive sleep apnea syndrome and cognition in Down syndrome. *Dev Med Child Neurol*. 2014;**56**(7):657–664. https://doi.org/10.1111/dmcn.12376
47. Souders MC, Zavodny S, Eriksen W, *et al*. Sleep in children with autism spectrum disorder. *Curr Psychiatry Rep*. 2017;**19**(6):34. https://doi.org/10.1007/s11920-017-0782-x
48. Carnett A, McLay L, Hansen S, France K, Blampied N. Sleep problems in children and adolescents with autism: type, severity and impact. *J Dev Phys Disabil*. 2021;**33**(6):977–991. https://doi.org/10.1007/s10882-020-09783-5
49. Chen H, Yang T, Chen J, *et al*. Sleep problems in children with autism spectrum disorder: a multicenter survey. *BMC Psychiatry*. 2021;**21**(1):406. https://doi.org/10.1186/s12888-021-03405-w

50. Limoges E, Mottron L, Bolduc C, Berthiaume C, Godbout R. Atypical sleep architecture and the autism phenotype. *Brain*. 2005;**128**(5):1049–1061. https://doi.org/10.1093/brain/awh425
51. Cortese S, Faraone SV, Konofal E, Lecendreux M. Sleep in children with attention-deficit/hyperactivity disorder: meta-analysis of subjective and objective studies. *J Am Acad Child Adolesc Psychiatry*. 2009;**48**(9):894–908. https://doi.org/10.1097/CHI.0b013e3181ac09c9
52. Gruber R, Sadeh A, Raviv A. Instability of sleep patterns in children with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*. 2000;**39**(4):495–501. https://doi.org/10.1097/00004583-200004000-00019
53. Ganglberger W, Nasiri S, Sun H, *et al*. Refining sleep staging accuracy: transfer learning coupled with scorability models. *Sleep*. 2024;**47**(11). https://doi.org/10.1093/sleep/zsae202

SLEEP, 2026, **00**, 1–16
https://doi.org/10.1093/sleep/zsag064
Original Article