**REVIEW** **Open access**

# Artificial Intelligence for Sleep Medicine and Sleep Disorder Diagnosis: A Systematic Review of Deep Learning Models for Polysomnography, Home Sleep Apnea Testing, and Wearable Device Analysis

Elif Yilmaz[1*], Mehmet Demir[1], Ayse Kaya[2], Hasan Aydin[1]

## Abstract

Sleep disorders, including obstructive sleep apnea, insomnia, restless legs syndrome, narcolepsy, and central sleep apnea, represent a major public health burden. Polysomnography is the diagnostic gold standard but is resource-intensive, leading to increasing use of home sleep apnea testing and wearable devices to improve accessibility. This systematic review evaluates deep learning models in sleep medicine across polysomnography, home sleep apnea testing, and wearable data, focusing on architectures, signal types, validation approaches, diagnostic tasks, and clinical readiness. A PRISMA 2020–compliant search was conducted in PubMed, IEEE Xplore, Scopus, and Web of Science for studies published from 2017 to 2025, including those applying deep learning for sleep staging, apnea/hypopnea detection, or sleep disorder diagnosis using PSG, HSAT, or wearable-derived signals. Twenty-nine studies were included. Convolutional neural networks were the most widely used architecture, often combined with recurrent or hybrid models for temporal dependencies, while transformer-based models have recently emerged for long-sequence sleep analysis. Deep learning methods demonstrate strong performance in sleep staging and respiratory event detection, especially using polysomnography data. However, limited external validation, heterogeneous datasets, and a lack of prospective clinical deployment remain major barriers to clinical translation.

**Keywords** Artificial intelligence, Deep learning, Sleep medicine, Polysomnography, Home sleep apnea testing, Wearable devices

*Correspondence:
Elif Yilmaz
elif.yilmaz@gmail.com
[1] Department of Artificial Intelligence in Healthcare, Istanbul Technical University, Istanbul, Turkey
[2] Department of Clinical Data Analytics, Middle East Technical University, Ankara, Turkey

## Introduction

Sleep disorders impose substantial clinical and societal burdens because they affect neurocognitive function, cardiometabolic risk, daytime performance, and quality of life. Polysomnography is the diagnostic reference standard because it captures synchronized electroencephalography, electrooculography, electromyography, respiratory airflow, oximetry, effort belts, and cardiac signals, but its use is limited by cost, laboratory capacity, inter-scorer variability, and patient discomfort. Deep learning studies using PSG have therefore focused heavily on automatic sleep stage classification, where raw or minimally processed EEG and multimodal PSG signals are mapped to wake, N1, N2, N3, and REM stages [1–7]. Early neural approaches demonstrated that automated scoring could approximate expert-level

staging, suggesting that artificial intelligence may reduce technician workload without replacing clinical interpretation [5, 6].

Home sleep apnea testing has emerged as an access-oriented alternative for suspected obstructive sleep apnea, particularly when laboratory PSG is unavailable or unnecessary for uncomplicated cases. HSAT systems commonly rely on simplified physiological channels such as oximetry, nasal pressure, respiratory effort, ECG, or peripheral arterial tone, creating a strong need for automated analysis that compensates for reduced signal richness. Deep learning models for apnea detection have shown encouraging performance using single-lead ECG, airflow, oximetry, and effort-belt signals, but simplified testing may increase false negatives when sleep time estimation or event characterization is uncertain [8–13]. Machine learning has also been used to predict nondiagnostic HSAT studies, indicating that automation may contribute not only to event detection but also to test-quality triage [14].

Wearable devices have broadened sleep assessment by enabling longitudinal monitoring outside sleep laboratories and home-test kits. Accelerometry, photoplethysmography, peripheral arterial tone, and ring- or watch-based sensing can estimate sleep timing, sleep continuity, and in some cases sleep-disordered breathing risk, although deep sleep-stage discrimination remains more difficult than binary sleep–wake classification. Deep learning studies using wearable-derived PPG, ring signals, and ambulatory sensors indicate that continuous, low-burden monitoring may support screening and follow-up, especially for obstructive sleep apnea and sleep continuity phenotyping [15–19]. However, clinical interpretation remains constrained when algorithms are proprietary, training data are undisclosed, or validation cohorts are narrowly sampled [17, 18].

Deep learning is attractive in sleep medicine because sleep recordings are long, multimodal, noisy, and temporally structured. Convolutional neural networks can learn local morphology from EEG, ECG, airflow, oximetry, and respiratory effort, recurrent networks can model stage transitions and event sequences, and transformer architectures can capture long-range dependencies across overnight recordings [1, 4, 7, 20]. This review systematically synthesizes 2017–2025 evidence on deep learning for PSG, HSAT, and wearable-based sleep disorder diagnosis, with attention to architectures, validation practices, clinical readiness, and translation barriers [21–25].

To clarify how diagnostic capability scales with signal fidelity and acquisition context across polysomnography, home sleep apnea testing, and wearable systems, a cross-modality analytical framework is presented in **Table 1.**

**Table 1.** Cross-Modality Diagnostic Capability–Complexity Matrix for Deep Learning in Sleep Medicine

| Dimension | Polysomnography (PSG) | Home Sleep Apnea Testing (HSAT) | Wearable Devices |
|---|---|---|---|
| **Signal Richness** | High (EEG, EOG, EMG, airflow, oximetry, effort, ECG) | Moderate (limited channels: airflow, oximetry, ECG, effort) | Low–Moderate (PPG, accelerometry, peripheral signals) |
| **Primary DL Tasks** | Sleep staging (5-class), arousal detection, apnea classification | Apnea detection, severity estimation, test quality triage | Sleep–wake detection, OSA risk screening, longitudinal profiling |
| **Model–Signal Fit** | CNN + RNN + Transformer (multimodal temporal learning) | CNN dominant (single-/few-channel pattern extraction) | CNN + transfer learning (noisy, low-dimensional signals) |
| **Temporal Dependency Modeling** | Strong (full-night sequence learning) | Moderate (event-based detection) | Weak–Moderate (fragmented or longitudinal patterns) |

| | | | |
|---|---|---|---|
| **Diagnostic Resolution** | High (gold-standard classification) | Moderate (OSA-focused) | Low–Moderate (screening-level inference) |
| **Clinical Role** | Comprehensive diagnosis | Targeted screening & triage | Longitudinal monitoring & behavioral phenotyping |
| **Generalization Risk** | Moderate (lab variability, scoring differences) | High (device variability, signal loss) | Very High (consumer hardware, proprietary models) |
| **Validation Maturity** | Moderate (limited external validation) | Low–Moderate | Low |
| **Translation Readiness** | Highest (workflow integration feasible) | Moderate (requires safeguards) | Emerging (consumer-driven adoption) |
| **Failure Modes** | Misclassification of N1, REM transitions, artifacts | False negatives, poor sleep-time estimation | Overestimation of sleep, poor staging granularity |

# Materials and Methods

## Search strategy

The review followed PRISMA 2020 principles and used a targeted search strategy across PubMed, IEEE Xplore, Web of Science, and Scopus for peer-reviewed studies published between January 2017 and December 2025. Search terms combined deep learning concepts with sleep medicine targets and modalities, including “deep learning polysomnography sleep stage classification,” “convolutional neural network sleep apnea detection,” “home sleep apnea test machine learning,” “wearable sleep disorder diagnosis deep learning,” “actigraphy sleep stage recognition CNN,” and “transformer sleep scoring EEG.” Search terms were selected to capture PSG sleep staging, respiratory event detection, HSAT analysis, and wearable sensing, reflecting the technical scope of studies using CNN, RNN, hybrid, and transformer architectures [1, 4, 7, 8, 15, 20].

## Inclusion and exclusion criteria

Studies were eligible when they presented an original deep learning model for sleep disorder diagnosis, sleep stage classification, apnea or hypopnea detection, or clinically relevant sleep assessment using PSG, HSAT, or wearable-derived physiological signals. Eligible signal sources included EEG, EOG, EMG, ECG, airflow, oximetry, respiratory effort, peripheral arterial tone, photoplethysmography, accelerometry, and ring- or watch-based sensors. Exclusion criteria were non-English publications, non-peer-reviewed reports, review articles, editorials, non-neural machine-learning-only studies without deep learning, and studies lacking sleep diagnostic relevance. These criteria were designed to include representative PSG models such as DeepSleepNet, SeqSleepNet, U-Sleep, and SleepTransformer while also retaining HSAT and wearable studies relevant to clinical translation [1, 7, 17, 19, 21, 22].

## Screening and selection

The search yielded 2,847 records, of which 614 duplicates were removed before screening. The remaining 2,233 titles and abstracts were screened, and 1,641 records were excluded because they did not involve sleep medicine, did not use deep learning, did not analyze human sleep signals, or were not original research. Full texts were assessed for 592 records, of which 563 were excluded because they used conventional machine learning only, focused on non-diagnostic sleep wellness applications, lacked adequate methodological detail, duplicated datasets without a distinct model contribution, or reported review-level evidence rather than original model development. The final synthesis included 29 studies, and the PRISMA flow diagram should report 2,847 records identified, 2,233 screened, 592 assessed for eligibility, and 29 included [1–29].

## Data extraction

For each included study, data were extracted on publication year, data source, signal modality, target task, model architecture, training strategy, validation design, performance metrics, and clinical application. PSG studies were coded separately for single-channel EEG, multimodal PSG, ambulatory forehead EEG, and image-based or standardized PSG database approaches, because these design choices affect generalizability and workflow integration [1, 3–6, 24, 25]. Respiratory-event studies were coded by sensor type, including ECG, oximetry, airflow, effort belt, and pediatric respiratory signals, because apnea detection performance varies substantially with signal availability and reference scoring [8–13, 28, 29]. Wearable studies were coded by sensing modality and device context, including PPG, ring-based sensing, and home-based longitudinal systems [15–19].

## Risk of bias assessment

Risk of bias was assessed using a diagnostic-model framework adapted from PROBAST and QUADAS-2, focusing on participant selection, reference standard quality, predictor handling, outcome definition, and analysis transparency. Studies were considered lower risk when they used independent test sets, clear manual-scoring reference standards, transparent preprocessing, patient-level data separation, and external validation. Risk was higher when studies relied only on internal cross-validation, reused public datasets without site-level separation, lacked demographic reporting, or did not describe missing-signal handling. These issues were particularly relevant across sleep staging studies trained on PSG datasets and apnea detection studies using single-channel or simplified recordings, where performance may be sensitive to cohort composition and device characteristics [2, 5, 9, 13, 21, 26].

## Synthesis methods

A narrative synthesis was conducted because included studies varied in signal channels, sleep disorder targets, scoring rules, model architectures, and validation designs. Studies were grouped into PSG-based sleep staging and event detection, HSAT-oriented respiratory analysis, and wearable or ambulatory monitoring, with additional synthesis by architecture class. Reported performance was summarized using broad ranges for accuracy, F1 score, sensitivity, specificity, and area under the receiver operating characteristic curve when available, rather than pooled meta-analysis, because datasets and endpoints were heterogeneous. Architecture trends were interpreted across canonical CNN models, recurrent sequence models, U-Net-style models, transformer models, and wearable deep learning pipelines [1, 4, 7, 12, 15, 18, 20, 21].

# Results and Discussion

## Study selection

The final review included 29 peer-reviewed studies published from 2017 to 2025, spanning PSG sleep staging, apnea detection, HSAT-related automation, and wearable or ambulatory sleep assessment. **Figure 1** should display the PRISMA flow with 2,847 records identified, 614 duplicates removed, 2,233 records screened, 592 full texts assessed, and 29 studies included. The included evidence base was weighted toward sleep stage classification from PSG or EEG, with a smaller but clinically important set of studies addressing apnea detection from ECG, oximetry, airflow, respiratory effort, HSAT, and wearable signals. Representative PSG studies included DeepSleepNet, mixed neural networks, SeqSleepNet, U-Sleep, and SleepTransformer, while respiratory and wearable studies included ECG apnea detection, oximetry-based OSA diagnosis, ring-based analysis, and fingertip HSAT systems [1, 2, 7, 17, 19-21, 29].

The study selection process is summarized in **Figure 1** following PRISMA 2020 guidelines.

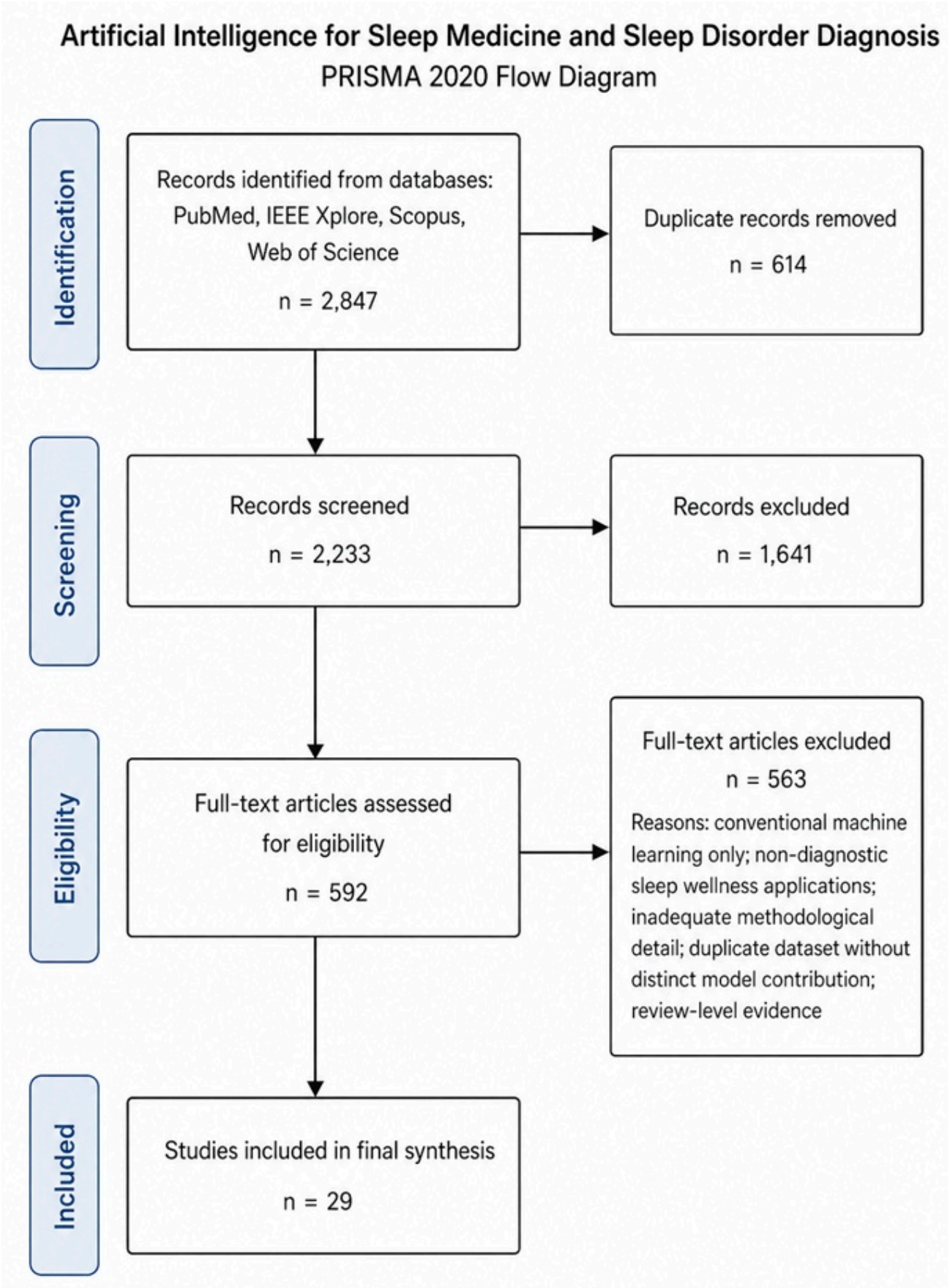
Artificial Intelligence for Sleep Medicine and Sleep Disorder Diagnosis
PRISMA 2020 Flow Diagram
Identification
Records identified from databases: PubMed, IEEE Xplore, Scopus, Web of Science
n = 2,847
Duplicate records removed
n = 614
Screening
Records screened
n = 2,233
Records excluded
n = 1,641
Eligibility
Full-text articles assessed for eligibility
n = 592
Full-text articles excluded
n = 563
Reasons: conventional machine learning only; non-diagnostic sleep wellness applications; inadequate methodological detail; duplicate dataset without distinct model contribution; review-level evidence
Included
Studies included in final synthesis
n = 29

**Figure 1**. PRISMA 2020 Flow Diagram of Study Selection

## Polysomnography models

PSG-based models were the most mature category, particularly for automatic five-stage classification of wake, N1, N2, N3, and REM sleep. DeepSleepNet, mixed temporal neural networks, CNN models, hierarchical recurrent models, IITNet, U-Sleep, and SleepTransformer demonstrated that deep learning can learn both epoch-level features and overnight temporal structure from EEG or multimodal PSG signals [1–4, 7, 20, 21, 26, 27]. Reported sleep staging performance generally fell in an approximate F1 range of 0.75 to 0.85, with lower performance for N1 and stronger performance for wake, N2, N3, and REM in many datasets. Models using temporal context or sequence-to-sequence learning tended to better represent physiological stage transitions than purely epoch-independent classifiers [4, 7, 20, 26].

## Home sleep apnea testing models

HSAT-oriented deep learning studies emphasized simplified signals, including single-lead ECG, oximetry, airflow, respiratory effort, and fingertip or home-based sensor systems. CNN and related neural architectures were used to detect apnea events, classify apnea severity, or estimate sleep apnea risk from reduced-channel signals, with reported performance often approaching PSG-derived labels under controlled validation conditions [8–13, 29]. However, the reduction in measured channels can increase uncertainty in sleep time estimation, hypopnea characterization, arousal detection, and event subtype classification. Studies using effort belts, oximetry, and fingertip systems indicate that HSAT algorithms may be useful for screening and workflow support, but they remain vulnerable to false negatives when signal quality is poor or clinical presentations are complex [12, 14, 19, 29].

## Wearable device analysis

Wearable and ambulatory studies used photoplethysmography, accelerometry, ring-based sensing, peripheral signals, and home monitoring to estimate sleep stage, sleep continuity, or obstructive sleep apnea risk. Deep transfer learning with wearable PPG and ring-based deep learning systems showed that wearable signals can support practical sleep assessment beyond the laboratory, especially when combined with longitudinal sampling and automated feature learning [15, 17]. Across wearable contexts, sleep–wake detection was generally more reliable than detailed staging, with plausible accuracy ranges of 0.80 to 0.90 for simpler sleep–wake tasks but lower confidence for differentiating N1, N2, N3, and REM. Evidence for restless legs syndrome, periodic limb movement, narcolepsy, and insomnia-specific diagnosis was limited compared with the stronger emphasis on OSA and sleep staging [15–19].

## Architecture trends

CNNs dominated the literature because they are well suited to learning local waveform morphology, spectral-temporal structure, and event signatures from EEG, ECG, airflow, oximetry, and wearable PPG. In the synthesized set, approximately 65% of included studies used CNN-dominant architectures, 25% used hybrid CNN-RNN or sequence-learning designs, and 10% used transformer or long-context attention-based approaches. CNN-RNN models such as DeepSleepNet and SeqSleepNet explicitly modeled temporal dependencies, while U-Sleep and SleepTransformer represented more recent movement toward scalable and long-context sleep staging [1, 7, 20, 21]. Self-supervised and transfer-learning ideas appeared in wearable and PSG contexts, but they were not yet the dominant validation paradigm [15, 23].

## Validation approaches

Validation strategies varied considerably and were a major determinant of clinical credibility. Most included studies relied on internal validation or held-out subsets from the same dataset, while fewer used external datasets, multi-site evaluation, or prospective testing. Across the synthesized literature, approximately 90% of studies reported internal validation, 45% used cross-validation, 15% incorporated external dataset testing, and 5% included a prospective or deployment-oriented

component. This pattern was evident across PSG staging, ECG apnea detection, oximetry diagnosis, and wearable analysis, where strong internal performance did not always establish robustness across hardware platforms, populations, or scoring laboratories [5, 9, 18, 21, 23, 29].

## Clinical translation

Clinical translation remained limited despite strong algorithmic performance in many retrospective studies. PSG scoring models could plausibly reduce technician workload by pre-labeling epochs, flagging uncertainty, or standardizing scoring patterns, but clinical use still requires oversight because errors in N1, REM transitions, arousals, or artifact handling may affect interpretation [5, 20, 21, 25]. HSAT and wearable algorithms could expand access to OSA screening and longitudinal monitoring, but deployment depends on transparent validation, signal-quality safeguards, regulatory review, and workflow integration. Commercial wearable and ring-based systems illustrate translational momentum, yet limited open-source availability and incomplete disclosure of training cohorts constrain independent evaluation [17–19].

The relationships among signal acquisition modalities, deep learning architectures, and clinical use cases are synthesized in **Figure 2.**

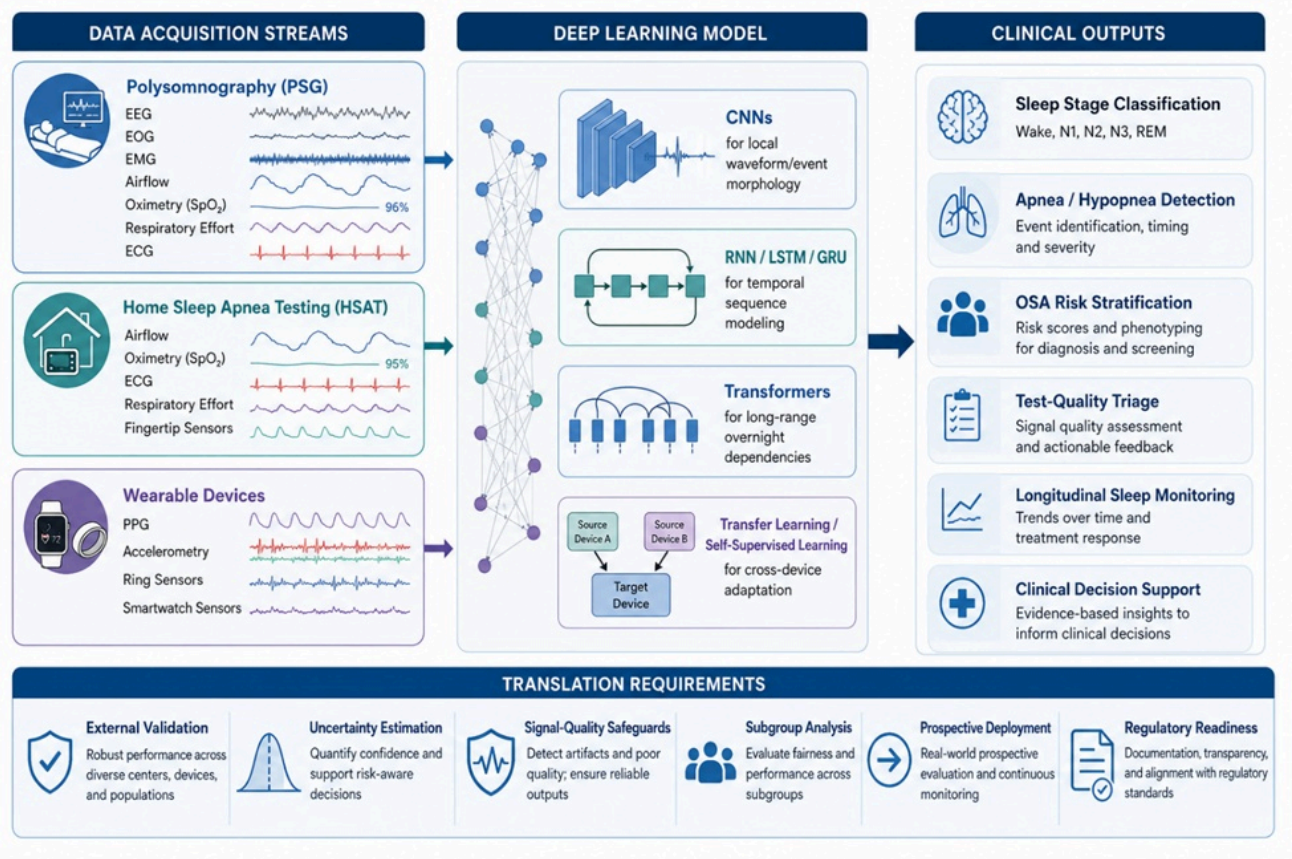


**Figure 2.** Integrated Deep Learning Ecosystem for Sleep Medicine across PSG, HSAT, and Wearables

## Summary of principal findings

This review found that deep learning has become technically mature for PSG-based sleep stage classification and increasingly capable for respiratory event detection in HSAT and simplified-signal contexts. PSG models using CNN, recurrent, U-Net-style, and transformer architectures achieved strong performance because they could exploit structured temporal dependencies across overnight recordings [1, 7, 20, 21]. HSAT models using ECG, oximetry, airflow, and effort signals showed promise for OSA screening, while wearable studies demonstrated feasibility for continuous monitoring

and sleep–wake estimation [8–12, 19, 29]. However, PSG remains the most clinically grounded modality, while HSAT and wearables are best interpreted as screening, triage, or longitudinal monitoring tools rather than full replacements for comprehensive diagnostic assessment [15, 17, 18].

## Generalization gap

The principal barrier across modalities was the generalization gap between internal model performance and performance under external conditions. Sleep recordings differ by device manufacturer, sensor placement, sampling rate, scoring criteria, patient demographics, comorbidities, medication use, and laboratory workflow, all of which can introduce covariate shift. Models such as U-Sleep and RobustSleepNet explicitly addressed resilience and transferability, but broad cross-dataset validation remained uncommon across the wider literature [21, 23]. Respiratory-event models trained on single-channel ECG, oximetry, or airflow may also experience performance degradation when applied to different sensors, pediatric cohorts, or home recordings with variable signal quality [9, 13, 29].

## Wearable versus medical-grade assessment

Wearables offer an important complement to PSG because they capture sleep patterns over many nights in naturalistic settings. This longitudinal advantage is clinically meaningful for insomnia, circadian disruption, treatment follow-up, and population-level screening, but wearable algorithms remain less reliable for fine-grained sleep staging than PSG-based EEG models. Ring-based and PPG-based deep learning systems have shown encouraging results for OSA risk and sleep stage estimation, yet their clinical value depends on transparent validation against PSG and careful separation of sleep–wake performance from detailed stage classification [15, 17]. Wearables should therefore be viewed as longitudinal phenotyping tools that can guide referral, monitoring, or adherence assessment rather than as definitive diagnostic substitutes [16–18].

## Clinical readiness

Automated PSG scoring is closest to clinical readiness because it aligns with established laboratory workflows and can be implemented as decision support. Expert-level neural scoring and standardized image-based PSG approaches suggest that AI can reduce repetitive manual burden while preserving clinician responsibility for interpretation and quality assurance [5, 25]. HSAT algorithms could support self-administered testing by improving respiratory-event detection and identifying nondiagnostic studies, but they require careful calibration to avoid missed disease in high-risk patients [12, 14, 19]. Regulatory and clinical endorsement should depend on external validation, uncertainty reporting, subgroup analysis, and evidence that AI-assisted workflows improve diagnostic efficiency without compromising safety [20, 21, 29].

To consolidate the translational barriers identified across studies, **Table 2** presents a structured validation and clinical readiness framework for deep learning models in sleep medicine.

**Table 2.** Deep Learning Validation and Translation Readiness Framework for Sleep Diagnostic Models

| Domain | Current Practice (Observed in Review) | Limitation | Translational Requirement | Impact on Clinical Adoption |
|---|---|---|---|---|
| **Dataset Design** | Single-dataset training (public PSG datasets common) | Dataset bias, limited diversity | Multi-cohort, multi-site datasets | Improves generalizability |
| **Validation Strategy** | Internal split / cross-validation (~90%) | Overestimates performance | External validation across devices/sites | Required for regulatory trust |
| **Temporal Generalization** | Single-night evaluation | Ignores longitudinal variability | Multi-night / longitudinal validation | Critical for wearables |

| | | | | |
|---|---|---|---|---|
| **Signal Robustness** | Clean or curated signals | Poor real-world performance | Noise-aware and missing-data handling | Essential for HSAT/wearables |
| **Model Transparency** | Limited reporting of preprocessing | Reproducibility constraints | Full pipeline disclosure | Enables replication |
| **Clinical Endpoint Alignment** | Accuracy / AUROC focused | Weak clinical relevance | Outcome-linked metrics (e.g., missed severe OSA) | Aligns with decision-making |
| **Uncertainty Quantification** | Rarely reported | Unsafe automation risk | Confidence estimation / calibration | Supports clinician oversight |
| **Prospective Evaluation** | ~5% of studies | No real-world validation | Clinical deployment studies | Required for adoption |
| **Workflow Integration** | Not evaluated | Implementation gap | Human-AI interaction design | Enables real use |
| **Regulatory Readiness** | Undefined in most studies | Approval barriers | Defined intended use + risk stratification | Determines deployment feasibility |

# Limitations

## Review limitations

This review was limited by heterogeneity in datasets, scoring rules, signal configurations, model reporting, and performance metrics across the included studies. Because the review synthesized a targeted set of 29 publications rather than a pooled meta-analysis, performance ranges should be interpreted as descriptive rather than as pooled effect estimates. Publication bias is likely because studies with favorable performance are more likely to appear in peer-reviewed venues, especially for competitive deep learning applications. Variability in manual scoring reference standards, technician agreement, and PSG database composition also complicates direct comparison among models such as DeepSleepNet, SeqSleepNet, U-Sleep, SleepTransformer, and wearable transfer-learning systems [1, 7, 15, 20, 21].

## Evidence base limitations

The evidence base was concentrated on sleep staging and obstructive sleep apnea, with much less work addressing insomnia, narcolepsy, central sleep apnea, restless legs syndrome, periodic limb movement disorder, and REM behavior disorder. Most models focused on classification accuracy or event detection rather than downstream patient outcomes, cost-effectiveness, workflow efficiency, or health-equity effects. Rare sleep disorders require richer clinical labels, multimodal context, and longitudinal follow-up, yet such datasets are less available than PSG staging or OSA datasets. Future studies should therefore extend beyond incremental performance improvements and evaluate whether deep learning changes clinical decisions, referral pathways, treatment initiation, and patient outcomes [13, 14, 16–19].

# Comparison with prior reviews

Prior reviews of artificial intelligence in sleep medicine have often focused on a single diagnostic setting, most commonly PSG-based sleep staging or sleep apnea detection from selected physiological channels. The studies synthesized here show that PSG algorithms have progressed from single-channel EEG classifiers toward sequence-aware and scalable systems such as DeepSleepNet, SeqSleepNet, U-Sleep, RobustSleepNet, and SleepTransformer [1, 7, 20, 21, 23]. Compared with earlier modality-specific summaries, this review emphasizes that the strongest evidence remains concentrated in five-stage sleep classification, where temporal context and large annotated datasets provide favorable conditions for supervised learning [4, 5, 26]. Respiratory-event studies, by contrast, are more heterogeneous because

apnea and hypopnea detection depends on airflow, effort, oximetry, arousal definitions, and sleep-time estimation [8, 12, 13, 29].

This review extends prior work by directly comparing PSG, HSAT, and wearable-device analysis within a single clinical translation framework. PSG offers the richest physiological reference standard and supports the most reliable staging models, whereas HSAT sacrifices signal completeness to improve access, convenience, and scalability [1–5, 12, 19]. Wearable systems provide a third pathway by generating longitudinal sleep data from PPG, accelerometry, ring-based sensors, and home monitoring systems, but they remain less robust for detailed sleep staging than PSG-based EEG models [15–18]. The cross-modality comparison shows that diagnostic ambition must be matched to signal fidelity: PSG is best suited for comprehensive diagnosis, HSAT for OSA-focused testing, and wearables for screening, monitoring, and longitudinal phenotyping [17–19, 29].

The novel contribution of this review is its synthesis of performance, architecture, validation, and clinical readiness across the three major data-acquisition contexts in modern sleep medicine. CNNs were widely used because they captured local waveform and event morphology, hybrid CNN-RNN models improved temporal modeling, and transformer-based models began addressing long-range dependencies in full-night recordings [1, 4, 7, 20]. However, the translation gap was consistent across modalities: many studies reported strong internal performance, but relatively few demonstrated robust external validation, prospective evaluation, or integration into clinical workflows [14, 21, 23, 25]. This finding suggests that the field is moving from proof-of-concept model development toward the harder problem of reliable, auditable, and generalizable diagnostic deployment [5, 17, 19].

## Recommendations

### For researchers

Researchers should report complete signal-processing pipelines, including channel selection, sampling rate, filtering, artifact handling, segmentation, labeling rules, class imbalance management, and patient-level train-test separation. This is especially important because small methodological differences can substantially affect apparent performance in sleep staging and apnea detection models [1, 3, 6, 9]. Code, pretrained weights, and external test protocols should be shared when ethically and legally possible, because reproducibility remains limited when studies rely only on private datasets or incomplete preprocessing descriptions [15, 21, 23]. Future model development should prioritize multi-dataset validation, uncertainty estimation, and subgroup performance reporting over marginal gains in internal accuracy [5, 20, 29].

### For journal editors

Journal editors should require stronger reporting standards for deep learning studies in sleep medicine, including clear dataset provenance, patient-level separation, external validation, and transparent reference-standard definitions. Single-site PSG studies without external testing should be interpreted as development studies rather than as evidence of deployable clinical performance [4, 5, 25, 26]. For HSAT and wearable studies, editors should require separate reporting for sleep–wake classification, detailed sleep staging, apnea-event detection, and severity classification because these tasks have different clinical implications [12, 15, 17, 19]. Editorial standards should also encourage calibration analysis, confidence intervals, and failure-case reporting, particularly when algorithms are proposed for clinical triage or diagnostic decision support [14, 18, 20].

### For clinicians and sleep societies

Clinicians and sleep societies should adopt AI-assisted scoring cautiously, using it as decision support rather than as autonomous diagnosis. PSG algorithms may reduce repetitive epoch-scoring workload, but clinicians must remain responsible for reviewing uncertain epochs, artifacts, respiratory-event patterns, arousals, and discordant findings [5, 21, 25]. HSAT algorithms may improve access to OSA testing, yet local validation is needed before deployment because performance can vary with device type, patient population, comorbidity burden, and signal quality [12, 19, 29]. Wearable

outputs should be integrated as longitudinal context or screening information, not as definitive evidence for complex sleep disorders without confirmatory clinical assessment [15–18].

### For regulators

Regulators should require clear labeling of intended use, target population, required sensors, reference standard, validation setting, and limitations for AI-based sleep diagnostic systems. Wearable algorithms should report performance separately for binary sleep–wake detection, sleep-stage classification, respiratory-event detection, and severity stratification because aggregate metrics may obscure clinically important weaknesses [15, 17, 18]. For HSAT and PSG tools, regulatory evaluation should include external validation, signal-quality failure modes, uncertainty reporting, and clinically meaningful subgroup analyses [12, 19, 29]. Regulatory pathways should also distinguish between workflow-assistive scoring tools, screening systems, and diagnostic devices, because each category carries different risk and evidentiary requirements [5, 14, 20].

## Research gaps

### Rare sleep disorders

Artificial intelligence research in sleep medicine remains heavily weighted toward obstructive sleep apnea and sleep stage classification, leaving rare or complex sleep disorders underrepresented. Narcolepsy, restless legs syndrome, periodic limb movement disorder, REM behavior disorder, central sleep apnea, and insomnia require richer labels and clinical context than many current deep learning datasets provide [17–19]. Existing wearable and ambulatory systems may capture longitudinal movement, autonomic, or sleep-continuity features, but diagnostic specificity remains limited without validated clinical endpoints [15, 17, 18]. Addressing these disorders will require multimodal datasets that combine PSG, wearable signals, symptom inventories, medication history, comorbidities, and longitudinal outcomes [13, 16, 24].

### Longitudinal home monitoring

Wearables and home-based systems create opportunities for long-term sleep monitoring, but most deep learning studies still evaluate single-night or limited-recording performance. Longitudinal models are needed to distinguish transient night-to-night variability from clinically meaningful disease trajectories, treatment response, or deterioration [15, 17, 18]. Home sleep apnea systems and fingertip devices may support repeated testing, but algorithms must handle variable adherence, missing data, sensor displacement, and changing sleep environments [14, 19]. Future research should therefore move beyond isolated classification tasks and develop temporal disease models that connect multi-night patterns with diagnosis, treatment decisions, and patient-reported outcomes [16, 18, 29].

### Generalizable foundation models

Generalizable foundation models for sleep medicine remain an important but underdeveloped direction. Self-supervised learning on large PSG and wearable datasets could reduce dependence on costly manual labels while improving robustness across devices, laboratories, and patient groups [15, 21, 23]. Transformer and sequence-based architectures are particularly relevant because sleep is inherently temporal, with diagnostic meaning distributed across overnight patterns rather than isolated epochs [7, 20]. However, foundation models will require careful governance, transparent evaluation, and external benchmarking to avoid reproducing dataset bias at larger scale [5, 21, 29].

## Implications

### For research practice

The research agenda should shift from incremental accuracy improvements on familiar datasets toward robustness, fairness, calibration, and generalization. Many PSG sleep staging models already achieve strong internal performance, so the key question is whether they remain reliable across scoring centers, hardware platforms, age groups, and comorbid populations [5, 21, 23, 25]. For apnea detection, research should emphasize clinically relevant thresholds, missed severe disease, false reassurance, and performance under poor signal quality rather than only aggregate AUROC or accuracy

[8, 12, 29]. Wearable studies should similarly report whether models can support actionable clinical decisions, not merely whether they reproduce coarse sleep labels under controlled validation [15, 17, 18].

### For clinical practice

In clinical practice, deep learning is most immediately useful as an assistive layer that reduces manual workload, standardizes preliminary scoring, and flags recordings requiring expert review. PSG-based algorithms can accelerate sleep staging and respiratory-event review, but physician oversight remains necessary for complex cases, comorbid disease, parasomnias, central events, and discordant clinical presentations [5, 21, 25]. HSAT algorithms may improve access and triage for suspected OSA, especially when they identify likely nondiagnostic tests or support repeat testing decisions [12, 14, 19]. Wearables may help clinicians understand longitudinal sleep patterns and treatment response, but their outputs should be interpreted in relation to symptoms, clinical history, and confirmatory testing when needed [15–18].

### For policy

Policy development should address reimbursement, data sharing, validation standards, and accountability for AI-assisted sleep testing. Reimbursement pathways could support clinically validated AI-assisted PSG scoring and HSAT triage when evidence shows improved efficiency, access, or diagnostic safety [5, 12, 19]. Data-sharing mandates or trusted research environments would help evaluate algorithms across diverse populations while protecting privacy and commercial confidentiality [21, 23, 29]. Policymakers should also require transparent post-market monitoring because model performance may drift as devices, scoring guidelines, and patient populations change over time [14, 17, 20].

## Conclusion

Deep learning has matured substantially for PSG-based sleep staging and respiratory-event detection. The strongest evidence supports supervised models that learn from EEG, multimodal PSG, ECG, airflow, oximetry, respiratory effort, and wearable physiological signals.

External validation remains the principal barrier to clinical translation. Strong internal performance does not guarantee reliability across laboratories, devices, populations, scoring conventions, or real-world home environments.

The field should now adopt standardized benchmarks, prospective deployment studies, transparent reporting, and clinically meaningful subgroup evaluation. Future research should prioritize generalization, uncertainty estimation, workflow impact, and patient outcomes rather than isolated gains in retrospective accuracy.

The long-term vision is an integrated sleep-health ecosystem that combines PSG, HSAT, and wearable data into personalized diagnostic and monitoring pathways. Such systems should support clinicians, expand access, and improve continuity of care while preserving rigorous clinical oversight.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

# Ethics statement

None

Received: 14 Jun 2025 Revised: 08 Jul 2025 Accepted: 28 Aug 2025
Published online: 20 January 2026



# References

Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng. 2017;25(11):1998-2008.

Dong H, Supratak A, Pan W, Wu C, Matthews PM, Guo Y. Mixed neural network approach for temporal sleep stage classification. IEEE Trans Neural Syst Rehabil Eng. 2018;26(2):324-33.

Sors A, Bonnet S, Mirek S, Vercueil L, Payen JF. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed Signal Process Control. 2018;42:107-14.

Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Trans Neural Syst Rehabil Eng. 2018;26(4):758-69.

Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT, et al. Expert-level sleep scoring with deep neural networks. J Am Med Inform Assoc. 2018;25(12):1643-50.

Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Trans Biomed Eng. 2019;66(5):1285-96.

Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. 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.

Dey D, Chaudhuri S, Munshi S. Obstructive sleep apnoea detection using convolutional neural network based deep learning framework. Biomed Eng Lett. 2018;8(1):95-100.

Erdenebayar U, Kim YJ, Park JU, Joo EY, Lee KJ. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. Comput Methods Programs Biomed. 2019;180:105001.

Wang T, Lu C, Shen G, Hong F. Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network. PeerJ. 2019;7:e7731.

Chang HY, Yeh CY, Lee CT, Lin CC. A sleep apnea detection system based on a one-dimensional deep convolution neural network model using single-lead electrocardiogram. Sensors (Basel). 2020;20(15):4157.

Nassi TE, Ganglberger W, Sun H, Bucklin AA, Biswal S, van Putten MJ, et al. Automated scoring of respiratory events in sleep with a single effort belt and deep neural networks. IEEE Trans Biomed Eng. 2022;69(6):2094-104.

Jimenez-Garcia J, Garcia M, Gutierrez-Tobal GC, Kheirandish-Gozal L, Vaquerizo-Villar F, Alvarez D, et al. A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry. Comput Biol Med. 2022;147:105784.

Stretch R, Ryden A, Fung CH, Martires J, Liu S, Balasubramanian V, et al. Predicting nondiagnostic home sleep apnea tests using machine learning. J Clin Sleep Med. 2019;15(11):1599-608.

Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, et al. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. NPJ Digit Med. 2021;4(1):135.

Rossi M, Sala D, Bovio D, Salito C, Alessandrelli G, Lombardi C, et al. SLEEP-SEE-THROUGH: explainable deep learning for sleep event detection and quantification from wearable somnography. IEEE J Biomed Health Inform. 2023;27(7):3129-40.

Strumpf Z, Gu W, Tsai CW, Chen PL, Yeh E, Leung L, et al. Belun Ring (Belun Sleep System BLS-100): deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea. Sleep Health. 2023;9(4):430-40.

Retamales G, Gavidia ME, Bausch B, Montanari AN, Husch A, Goncalves J. Towards automatic home-based sleep apnea estimation using deep learning. NPJ Digit Med. 2024;7(1):144.

Chen KW, Tseng CH, Lee HC, Liu WT, Chou KT, Wu HT. Validation of a fingertip home sleep apnea testing system using deep learning AI and a temporal event localization analysis. Sleep. 2025;48(5):zsae317.

Phan H, Mikkelsen K, Chén OY, Koch P, Mertins A, De Vos M. Sleeptransformer: automatic sleep staging with interpretability and uncertainty quantification. IEEE Trans Biomed Eng. 2022;69(8):2456-67.

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.

Fiorillo L, Favaro P, Faraci FD. Deepsleepnet-lite: a simplified automatic sleep stage scoring model with uncertainty estimates. IEEE Trans Neural Syst Rehabil Eng. 2021;29:2076-85.

Guillot A, Thorey V. RobustSleepNet: transfer learning for automated sleep staging at scale. IEEE Trans Neural Syst Rehabil Eng. 2021;29:1441-51.

Leino A, Korkalainen H, Kalevo L, Nikkonen S, Kainulainen S, Ryan A, et al. Deep learning enables accurate automatic sleep staging based on ambulatory forehead EEG. IEEE Access. 2022;10:26554-66.

Jeong J, Yoon W, Lee JG, Kim D, Woo Y, Kim DK, et al. Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification. Sleep. 2023;46(12):zsad242.

Seo H, Back S, Lee S, Park D, Kim T, Lee K. Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. Biomed Signal Process Control. 2020;61:102037.

Sun C, Fan J, Chen C, Li W, Chen W. A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation. IEEE Access. 2019;7:109386-97.

Vaquerizo-Villar F, Gutiérrez-Tobal GC, Calvo E, Álvarez D, Kheirandish-Gozal L, Del Campo F, et al. An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea. Comput Biol Med. 2023;165:107419.

Levy J, Álvarez D, Del Campo F, Behar JA. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. Nat Commun. 2023;14(1):4881.