

Human-Centered AI for Sleep EEG: A Systematic Review of Translational and Cross-Species Evidence

Cole Garrison¹, Zeina Nweashe¹, and Xiaodong Qu¹

Department of Computer Science,
George Washington University, Washington, DC, USA
(cole.garrison, zeina.nweashe, x.qu)@gwu.edu

Abstract. Human-centered artificial intelligence (AI) is increasingly applied to electroencephalography (EEG) for sleep staging and sleep disorder detection, offering scalable screening and automated monitoring. In parallel, mouse EEG studies provide mechanistic insight into sleep regulation under controlled experimental conditions. Despite rapid advances in both domains, synthesis across species and integration of human-centered evaluation remain limited.

This PRISMA-guided systematic review examines 42 studies of AI-based sleep EEG across human and mouse research, with an explicit human-centered AI perspective. We synthesize (1) effectiveness evidence and validation practices, (2) biomarkers and model classes used for prediction and monitoring, and (3) translational connections between mechanistic and clinical contexts. Reported performance is generally strong within task-specific settings (e.g., 75–85% accuracy for human sleep staging; AUC 0.80–0.92 for apnea detection), yet fewer than half of studies report external validation, and formal user evaluations are rare (fewer than five studies).

Across species, hybrid approaches combining physiologically grounded biomarkers with deep learning representations are increasingly common. However, interaction validity—including interpretability, trust calibration, and workflow integration—is inconsistently assessed. We argue that advancing AI-assisted sleep EEG from high-performing models to clinically deployable systems requires alignment across mechanistic, representational, clinical, and interaction layers of evaluation.

Keywords: Human-centered AI · EEG · Sleep EEG · Clinical workflow · Interpretability · Translational research · Clinical decision support

1 Introduction

Sleep disorders affect a substantial proportion of the global population and are associated with significant health, cognitive, and societal consequences. Electroencephalography (EEG) remains a central modality for assessing sleep architecture and detecting abnormalities because it provides a direct measure of

neural dynamics across sleep stages. Recent survey and systematic review work has further documented rapid growth in deep learning, transformer-based, and hybrid EEG modeling approaches, alongside persistent challenges in comparability and reporting across studies [26,34]. Recent advances in artificial intelligence (AI), particularly machine learning and deep learning, have enabled automated analysis of EEG signals for sleep staging, disorder detection, and physiological characterization.

Despite rapid methodological progress, the literature remains fragmented across research domains. Clinical studies primarily focus on diagnostic performance and deployment-oriented evaluation in human populations, while animal research—especially mouse studies—emphasizes mechanistic insight, controlled experimentation, and biomarker discovery. Although both domains contribute valuable knowledge, their findings are rarely synthesized within a unified framework, limiting the ability to translate physiological insights into clinically actionable AI systems.

Related human-in-the-loop EEG systems and real-time personalization studies also suggest that user interaction constraints and feedback mechanisms can substantially shape deployment feasibility [4,39].

At the same time, growing interest in human-centered clinical AI highlights that performance alone is insufficient for safe adoption. Interpretability, workflow integration, and trust calibration play critical roles in determining whether AI-assisted EEG tools can be effectively incorporated into clinical practice. However, these considerations are often discussed separately from methodological and translational analyses, creating a gap between technical development and real-world deployment.

Our framing is also informed by prior EEG-centered reviews that emphasize how taxonomy choices and evaluation protocols shape what conclusions can be drawn from heterogeneous literature [26,59].

This review addresses these gaps by providing a cross-species synthesis of AI-based EEG research for sleep analysis, integrating methodological trends, biomarker design, human-centered considerations, and translational implications. By jointly examining human and mouse studies, we aim to identify common patterns, differences in evaluation practices, and opportunities for bridging mechanistic understanding with clinical application.

Research Questions (RQs).

1. **RQ1:** What performance patterns and validation practices characterize AI-based EEG methods for sleep disorder detection across human and mouse studies?
2. **RQ2:** What biomarkers, feature representations, and AI model classes are most commonly used for prediction and monitoring, and how are they distributed across species?
3. **RQ3:** What human–AI interaction considerations (e.g., interpretability, trust calibration, workflow integration) are reported in this literature, and how do they shape the deployment of AI-assisted EEG systems?

4. **RQ4:** How can insights from mouse studies inform human clinical applications, and what design guidelines emerge for human-centered translational EEG AI?

Contributions. This review makes three primary contributions. First, it provides a structured comparison of methodological and evaluation practices across human and mouse EEG studies, highlighting differences in performance reporting and validation strategies. Second, it synthesizes the landscape of biomarkers, feature representations, and modeling approaches used in sleep EEG analysis, identifying emerging trends in representation learning and interpretability. Third, it integrates translational and human-centered perspectives, outlining design considerations for AI-assisted EEG systems that bridge mechanistic research and clinical deployment.

Together, these contributions aim to support the development of AI methods that are not only technically robust but also physiologically grounded, clinically relevant, and aligned with real-world decision-making contexts.

This framing motivates examination of alignment across mechanistic, representational, clinical, and interaction dimensions.

2 Related Work

Prior research on AI-assisted EEG analysis for sleep disorders spans several intersecting themes, including algorithmic performance evaluation, biomarker and feature development, translational neuroscience, and human-centered deployment considerations. Existing reviews provide valuable foundations, yet they often focus on a single species, disorder, or methodological perspective. This section synthesizes these strands to position the present review and highlight gaps motivating a cross-species, human-centered analysis.

2.1 AI-Based EEG Analysis and Performance Evaluation

A substantial body of literature has examined machine learning and deep learning methods for automated sleep staging, event detection, and disorder classification using EEG and polysomnography signals. Broad surveys of computational sleep analysis outline the evolution of algorithmic approaches and emphasize variability in benchmarking practices and evaluation protocols [12]. Complementary signal-processing reviews detail classical EEG feature families, including spectral, time-frequency, and entropy-based representations, which remain important baselines and often serve as inputs to hybrid machine learning pipelines [33].

More recent synthesis efforts provide comprehensive overviews of automated sleep analysis pipelines, summarizing commonly used datasets, modeling approaches, and evaluation metrics across a large body of studies [60]. Earlier deep learning-focused reviews document the field’s transition toward end-to-end representation learning and highlight how architectural advances have shaped performance improvements in sleep staging and disorder detection [8]. Complementary mapping-style reviews have similarly emphasized the rapid diversification

of EEG model classes and the need for clearer reporting of evaluation design choices [34]. The most recent systematic reviews continue to report heterogeneity in datasets, validation protocols, and reporting standards, underscoring the need for more structured comparisons of performance patterns across studies [57].

Recent work also highlights the expanding ecosystem of EEG-based sleep technologies beyond traditional laboratory PSG. Related systematic reviews of consumer-grade EEG in interactive settings further highlight practical constraints (e.g., usability, calibration, and real-time feedback) that shape deployment beyond algorithmic accuracy alone [59]. Systematic reviews of wearable EEG systems report increasing feasibility, high user compliance, and promising accuracy for sleep staging, suggesting a shift toward scalable and home-based monitoring solutions [14]. Complementing these reviews, empirical benchmarking studies comparing multiple machine learning classifiers demonstrate that optimized feature engineering and ensemble methods can achieve high diagnostic performance, while also illustrating how results depend on dataset characteristics and experimental design [44].

Additional methodological overviews further emphasize the rapid diversification of architectures and learning paradigms in automated sleep analysis, documenting trends in model design, training strategies, and evaluation practices across recent studies [6,24,28].

Within clinical sleep disorder detection, obstructive sleep apnea has received particular attention. Reviews of diagnostic sensors and algorithms situate EEG-based methods within multimodal monitoring ecosystems and highlight ongoing challenges in dataset heterogeneity, labeling practices, and cross-study comparability [31,49]. More recent systematic reviews further categorize datasets, modeling approaches, and evaluation metrics, underscoring persistent issues such as subject-wise generalization and incomplete reporting of validation strategies [13]. Collectively, these observations motivate closer examination of performance patterns and validation practices across both human and animal studies.

2.2 Biomarkers, Feature Representations, and Modeling Trends

At the methodological level, both classical machine learning and deep neural architectures have been applied to EEG for sleep staging and disorder detection. Representative work includes automated sleep stage classification pipelines evaluated on clinical datasets [9]. Studies on apnea classification demonstrate that engineered frequency-band features can remain competitive under certain conditions, suggesting that performance gains may arise from both representation learning and preprocessing choices [53,63].

Recent empirical and quantitative studies further explore the role of feature engineering and physiological biomarkers in sleep analysis. Comparative investigations of EEG feature extraction strategies highlight the contributions of spectral, statistical, and nonlinear features to classification performance [50]. Quantitative EEG analyses provide additional insight into physiologically interpretable markers associated with sleep stages and transitions [19]. Meanwhile, studies on

channel selection demonstrate that reduced-electrode configurations can maintain competitive performance while improving practicality for real-world monitoring scenarios [29].

Recent reviews and methodological syntheses further emphasize the diversity of feature representations and their interaction with modeling approaches. Emerging surveys document how spectral, time-frequency, and nonlinear biomarkers remain foundational descriptors of sleep dynamics, while hybrid feature pipelines increasingly combine physiologically interpretable features with data-driven learning [48]. Automated analysis frameworks highlight the growing role of end-to-end feature learning in extracting discriminative EEG signatures for sleep tasks [10]. Comparative evaluations of automated EEG pipelines further underscore how performance depends on the alignment between feature representation, preprocessing strategy, and task definition [62].

Systematic mapping studies further highlight the growing dominance of deep learning approaches and provide taxonomies linking EEG biomarkers, neural architectures, and diagnostic tasks, while also identifying gaps in explainability and cross-disorder generalization [46].

Recent advances in deep learning have further expanded the landscape of representation learning for sleep EEG analysis, consistent with hybrid EEG modeling trends reported in recent transformer-focused reviews [65,26]. Hybrid architectures combining convolutional and transformer components have been proposed to jointly model local signal characteristics and long-range temporal dependencies, demonstrating improved performance for both sleep staging and disorder prediction tasks [54]. Other work continues to refine convolutional network designs, highlighting how architectural choices such as receptive field configuration and hierarchical feature extraction influence classification accuracy and robustness [32]. At the same time, lightweight and efficiency-oriented deep models are emerging to support scalable and real-world monitoring scenarios, reflecting a growing emphasis on deployability alongside performance [21].

Recent work has also expanded the scope of physiologically grounded biomarkers and feature representations. Reviews of sleep EEG microstructure emphasize that traditional severity indices such as the apnea–hypopnea index may not fully capture clinically relevant neural dynamics, motivating the use of microstructural markers including sleep spindles, K-complexes, cyclic alternating pattern, and quantitative REM signatures [16]. Connectivity-based feature extraction approaches demonstrate how inter-channel relationships and network-level representations can reveal disorder-related patterns [45]. Nonlinear and geometry-aware approaches further extend this space, with topological data analysis providing complementary frameworks for capturing complex temporal structure in EEG signals [47].

Complementing representation-focused research, several studies provide systematic comparisons between traditional machine learning and deep learning paradigms for sleep EEG analysis [2,5,51]. These comparisons emphasize that methodological choice depends on deployment constraints such as data volume, computational resources, and the need for explainable outputs.

In addition, attention-based CNN variants have been used to capture both averaged voltage patterns and localized temporal changes in EEG recordings, further reflecting the broader shift toward attention mechanisms in EEG modeling [61]. Across studies, commonly reported biomarkers include spectral power distributions, time-frequency signatures, entropy-based measures, and event-related features such as micro-arousals. More recent work increasingly leverages learned representations from convolutional, recurrent, and transformer-based architectures, reflecting a shift toward end-to-end modeling. However, the relative prevalence of these approaches and their distribution across species remain undersynthesized, highlighting the need for a systematic taxonomy of biomarkers and model classes.

2.3 Translational Perspectives in Cross-Species Sleep Research

Mouse EEG research has historically focused on mechanistic insight and controlled experimentation, enabling causal inference unavailable in many clinical studies. Translational sleep science frameworks emphasize the importance of integrating preclinical models with human observational research to identify biomarkers that are both physiologically grounded and clinically meaningful [15].

Recent work further highlights methodological challenges in bridging experimental and clinical domains. Comparative discussions of sleep measurement and benchmarking practices underscore differences in acquisition protocols, scoring criteria, and evaluation paradigms across species, which complicate direct comparison and model transfer [25]. Emerging conceptual syntheses also highlight how translational success depends on aligning experimental designs, data representations, and clinical endpoints across research settings [17].

At a mechanistic level, studies of neural representations and circuit-level dynamics provide insight into how sleep-related activity patterns generalize across species [3,18,22]. These perspectives suggest that translational progress may depend on identifying shared physiological mechanisms while accounting for species-specific differences in sleep architecture and recording methodologies. Despite this potential, explicit integration of cross-species evidence in AI-driven EEG research remains limited, motivating the need for systematic synthesis.

2.4 Human-Centered Considerations in Clinical AI

Beyond technical performance, the deployment of AI-assisted EEG systems depends on how model outputs are integrated into clinical workflows. Human-AI interaction research emphasizes interpretability, transparency, and trust calibration as key determinants of safe and effective adoption [7].

Empirical studies of AI-assisted sleep staging systems show that explanation interfaces grounded in clinician workflows can improve both user performance and confidence [20]. Conceptual frameworks further emphasize that human-AI collaboration evolves through feedback and adaptation, framing clinical AI as a socio-technical system rather than a static tool [35]. Systematic perspectives

on explainable AI in healthcare identify gaps in evaluation practices and call for standardized reporting of explanation quality and usability [1].

Workflow-oriented analyses outline how attribution methods and visualization techniques can be integrated into diagnostic pipelines [30], while human-factors studies demonstrate that clinician interaction with AI systems is shaped by usability, cognitive workload, and trust calibration [55]. Governance-focused perspectives further highlight the interplay between interpretability, accountability, and responsibility in clinical decision making [36].

In healthcare contexts, reporting frameworks such as DECIDE-AI highlight the importance of documenting intended users, decision pathways, and evaluation stages for AI-driven decision support [52]. In sleep medicine, these considerations are particularly salient because clinicians must interpret algorithmic outputs alongside other physiological signals and patient histories. However, existing EEG-focused reviews rarely examine whether and how studies address these human-centered deployment factors, indicating an important gap between technical development and real-world use.

2.5 Positioning of This Review

Taken together, prior work provides extensive insight into algorithmic methods, biomarker development, translational neuroscience, and human-centered clinical AI. Yet several gaps remain: existing reviews often focus on either human or animal studies, emphasize technical performance without systematically comparing validation practices, lack unified taxonomies across species, and rarely synthesize human–AI interaction considerations alongside methodological findings.

The present review addresses these gaps by integrating evidence across species, examining methodological and performance patterns, identifying recurring biomarkers and model classes, and synthesizing human-centered implications for deployment. This positioning motivates the RQ-driven synthesis presented in the subsequent sections.

3 Method

3.1 Protocol and Reporting Standard

This systematic literature review was conducted in accordance with the PRISMA 2020 reporting guidelines to ensure transparency and methodological rigor. The review protocol defined the research questions, search strategy, inclusion and exclusion criteria, and data extraction framework prior to full-text screening. The study selection process is summarized in Fig. 1.

3.2 Search Strategy

A structured search strategy was developed to identify peer-reviewed studies applying artificial intelligence (AI) methods to EEG-based analysis of sleep disorders. Searches were conducted between December 2025 and January 2026.

Table 1. Keyword groups and qualitative frequency in search results.

Category	Example keywords	Freq.
AI / ML	machine learning; deep learning; neural network; transformer	H
EEG	electroencephalography; EEG signal; EEG analysis; brain activity	H
Sleep disorders	sleep disorder; apnea; insomnia; narcolepsy; REM sleep	M–H
Cross-species	mouse model; rodent; human study; translational	M
Performance	accuracy; AUC; F1-score; classification; prediction	M
Features	wavelet; spectral; time-frequency; entropy	M–L

Five databases were queried: PubMed, Web of Science, IEEE Xplore, ACM Digital Library, and Google Scholar. Boolean combinations were constructed across three primary concept groups:

- **AI-related terms:** “machine learning,” “deep learning,” “neural network,” “transformer”
- **EEG-related terms:** “electroencephalography,” “EEG signal,” “EEG analysis,” “brain activity”
- **Sleep-related terms:** “sleep disorder,” “sleep apnea,” “insomnia,” “narcolepsy,” “REM sleep”

Search queries were adapted to the syntax of each database. For Google Scholar, results were sorted by relevance and the first 100 records were screened, consistent with prior systematic review practices to balance coverage and feasibility.

Although no strict lower publication year bound was imposed, studies prior to approximately 2015 were rarely eligible due to the focus on modern AI and deep learning approaches. Studies lacking sufficient methodological detail or quantitative evaluation were excluded during screening.

Table 1 summarizes keyword groups and qualitative frequency observed during search and screening.

3.3 Screening and Study Selection

The initial search returned 120 records for title and abstract screening. After removal of duplicate records ($n = 12$), 92 records proceeded to eligibility assessment. Seventy full-text articles were reviewed in detail. A total of 42 studies met all inclusion criteria and were included in the final synthesis, as illustrated in Fig. 1.

Title/abstract screening and full-text eligibility assessment were conducted by a primary reviewer, with a second reviewer independently verifying inclusion decisions at the full-text stage. Disagreements were resolved through discussion until consensus was reached.

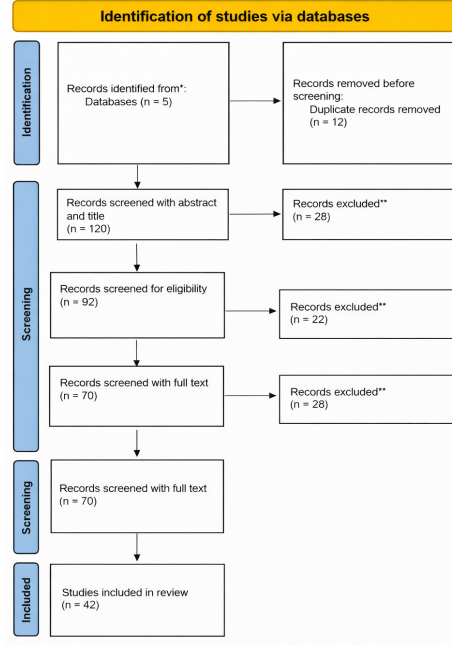


Fig. 1. PRISMA 2020 flow diagram showing identification of studies via databases and screening.

3.4 Inclusion and Exclusion Criteria

Studies were included if they:

- Applied artificial intelligence methods (machine learning, deep learning, or hybrid approaches) to EEG data;
- Focused on sleep disorders (e.g., sleep apnea, insomnia, narcolepsy, parasomnias);
- Used human or mouse subjects with clearly described EEG methodology;
- Were published in peer-reviewed journals or full conference proceedings;
- Reported quantitative performance metrics (e.g., accuracy, F1-score, AUC);
- Were available in full-text English.

Studies were excluded if they relied solely on classical statistical analyses without machine learning components, lacked sufficient EEG methodological detail, or did not provide quantitative evaluation. Earlier studies predating modern AI approaches were excluded when methodological reporting did not meet contemporary standards.

Table 2 summarizes the screening criteria.

Table 2. Inclusion and exclusion criteria applied during the screening process.

Inclusion / Exclusion Criteria	Include?
Applies artificial intelligence (ML, DL, or hybrid models) to EEG data	Yes
Focuses on sleep disorders (e.g., sleep apnea, insomnia, narcolepsy, parasomnias)	Yes
Uses data from human or mouse subjects with clear EEG-based methodology	Yes
Published in a peer-reviewed journal or full conference paper	Yes
Includes quantitative performance evaluation (e.g., accuracy, F1-score, AUC)	Yes
Written in English with full text available	Yes
Non-AI or purely statistical methods without machine learning components	No
Lacking sufficient EEG methodological detail or modern evaluation standards	No

3.5 Data Extraction and Coding

A structured data extraction framework was developed to align with the research questions. The coding schema was piloted on 15 randomly selected eligible studies to refine variable definitions and ensure clarity before full extraction.

Each included study was coded for:

- Species (human, mouse, or cross-species);
- Sleep disorder focus and task type (e.g., staging, apnea detection);
- EEG acquisition characteristics (e.g., channel configuration, sampling rate when reported);
- Feature representation category (spectral, time-frequency, entropy, statistical, learned representation, etc.);
- AI model class (e.g., classical ML, CNN, RNN, Transformer, hybrid);
- Validation strategy (e.g., cross-validation, subject-wise split, external test set);
- Performance metrics and reported ranges;
- Translational framing (mechanistic discovery, clinical validation, cross-species mapping);
- Human-centered deployment signals (e.g., interpretability mechanisms, explanation interfaces, workflow integration, user evaluation).

Extraction was conducted by a primary reviewer and verified by a second reviewer. Discrepancies were resolved through discussion until consensus was achieved.

3.6 Synthesis Approach

Given heterogeneity across datasets, tasks, and evaluation protocols, a narrative synthesis combined with structured descriptive aggregation was performed rather than a pooled meta-analysis.

For RQ1, reported performance metrics (e.g., accuracy, AUC, F1-score) were grouped by task type and species, and performance ranges were summarized to characterize patterns in effectiveness and validation practices.

For RQ2, feature representations and AI model classes were categorized into a unified taxonomy to identify dominant methodological trends across species.

For RQ3, studies were analyzed for explicit reporting of human–AI interaction considerations. Approximately 25–50% of included studies referenced interpretability, explanation, or workflow integration, though formal user studies were comparatively rare.

For RQ4, translational evidence was mapped across mechanistic, representation-level, clinical validation, and deployment-oriented stages to identify recurring challenges and design implications.

This cross-sectional synthesis enables comparison of methodological trends, evaluation rigor, and human-centered considerations across human and mouse EEG research without assuming statistical homogeneity.

4 Results

4.1 Overview of Included Studies

A total of 42 studies met inclusion criteria, including 33 human-focused studies and 9 mouse-based studies. No included study directly used shared human and mouse datasets within a single empirical framework; translational insights were primarily conceptual or representation-level rather than dataset-level integration.

Human studies predominantly addressed diagnostic tasks such as sleep staging and sleep apnea detection, while mouse studies emphasized automated state scoring and mechanistic biomarker discovery under controlled laboratory conditions.

4.2 RQ1: Effectiveness of AI-Based EEG Methods Across Species

Performance Patterns in Human Studies Human studies most frequently evaluated sleep staging and sleep apnea detection. Reported performance ranges varied by task type:

Performance was generally strong within task-specific settings. However, validation practices varied considerably. Approximately 20–40% of studies reported evaluation on external or independent test sets, while fewer than 20% explicitly employed strict subject-wise cross-validation protocols. This heterogeneity limits direct cross-study comparability and highlights ongoing challenges in generalization assessment.

Table 3. Representative performance ranges reported across included studies.

Task Type	Metric	Typical Range
Sleep staging (human)	Accuracy	75%–85%
Apnea detection (human)	AUC	0.80–0.92
Mouse state classification	Accuracy	85%–95%

Table 4. Approximate distribution of AI model classes across included studies.

Model Class	Approximate Prevalence
CNN-based models	30–50%
Transformer / hybrid deep models	>30%
Classical ML (SVM, RF, etc.)	20–40%

Model Distribution Across included studies, deep learning architectures were dominant, particularly convolutional neural networks (CNNs), which appeared in approximately 30–50% of studies. Transformer-based or hybrid architectures accounted for more than 30% of recent studies, reflecting growing interest in modeling long-range temporal dependencies. Classical machine learning methods (e.g., SVM, random forest) appeared in approximately 20–40% of studies, often in combination with engineered features.

Importantly, performance differences were not attributable solely to model architecture. Feature engineering, preprocessing pipelines, and dataset composition were frequently cited as influential factors in reported outcomes.

Comparative Interpretation Across Species Mouse studies emphasized reproducibility and physiological alignment with expert scoring rather than clinical utility metrics. Reported accuracy ranges (85–95%) were typically evaluated under controlled laboratory conditions. Differences in EEG montage, sampling rate, stage definitions, and experimental objectives complicate absolute cross-species comparison. Therefore, effectiveness is best interpreted as domain-specific robustness rather than directly comparable numerical superiority.

4.3 RQ2: Biomarkers and AI Techniques for Prediction and Monitoring

Feature representations spanned both physiologically grounded biomarkers and data-driven learned representations. Spectral and time-frequency features were the most consistently reported engineered features across both species. Entropy-based and nonlinear complexity measures appeared less frequently but were often used in sleep quality or microstructure analysis contexts.

Learned representations derived from CNN, RNN, and transformer architectures were increasingly common, particularly in human diagnostic studies.

Table 5. Taxonomy of EEG biomarkers and feature representations identified across included studies.

Category	Description	Common Context
Spectral features	Frequency-band power (delta, theta, alpha, beta)	Staging, apnea detection
Time-frequency features	Wavelet/STFT patterns	Transition detection
Entropy / nonlinear	Signal complexity measures	Sleep quality analysis
Statistical features	Amplitude and variance summaries	Baseline modeling
Physiological biomarkers	qEEG indices, stage signatures	Interpretable diagnostics
Learned representations	CNN/RNN/Transformer embeddings	End-to-end prediction
Channel optimization	Reduced electrode setups	Wearable monitoring

Table 6. Reporting frequency of human-centered AI considerations.

Human-AI Reporting Element	Approximate Count (n=42)
Interpretability / explanation discussed	10–15
Formal user study or clinician evaluation	<5
Workflow integration discussion	10–15

Hybrid approaches combining interpretable features with deep learning were frequently described as balancing predictive accuracy with clinical interpretability.

Table 5 summarizes the unified taxonomy of feature categories across included studies.

Across studies, convergence toward hybrid modeling strategies was observed, particularly in human clinical contexts where interpretability was valued alongside predictive performance.

4.4 RQ3: Human–AI Interaction Considerations in Clinical EEG Systems

Human-centered reporting was less prevalent than technical performance reporting. Approximately 10–15 of the 42 included studies explicitly discussed interpretability, explanation interfaces, or workflow integration. Fewer than five studies reported formal user evaluations or clinician-in-the-loop experiments.

Most human-centered discussion focused on explanation design, visualization clarity, and trust calibration. Structured evaluation of cognitive workload, accountability mechanisms, or longitudinal adoption remained comparatively rare.

4.5 RQ4: Translational Insights Across Species

Although no included study directly integrated human and mouse EEG datasets within a single empirical framework, several studies described representation-

level or biomarker-level parallels across species. Translational pathways typically followed staged progression from mechanistic discovery (animal models) to representation alignment and clinical validation.

Barriers to translation included heterogeneity in acquisition protocols, labeling conventions, and sampling rates. Nonetheless, the literature suggests that physiologically grounded biomarkers identified in animal research can inform hypothesis generation and representation design in human diagnostic models.

Overall, translational integration remains conceptual and methodological rather than dataset-level, underscoring the need for standardized cross-species evaluation frameworks and harmonized reporting practices.

5 Discussion

This discussion synthesizes findings across research questions and introduces a conceptual framing to clarify how cross-species evidence, methodological rigor, and human-centered deployment intersect in AI-assisted EEG systems for sleep research.

5.1 A Translational Alignment Model for Human-Centered EEG AI

The results suggest that translational progress in AI-based EEG analysis depends on alignment across multiple layers of evidence and evaluation. We therefore propose a *Translational Alignment Model* comprising four interdependent forms of validity:

1. **Mechanistic validity:** Alignment between EEG features and underlying neurophysiological processes, often derived from controlled animal studies.
2. **Representational validity:** Stability and transferability of learned or engineered feature representations across datasets and contexts.
3. **Clinical validity:** Robust performance under clinically realistic validation protocols, including subject-wise splits and external test sets.
4. **Interaction validity:** Safe and effective integration into human workflows, including interpretability, trust calibration, and usability.

Current literature demonstrates strength in mechanistic and task-specific clinical validity but reveals variability in representational generalization and limited systematic evaluation of interaction validity. Translational success therefore requires not only accurate models but alignment across these four layers. Related human-in-the-loop EEG systems and real-time personalization studies further illustrate how interaction constraints, calibration procedures, and feedback loops influence practical deployment feasibility beyond model accuracy alone [4,23,27,39,56,58,64].

5.2 Reframing Performance Evaluation Beyond Accuracy (RQ1)

While reported performance ranges indicate that AI-based EEG methods are mature and effective within defined tasks, evaluation practices remain heterogeneous. Fewer than half of studies reported external validation, and strict subject-wise cross-validation was comparatively uncommon.

This pattern suggests that performance metrics alone are insufficient indicators of translational readiness. Accuracy, AUC, and F1-score reflect task-level discrimination, but they do not necessarily capture robustness under distributional shifts or real-world deployment constraints.

We argue that future evaluation frameworks should explicitly incorporate:

- Clear reporting of validation splits (record-wise vs subject-wise),
- External dataset testing when feasible,
- Sensitivity analyses under varied acquisition conditions,
- Reporting of uncertainty estimates.

Such practices would strengthen representational and clinical validity within the Translational Alignment Model. More broadly, recent work on human-centered AI evaluation emphasizes that deployment readiness depends on aligning technical validation with structured interaction-level assessment and accountability considerations [11,41,40,42,37,38,43].

5.3 Hybrid Representation Design and Interpretability (RQ2)

The convergence toward hybrid modeling strategies reflects an implicit recognition that performance and interpretability are not mutually exclusive goals. Spectral and time-frequency biomarkers provide physiologically interpretable anchors, while learned representations enhance discriminative capacity. Recent transformer-based and hybrid EEG architectures similarly reflect this dual emphasis, combining convolutional feature extraction with attention mechanisms to balance spatial-temporal modeling and interpretability [26,61,65].

Translational opportunities lie in deliberately coupling interpretable biomarkers with data-driven representations. This approach may support both mechanistic validity and interaction validity, particularly in clinical settings where explainability influences adoption.

More consistent reporting of preprocessing pipelines, feature extraction methods, and representation choices would further enable reproducibility and cross-study synthesis.

5.4 Interaction Validity as a Core Evaluation Dimension (RQ3)

Approximately one-quarter to one-third of included studies discussed interpretability or workflow integration, yet fewer than five reported formal user evaluations. This imbalance suggests that interaction validity remains under-assessed relative to predictive performance.

We contend that interaction quality should be considered a primary evaluation dimension rather than an auxiliary consideration. In clinical EEG workflows, AI systems function as decision-support tools embedded within socio-technical environments. Effective deployment depends on:

- Explanation mechanisms aligned with clinician reasoning,
- Transparent communication of model uncertainty,
- Mechanisms for verification, override, or feedback,
- Cognitive workload considerations.

Without systematic evaluation of these dimensions, high-performing models may face barriers to safe adoption.

5.5 Translational Opportunities and Structural Challenges (RQ4)

Although no included study directly integrated shared human and mouse EEG datasets within a unified empirical framework, representation-level parallels were frequently discussed. This suggests that cross-species translation currently operates at the level of biomarker interpretation and architectural design rather than dataset integration.

We view this not as a failure but as an opportunity. Future progress may depend on developing harmonized annotation protocols, standardized EEG ontologies, and shared benchmarking frameworks capable of bridging mechanistic and clinical contexts.

At a field level, establishing minimal reporting standards for cross-species representation mapping could accelerate translational alignment and reduce fragmentation across research domains.

5.6 Design Recommendations for Human-Centered EEG AI

Based on synthesized findings, Table 7 outlines design and reporting recommendations to support translational alignment and human-centered deployment.

These recommendations operationalize interaction validity and provide concrete steps for strengthening translational alignment.

5.7 Societal and Practical Considerations

AI-assisted EEG systems have potential to expand access to sleep diagnostics through wearable and home-based monitoring. However, risks include dataset bias, uneven performance across demographic groups, and challenges in accountability.

Addressing these concerns requires both technical robustness and structured human-centered evaluation to ensure equitable and responsible deployment.

Table 7. Design and reporting recommendations for human-centered EEG AI.

Dimension	Recommendation
Validation transparency	Clearly report subject-wise vs record-wise splits and use external test sets when possible
Representation reporting	Document preprocessing pipelines and feature categories explicitly
Interpretability	Describe explanation mechanisms and their physiological grounding
Workflow integration	Specify intended user roles and decision pathways
Uncertainty communication	Report confidence estimates or reliability indicators
User evaluation	Incorporate clinician or end-user testing where feasible

5.8 Limitations and Methodological Reflexivity

This review is constrained by heterogeneity in datasets, task definitions, and reporting standards, which precluded formal meta-analysis. Although screening and coding were conducted with verification and consensus procedures, interpretive synthesis necessarily involves subjective judgment.

Limiting Google Scholar screening to the first 100 relevance-ranked results may have excluded some peripheral studies. In addition, restricting inclusion to English-language publications introduces potential language bias. These considerations underscore the importance of transparent reporting and ongoing refinement of systematic review methodologies in rapidly evolving AI domains.

5.9 Future Directions

Future research should prioritize:

- Cross-dataset benchmarking and shared annotation standards,
- Development of representation-level alignment frameworks across species,
- Integration of usability and clinician-in-the-loop evaluation into technical validation,
- Longitudinal studies examining trust calibration and workflow adaptation,
- Harmonized reporting guidelines that incorporate interaction validity alongside predictive performance.

Advancing AI-based EEG systems from high-performing models to human-centered clinical tools will require coordinated progress across mechanistic, representational, clinical, and interaction layers.

6 Conclusion

We conducted a PRISMA-guided systematic review of AI-based EEG methods for sleep disorder research across human and mouse studies, with an explicit human–AI interaction perspective.

Answer to RQ1: AI-based EEG approaches generally report strong performance for sleep staging and disorder detection tasks. However, comparisons across studies remain constrained by heterogeneous validation strategies, limited external testing, and inconsistent reporting of generalization protocols.

Answer to RQ2: Both physiologically grounded biomarkers (e.g., spectral and time-frequency features) and learned representations are widely used across species. Recent work increasingly adopts hybrid modeling strategies that balance predictive accuracy with interpretability and clinical feasibility.

Answer to RQ3: Human-centered considerations are reported less consistently than technical metrics. While interpretability and workflow integration are increasingly recognized, formal user evaluations and structured interaction assessments remain relatively rare.

Answer to RQ4: Translational connections between mouse and human EEG research are primarily representation- and biomarker-level rather than dataset-level, highlighting the need for harmonized reporting practices and cross-domain evaluation frameworks.

Overall, the evidence suggests that advancing AI-assisted EEG systems from high-performing models to clinically deployable tools requires integrating mechanistic insight, robust validation practices, and explicit human-centered design and evaluation strategies.

References

1. Abbas, Q., Jeong, W., Lee, S.W.: Explainable ai in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. In: *Healthcare*. vol. 13, p. 2154. MDPI (2025)
2. Abibullaev, B., Keutayeva, A., Zollanvari, A.: Deep learning in eeg-based bcis: A comprehensive review of transformer models, advantages, challenges, and applications. *IEEE Access* **11**, 127271–127301 (2023)
3. Ahmad, J., Ellis, C., Leech, R., Voytek, B., Garces, P., Jones, E., Buitelaar, J., Loth, E., Dos Santos, F.P., Amil, A.F., et al.: From mechanisms to markers: novel noninvasive eeg proxy markers of the neural excitation and inhibition system in humans. *Translational psychiatry* **12**(1), 467 (2022)
4. Aleisa, N., Lu, Y., Gao, L., Nweashe, Z., Flanagan, E., Winchester, M., Ampomah, S.J., Qu, X.: Accessible eeg game control: Real-time personalization with consumer-grade eeg hardware. In: *International Conference on Human-Computer Interaction*. pp. 120–135 (2026)
5. Almutairi, H., Hassan, G.M., Datta, A.: Machine-learning-based-approaches for sleep stage classification utilising a combination of physiological signals: a systematic review. *Applied Sciences* **13**(24), 13280 (2023)
6. Alshammari, T.S.: Applying machine learning algorithms for the classification of sleep disorders. *IEEE Access* **12**, 36110–36121 (2024)
7. Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., et al.: Guidelines for human-ai interaction. In: *Proceedings of the 2019 chi conference on human factors in computing systems*. pp. 1–13 (2019)
8. de Bardeci, M., Ip, C.T., Olbrich, S.: Deep learning applied to electroencephalogram data in mental disorders: A systematic review. *Biological Psychology* **162**, 108117 (2021)
9. Chen, T., Huang, H., Pan, J., Li, Y.: An eeg-based brain-computer interface for automatic sleep stage classification. In: *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. pp. 1988–1991. IEEE (2018)
10. Dequidt, P., Seraphim, M., Lechervy, A., Gaez, I.I., Brun, L., Etard, O.: Automatic sleep stage classification on eeg signals using time-frequency representation. In: *International Conference on Artificial Intelligence in Medicine*. pp. 250–259. Springer (2023)
11. Dou, G., Zhou, Z., Qu, X.: Time majority voting, a pc-based eeg classifier for non-expert users. In: *International Conference on Human-Computer Interaction*. pp. 415–428. Springer (2022)
12. Fallmann, S., Chen, L.: Computational sleep behavior analysis: A survey. *IEEE Access* **7**, 142421–142440 (2019)
13. Fathima, S., Ahmed, M.: Sleep apnea detection using eeg: a systematic review of datasets, methods, challenges, and future directions. *Annals of biomedical engineering* **53**(5), 1043–1067 (2025)
14. de Gans, C.J., Burger, P., Van den Ende, E., Hermanides, J., Nanayakkara, P., Gemke, R., Rutters, F., Stenvers, D.: Sleep assessment using eeg-based wearables—a systematic review. *Sleep Medicine Reviews* **76**, 101951 (2024)
15. Grandner, M.A., Fernandez, F.X.: The translational neuroscience of sleep: a contextual framework. *Science* **374**(6567), 568–573 (2021)
16. Gu, Y., Gagnon, J.F., Kaminska, M.: Sleep electroencephalography biomarkers of cognition in obstructive sleep apnea. *Journal of Sleep Research* **32**(5), e13831 (2023)

17. Hermann, D.M., Bacigaluppi, M., Bassetti, C.L., Bassotti, G., Boltze, J., Chan, A., Dalkara, T., Denes, A., Diez-Tejedor, E., Dodel, R., et al.: Most prominent challenges in translational neuroscience and strategic solutions to bridge the gaps: Perspectives from an editorial board interrogation. *Exploration of neuroscience* **4**, 1006106 (2025)
18. Herzog, L.E., Wang, L., Yu, E., Choi, S., Farsi, Z., Song, B.J., Pan, J.Q., Sheng, M.: Mouse mutants in schizophrenia risk genes *grin2a* and *akap11* show eeg abnormalities in common with schizophrenia patients. *Translational Psychiatry* **13**(1), 92 (2023)
19. Hussain, I., Hossain, M.A., Jany, R., Bari, M.A., Uddin, M., Kamal, A.R.M., Ku, Y., Kim, J.S.: Quantitative evaluation of eeg-biomarkers for prediction of sleep stages. *Sensors* **22**(8), 3079 (2022)
20. Hwang, J., Lee, T., Lee, H., Byun, S.: A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study. *Journal of medical Internet research* **24**(1), e28659 (2022)
21. Ito, A., Tanaka, T.: Sleepsatelightftc: A lightweight and interpretable deep learning model for single-channel eeg-based sleep stage classification. *IEEE Access* (2025)
22. Keinath, A.T., Mosser, C.A., Brandon, M.P.: The representation of context in mouse hippocampus is preserved despite neural drift. *Nature communications* **13**(1), 2415 (2022)
23. Key, M.L., Mehtiyev, T., Qu, X.: Advancing eeg-based gaze prediction using depth-wise separable convolution and enhanced pre-processing. In: *International Conference on Human-Computer Interaction*. pp. 3–17. Springer (2024)
24. Khanmohmmadi, S., Khatibi, T., Tajeddin, G., Akhondzadeh, E., Shojaei, A.: Revolutionizing sleep disorder diagnosis: A multi-task learning approach optimized with genetic and q-learning techniques. *Scientific Reports* **15**(1), 16603 (2025)
25. Klerman, E.B., Kramer, A., Zee, P.C.: From bench to bedside and back again: translating circadian science to medicine. *Journal of biological rhythms* **38**(2), 125–130 (2023)
26. Konkar, A., Qu, X.: A review of transformer-based and hybrid deep learning approaches for eeg analysis. In: *International Conference on Human-Computer Interaction*. pp. 391–404. Springer (2026)
27. Li, W., Zhou, N., Qu, X.: Enhancing eye-tracking performance through multi-task learning transformer. In: *International Conference on Human-Computer Interaction*. pp. 31–46. Springer (2024)
28. Liu, P., Qian, W., Zhang, H., Zhu, Y., Hong, Q., Li, Q., Yao, Y.: Automatic sleep stage classification using deep learning: signals, data representation, and neural networks. *Artificial Intelligence Review* **57**(11), 301 (2024)
29. Lyu, J., Shi, W., Zhang, C., Yeh, C.H.: A novel sleep staging method based on eeg and ecg multimodal features combination. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **31**, 4073–4084 (2023)
30. Maurer, M.C., Metsch, J.M., Hempel, P., Bender, T., Spicher, N., Hauschild, A.C.: Explainable artificial intelligence on biosignals for clinical decision support. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pp. 6597–6604 (2024)
31. Mostafa, S.S., Mendonça, F., G. Ravelo-García, A., Morgado-Dias, F.: A systematic review of detecting sleep apnea using deep learning. *Sensors* **19**(22), 4934 (2019)
32. Mostafaei, S.H., Tanha, J., Sharafkhaneh, A.: A novel deep learning model based on transformer and cross modality attention for classification of sleep stages. *Journal of biomedical informatics* **157**, 104689 (2024)

33. Motamedi-Fakhr, S., Moshrefi-Torbati, M., Hill, M., Hill, C.M., White, P.R.: Signal processing techniques applied to human sleep eeg signals—a review. *Biomedical Signal Processing and Control* **10**, 21–33 (2014)
34. Murungi, N.K., Pham, M.V., Dai, X., Qu, X.: Trends in machine learning and electroencephalogram (eeg): a review for undergraduate researchers. In: *International Conference on Human-Computer Interaction*. pp. 426–443. Springer (2023)
35. Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., Rinzivillo, S.: Co-design of human-centered, explainable ai for clinical decision support. *ACM Transactions on Interactive Intelligent Systems* **13**(4), 1–35 (2023)
36. Panigutti, C., Beretta, A., Giannotti, F., Pedreschi, D.: Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. pp. 1–9 (2022)
37. Qu, X.: Time Continuity Voting for Electroencephalography (EEG) Classification. Ph.D. thesis, Brandeis University (2022)
38. Qu, X., Damoah, A., Sherwood, J., Liu, P., Jin, C.S., Chen, L., Shen, M., Aleisa, N., Hou, Z., Zhang, C., et al.: A comprehensive review of ai agents: Transforming possibilities in technology and beyond. *arXiv preprint arXiv:2508.11957* (2025)
39. Qu, X., Hickey, T.J.: Eeg4home: A human-in-the-loop machine learning model for eeg-based bci. In: *International Conference on Human-Computer Interaction*. pp. 162–172. Springer (2022)
40. Qu, X., Liu, P., Li, Z., Hickey, T.: Multi-class time continuity voting for eeg classification. In: *International Conference on Brain Function Assessment in Learning*. pp. 24–33. Springer (2020)
41. Qu, X., Liukasemsarn, S., Tu, J., Higgins, A., Hickey, T.J., Hall, M.H.: Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on eeg patterns. *Frontiers in psychiatry* **11**, 541659 (2020)
42. Qu, X., Mei, Q., Liu, P., Hickey, T.: Using eeg to distinguish between writing and typing for the same cognitive task. In: *International Conference on Brain Function Assessment in Learning*. pp. 66–74. Springer (2020)
43. Qu, X., Sherwood, J., Liu, P., Aleisa, N.: Generative ai tools in higher education: A meta-analysis of cognitive impact. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. pp. 1–9 (2025)
44. Rahman, M.A., Jahan, I., Islam, M., Jabid, T., Ali, M.S., Rashid, M.R.A., Islam, M.M., Ferdaus, M.H., Rasel, M.M.K., Jahan, M.R., et al.: Improving sleep disorder diagnosis through optimized machine learning approaches. *IEEE Access* **13**, 20989–21004 (2025)
45. Rajeswari, J., Jagannath, M.: Brain connectivity analysis based classification of obstructive sleep apnea using electroencephalogram signals. *Scientific Reports* **14**(1), 5561 (2024)
46. Rivera, M.J., Teruel, M.A., Mate, A., Trujillo, J.: Diagnosis and prognosis of mental disorders by means of eeg and deep learning: a systematic mapping study. *Artificial Intelligence Review* **55**(2), 1209–1251 (2022)
47. Sathyanarayana, A., Manjunath, S., Perea, J.A.: Topological data analysis based characteristics of electroencephalogram signals in children with sleep apnea. *Journal of sleep research* **34**(6), e70017 (2025)
48. Sharma, R., Meena, H.K.: Emerging trends in eeg signal processing: A systematic review. *SN Computer Science* **5**(4), 415 (2024)

49. Shokouinejad, M., Fernandez, C., Carroll, E., Wang, F., Levin, J., Rusk, S., Glat-tard, N., Mulchrone, A., Zhang, X., Xie, A., et al.: Sleep apnea: a review of diag-nostic sensors, algorithms, and therapies. *Physiological measurement* **38**(9), R204–R252 (2017)
50. Siddiqa, H.A., Qureshi, M.F., Khurshid, A., Xu, Y., Wang, L., Abbasi, S.F., Chen, C., Chen, W.: Eeg electrode setup optimization using feature extraction techniques for neonatal sleep state classification. *Frontiers in Computational Neuroscience* **19**, 1506869 (2025)
51. Siddiqui, M.I.H., Sakib, A.H., Akter, S., Debnath, J., Mahmud, M.R.: Comparative analysis of traditional machine learning vs deep learning for sleep stage classifica-tion. *International Journal of Science and Research Archive [Internet]* pp. 1778–89 (2025)
52. Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., et al.: Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai. *bmj* **377** (2022)
53. Vimala, V., Ramar, K., Ettappan, M.: An intelligent sleep apnea classification system based on eeg signals. *Journal of medical systems* **43**(2), 36 (2019)
54. Wan, C., Nnamdi, M.C., Shi, W., Smith, B., Purnell, C., Wang, M.D.: Advancing sleep disorder diagnostics: A transformer-based eeg model for sleep stage classifica-tion and osa prediction. *IEEE Journal of Biomedical and Health Informatics* **29**(2), 878–886 (2024)
55. Wang, L., Zhang, Z., Wang, D., Cao, W., Zhou, X., Zhang, P., Liu, J., Fan, X., Tian, F.: Human-centered design and evaluation of ai-empowered clinical decision support systems: a systematic review. *Frontiers in Computer Science* **5**, 1187299 (2023)
56. Wang, Y., Xie, W., Dou, S., Qu, X.: Emigc: An eeg motor imagery controller for real-time gameplay. In: *International Conference on Human-Computer Interaction*. pp. 63–76. Springer (2025)
57. Wara, T.U., Fahad, A.H., Das, A.S., Shawon, M.M.H.: A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence. *Heliyon* **11**(12) (2025)
58. Xie, W., Dou, J., Wang, Y., Qu, X.: From theory to play: A review of eeg-controlled directional games and evaluation of custom-developed bci games. In: *International Conference on Human-Computer Interaction*. pp. 179–191 (2025)
59. Xie, W., Wang, Y., Dou, S., Qu, X.: A systematic review of consumer-grade eeg applications in directional game control via motor imagery. In: *International Con-ference on Human-Computer Interaction*. pp. 85–105. Springer (2026)
60. Yazdi, M., Samaee, M., Massicotte, D.: A review on automated sleep study. *Annals of Biomedical Engineering* **52**(6), 1463–1491 (2024)
61. Yi, L., Qu, X.: Attention-based cnn capturing eeg recording’s average voltage and local change. In: *International Conference on Human-Computer Interaction*. pp. 448–459. Springer (2022)
62. Zhang, R., Jiang, R., Hu, H., Gao, Y., Xia, W., Song, B.: Automatic sleep staging method using eeg based on stft and residual network. *IEEE Access* **13**, 1778–1789 (2024)
63. Zhao, X., Wang, X., Yang, T., Ji, S., Wang, H., Wang, J., Wang, Y., Wu, Q.: Classification of sleep apnea based on eeg sub-band signal characteristics. *Scientific Reports* **11**(1), 5824 (2021)

64. Zhou, Z., Dou, G., Qu, X.: Brainactivity1: A framework of eeg data collection and machine learning analysis for college students. In: International Conference on Human-Computer Interaction. pp. 119–127. Springer (2022)
65. Zhu, C., Xu, Y., Qu, X.: Gvit: Combining convolutional and transformer layers for spatial-temporal eeg analysis. In: International Conference on Human-Computer Interaction. pp. 433–442. Springer (2025)