Explainable AI for Sleep Disorder Diagnostics

Nirubama Vijayakumar[0009-0004-3850-2564] and N. Arivazhagan[0000-0001-9515-4237]

SRM Institute of Science and Technology, Chengalpattu, 603203, India

nm1203@srmist.edu.in
arivazhn@srmist.edu.in

Abstract. Millions of people in worldwide suffer from sleep problems, which will affect total quality of life. Due to the recent advancements in artificial intelligence (AI), it has introduced innovative diagnostic tools for sleep disorders, the machine learning models are used to analyse complex datasets from wearable devices, polysomnography (PSG), and other sleep-monitoring technologies. But, the "black-box" nature of these AI models limits their adoption in clinical settings because of the absence of interpretability and transparency. These challenges are addressed by Explainable AI (XAI) by giving insights for the decision-making processes of AI models. This review explores the state-of-the-art XAI techniques, their benefits, limitations, and future directions applied to sleep disorder diagnostics. The article also discusses about the integration of XAI with wearable technology and clinical progress to increase diagnostic accuracy and thereby increasing the patient trust.

Keywords: First Keyword, Second Keyword, Third Keyword.

1 Introduction

About 30% of global population are affected by sleep problems such narcolepsy, insomnia, sleep apnea, and restless legs syndrome. To reduce health related problems such as diabetes, mental health issues, and cardiovascular illnesses, it is necessary to diagnosis in early stage. Even while traditional diagnostic methods such as PSG and other conventional diagnostic techniques remains gold standard, many patients cannot afford them, and they need a lot of resources. AI has become a substitute, providing physicians with automated sleep data analysis. By the improvement of AI-based diagnostics' openness, interpretability, and credibility, explainable AI offers answers to these challenges.

Research Paper
DOI: https://doi.org/10.46793/BISEC25.454V
Part of ISBN: 978-86-89755-40-4

2 AI Methods

It is hard to detect sleep disorders automatically because physiological signs might differ greatly. Because of these differences, it is challenging to develop effective sleep problem detection models that assist human specialists in diagnosis and therapy monitoring. Eight sleep problems during the review, including insomnia and sleep apnea were discovered. The detection operation was performed using 24 different methods. These methods have changed over time; before to 2017, only conventional machine learning) was employed. Both machine learning and deep learning techniques were applied to the diagnosis of sleep disorders starting in 2018. Since DL algorithms require far more data for training and testing than ML algorithms, their strong development has major implications for future detection systems. According to the findings of the review, the kind and quantity of labeled data are essential for the development of future sleep problem detection systems as they will influence the selection of the AI algorithm that creates the required level of decision support. More labeled data will aid in representing the differences in symptoms [1].

Because they are viewed as "black boxes," automatic deep-learning models for sleep scoring in kids with obstructive sleep apnea are not widely employed in clinical settings. The goal was to use single-channel electroencephalogram data to create a deep learning model for children's sleep staging that was both accurate and interpretable. EEG signals from a clinical sleep database (n = 980) and the Childhood Adenotonsillectomy Trial dataset (n = 1637) were used. In order to automatically categorize sleep phases from a single-channel EEG data, three different deep-learning architectures were investigated. Then, an explainable artificial intelligence method called Gradient-weighted Class Activation Mapping was used to analyze the individual EEG patterns that contributed to each anticipated sleep stage. For automated sleep stage recognition in the CHAT test set, a conventional neural network outperformed the other studied architectures (accuracy = 86.9% and five-class kappa = 0.827). Additionally, there was a high degree of agreement between the CNN-based prediction of total sleep duration and the clinical dataset (intra-class correlation value = 0.772). The EEG characteristics linked to each sleep state were successfully emphasized by XAI method utilizing Grad-CAM, highlighting their impact on the CNN's decision-making process in both datasets. Therefore, automated sleep staging in pediatric sleep apnea testing may be made possible by using an explainable CNN-based deep-learning model in the clinical setting [2].

Identifying sleep phases is crucial to diagnosing sleep disorders. This categorization has historically been carried out by polysomnography recordings. However, automated methods have been developed to lessen the labor- intensive aspect of this operation. The dataset includes PSG recordings from 2056 participants in the National Sleep Research Resource's multi-ethnic study of atherosclerosis cohort. " A total of 1731 of these participants are categorized as excellent sleepers, whereas 130 have insomnia, 39 have PLM, and 156 have sleep apnea. By applying wavelet-based Hjorth parameters to create

a machine-learning model with explainable artificial intelligence capabilities, this work suggests an automated computerized method for classifying sleep phases. It has been used to extract subbands from 30-second electroencephalogram epochs using an optimum biorthogonal wavelet filter bank . To get the best result, three EEG channels are used: Fz_Cz, Cz_Oz, and C4_M1. Following their extraction from SBs, the Hjorth parameters were then input into several machine learning methods. The model used characteristics from all channels and an ensembled bagged trees (EnBT) classifier for participants with the disorders indicated above. The insomniac, PLM, apniac, excellent sleepers, and full datasets yield the maximum accuracy, which is 86.8%, 87.3%, 85.0%, 84.5%, and 83.8%, respectively. To increase the accuracy of sleep stage categorization and advance knowledge of sleep disorders such insomnia, PLM, and sleep apnea by utilizing these methods and datasets [3].

A considerable percentage of adults suffer from sleep apnea , a common sleep problem. In order to identify apnea and normal events using single lead electrocardiography , instantaneous heart rate, and ECG-derived breathing, the suggested technique coarse-grains a signal at several scales utilizing the well-known multiscale entropy methods. A moderately accurate sleep apnea detection model with an accuracy ranging from 70% to 100% based on two different probabilistic thresholds of 50% and 70%, with a reduction in false positive rate from 28% to 14% for applications in the development of AI-based IoT connected smart wearable devices were used. This model is based on the concept of nonlinear dynamical systems analysis for feature extraction along with machine learning approach. Additional model validation using out-of-distribution datasets is necessary to assess the efficacy of the suggested approach in accurately detecting sleep apnea [4].

A repetitive physiological electroencephalogram activity that occurs in the brain during sleep is known as the cyclic alternating pattern . In order to categorize CAP A & B phases and phases A sub-phases (A1, A2, A3), an automated, computerized method for creating a machine learning model with explainable artificial intelligence capabilities utilizing wavelet-based Hjorth parameters is suggested.To shed light on the model, feature ranking based on SHAP was used. To build the model, single-channel standardized EEG recordings from patients with five different forms of sleep disorders—narcolepsy, nocturnal frontal lobe epilepsy , insomnia, periodic leg movement disorder , and rapid eye movement behavior disorder — as well as healthy participants were used. K- nearest neighbors and ensemble bagged trees (EbagT) classifiers yield the greatest results. For categorizing phases A and B, the suggested model's average classification accuracy was 91.6% for healthy participants and 94.33%, 86.3%, 88.68%, 84.43%, and 88.5% for patients with narcolepsy, RBD, PLM, NFLE, and insomnia, respectively. When classifying A subphases (A1, A2, A3), our model's average classification accuracy was 92.85% for healthy individuals and 93.9%, 84.9%, 88.0%, 80.92%, and 89.41% for narcolepsy, RBD, PLM, NFLE, and insomnia participants, respectively. The suggested approach could make it easier for sleep specialists to use the microstructure of sleep to automatically assess an individual's quality of sleep [5].

Five somnographic-like signals that are developed using optical, differential air-pressure, and acceleration signals obtained by a chest-worn sensor. In order to forecast the overall signal quality (normal, corrupted), three breathing- related patterns (normal, apnea, irregular), and three sleep-related patterns (normal, snoring, noisy), this solves a three-fold classification issue. The created architecture produces extra information in the form of quantitative (confidence indices) and qualitative (saliency maps) data to enhance explainability and enhance prediction interpretation. Twenty healthy participants were observed as they slept for about ten hours each night. The training dataset was constructed by manually labeling somnographic-like signals based on the three class groupings. To assess the prediction performance-c and the consistency of the findings, analyses were conducted both record- wise and subject-wise. The network's accuracy in differentiating between normal and corrupted signals was 0.96.The accuracy of breathing pattern prediction was higher (0.93) than that of sleep pattern prediction (0.76). Apnea was predicted more accurately (0.97) than abnormal breathing (0.88).The differentiation between noise occurrences (0.61) and snoring (0.73) was less successful in the sleep pattern condition [6].

A deep-learning architecture combining convolutional neural networks and recurrent neural networks are used to analyze through out night airflow and oximetry signals for identifying pediatric obstructive sleep apnea (OSA). Data from the Childhood Adenotonsillectomy Trial public database (1,638 subjects) and a proprietary database (974 subjects) were distributed into 30-minute intervals, handled by the model to estimate apneic events, and used for calculating the apnea-hypopnea index (AHI) and categorize OSA harshness into four levels. The model made high deal in AHI regression with an intraclass correlation coefficient go beyond 0.9 and diagnostic accuracies over 84% for AHI cutoffs of 1, 5, and 10 events per hour. OSA sincerity category accuracies were 74.51% for CHAT and 62.31% for the proprietary dataset, with Cohen's Kappa values of 0.6231 and 0.4495, representing moderate to substantial agreement. Grad-CAM visualizations draw attention to critical features such as airflow cessations and SpO2 drops but often forgot hypopneas connected to arousals. Restrictions included dataset specificity affecting generalizability, varying performance across populations, and faces in model interpretability. Despite these constraints, the CNN + RNN model demonstrates potential as an accurate and interpretable diagnostic tool for pediatric OSA in clinical settings [7].

A deep learning architecture to analyze signals from a chest-worn sensor capturing optical, differential air- pressure, and acceleration data, converting them into five somnographic-like signals to address a three-fold classification problem: forecasting signal quality (normal or corrupted), identifying breathing patterns (normal, apnea, irregular), and detecting sleep patterns (normal, snoring, noise) . The dataset, including though out night monitoring of twenty healthy subjects with manually labeled signals, revealed high accuracy in unique normal from corrupted signals (0.96) and excellent performance in predicting apnea (0.97), however irregular breathing was less accurate (0.88). Sleep pattern estimate was weaker (0.76), specifically in distinguishing snoring (0.73) and noise (0.61). In spite of the small sample size restraining generalizability, the study

highlighted the importance of explainability in deep learning over saliency maps and confidence indices, marking a significant step toward clinical AI applications for sleep disorder discovery while emphasizing the need for further validation in greater, more diverse populations [8].

An explainable sleep staging algorithm employing single-channel EEG data to simplify sleep stage evaluation. Key techniques include signal decomposition through band-pass filters, an attention mechanism to consentrates on significant EEG features, and cross-evaluation utilizing data from 80 subjects to ensure robustness. The model achieved an average F1-score of 72.66 (±22.24), representing balanced precision and recall in classifying five sleep stages. Explainability was a critical feature, with the attention mechanism underlining influential EEG components, enhancing trust and understanding in clinical treatments. Restrictions include potential data oversimplification, variability in F1-scores, limited generalizability, lack of longitudinal data, and explainability challenges. Although these, the algorithm demonstrates promise as a more effective and interpretable alternative to traditional polysomnography-based techniques for sleep staging, advancing diagnostic processes [9].

The study introduces an AI-driven system, DreamGuardian, designed for the early finding and estimate of sleep apnea by machine learning algorithms and advanced signal processing techniques. The methodology involves acquiring a diverse dataset from sleep clinics and repositories, formatting the data, and selecting key features such as heart rate, BMI, age, gender, respiratory rate, ESS score, AHI, and oxygen saturation. Non-invasive sensors integrated into a wearable device prototype monitor physiological signals such as respiratory patterns, heart rate variability, and movement, with standardized data collection protocols ensuring ethical compliance. While the system demonstrates potential for accurate sleep apnea finding, the study highlights restrictions, including variability introduced through the diverse dataset, dependence on sensor accuracy, and limited accessibility, which could hinder widespread adoption. Additionally, the lack of specific quantitative performance metrics causes its real-world effectiveness unclear. Although these challenges, the DreamGuardian system represents a promising innovation in addressing sleep apnea, with chances for future refinement and broader application [10].

The use of artificial intelligence -based models, mainly machine learning and deep learning, to automate the finding and classification of sleep disorders, targeting to increase diagnostic speed and accuracy compared to traditional, human-dependent methods, which are time-consuming and costly. A comparative analysis of existing research underlines key outcomes and challenges, including variability in data quality and availability, which significantly impact AI model performance. The complexity of sleep disorders and the necessity for large, diverse datasets for efficient training are also identified as critical obstacles. Although these restrictions, the study proposes a framework for future research to discover and classify multiple sleep disorders utilizing advanced AI models, streamlining diagnostics and expanding patient outcomes [11].

The study applied a novel deep learning architecture to identify EEG biomarkers for assessing the severity of Obstructive Sleep Apnea . The model was trained by applying features derived from EEG signals and other biomarkers, like desaturation area, arousal events, and respiratory event duration, to detect an objective metric for OSA severity assessment . But, the study has restrictions, including the absence of consensus on a common efficient metric for OSA severity, which may affect the generalizability of the findings across several populations and settings. Moreover, the reliance on specific EEG frequency bands (0–8 Hz) may overlook other potentially relevant features in the EEG data . The results demonstrated that the model efficiently identified significant EEG features from posterior and medial regions that correlate with apnea-hypopnea severity, paving the way for the application of Explainable Artificial Intelligence in making OSA severity assessments more objective and facilitating the discovery of EEG biomarkers in various tasks [12].

3 XAI Techniques in Sleep Disorder Diagnostics

Several XAI techniques have been employed to interpret AI models in sleep disorder diagnostics. These include:

- Feature Importance Methods:

Methods such as SHapley Additive Explanations and Local Interpretable Model-agnostic Explanations (LIME) provide insight on which characteristics—such as heart rate, oxygen saturation, and EEG signals—have the most influence on a model's predictions.

- Integrated Gradients (IG):

IG is a well-liked technique for deep learning models that links a model's output to its input characteristics. This makes it useful for examining time-series data, like EEG recordings.

- Saliency Maps:

visualizations that draw attention to the areas of input data like EEG or ECG signals that are the most bearing in the decision-making process.

- Counterfactual Explanations:

Provide clinicians practical insights by illustrating how slight change to input data could affect diagnostic results. Overview of XAI Techniques in Sleep Disorder Diagnostics

The proposed system integrates AI and XAI techniques to ensure accurate, interpretable, and clinically reliable diagnosis of sleep disorders. the workflow of the *Explainable AI for Sleep Disorder Diagnostics* system is designed. It proceeds by the following major stages:

1. Data Collection

Gathering physiological data from multiple sources such as wearable devices and polysomnography (PSG) are the beginning process. The vital signals including electroencephalogram (EEG), electrocardiogram (ECG), oxygen saturation (SpO_2), respiratory rate, and body movement are recorded by these sources. These different data serve as the foundation for AI-based sleep disorder analysis.

2. Dataset Creation

The collected raw data is mound into structured datasets. These datasets include labeled information which corresponds to various sleep stages and disorders like insomnia, sleep apnea, narcolepsy, and restless leg syndrome. For supervised machine learning and deep learning model training labeling is essential.

3. Data Preprocessing

This step includes cleaning and preparing the data for analysis. Noise and artifacts are removed by filtering techniques and handling the missing values. For analysis signals are segmented into time-series windows. To ensure the input data is of high quality and ready for modeling, feature extraction is then performed to identify key patterns from EEG, ECG, and SpO_2 signals

4. Prediction Model Development

To detect and classify sleep disorders machine learning (ML) and deep learning (DL) models are developed. Techniques like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and ensemble learning methods are applied. These models classify sleep stages by learning to differentiate normal and abnormal sleep patterns, predict apnea events.

5. Explainable AI (XAI) Integration

XAI methods like SHAP, LIME, Integrated Gradients, Saliency Maps, and Counterfactual Explanations are integrated to overcome the “black-box” nature of AI models. These methods help in visualizing and interpret how different physiological features influence the AI model’s decisions, enhancing transparency and trust.

6. Performance Evaluation and Clinical Validation

Using performance metrics such as accuracy, F1-score, and kappa coefficient the developed models are evaluated. To ensure reliability clinical experts validate the model outputs. With known physiological and clinical knowledge, the integration of XAI enables medical professionals to verify that the model’s reasoning aligns.

7. Output and Interpretation

Finally, with visual explanations and confidence levels the system provides diagnostic insights, allowing both clinicians and patients to understand and trust the diagnostic outcomes. In clinical settings this interpretability enhances patient trust, supports personalized treatment planning, and promotes adoption. The following Fig. 1 illustrates the workflow of the proposed system.

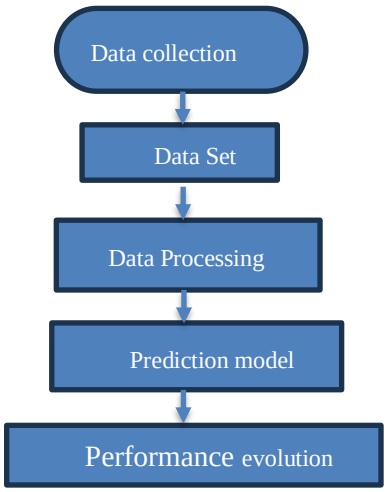

Fig.1. Conceptional workflow of Explainable AI in Sleep Disorder Diagnostics

4 Integration with Wearable Technology

The widespread use of wearable technology has increased accessibility to ongoing sleep monitoring. Numerous physiological indications, such as blood oxygen levels, body motions, and heart rate variability, are recorded by these devices. XAI is important for understanding the vast data produced by wearables and making sure that the results of AI models are intelligible and useful for both patients and clinicians . By applying SHAP on wearable data, for example, researchers can identify specific patterns linked to sleep apnea events, enhancing early detection and treatment.

5 Applications of XAI in Sleep Disorder Diagnostics

Obstructive Sleep Apnea (OSA):
Research has shown that by analyzing oxygen desaturation and airflow patterns, the effectiveness of XAI is useful in detecting apnea episodes. Also, XAI enhancing the trust by validating AI predictions against healthcare recommendations.
Insomnia:
Based on lifestyle and sleep patterns, machine learning models have been used to predict the severity of insomnia. XAI techniques like counterfactual explanations make clear how factors like screen time or coffee consumption affect predictions.
Narcolepsy:
By analyzing PSG data using deep learning models XAI methods help clinicians understand the neural mechanisms linked with excessive daytime sleepiness and sleep onset REM periods.
Restless Leg Syndrome (RLS):
Periodic limb movements and related brain activity are Key indicators for RLS, can be found using feature attribution approaches. The following table.1 portrays the comparison of XAI techniques in sleep disorder diagnostics

Table 1. Comparison of XAI Techniques in Sleep Disorder Diagnostics

XAI Technique	Application	Advantages	Limitations
SHAP	The significance of features for OSA and insomnia	High interpretability and independence from models	computationally costly
LIME	Localized justifications for wearable information	Easy to execute and adaptable	Restricted to regional justifications
Integrated Gradients	Analysis of time series for PSG data	compatibility with deep learning, strong	Differentiable models are necessary.
Saliency Maps	Visual aids for analyzing signals	Simple to understand and intuitive	Lacks global insights and is susceptible to noise
Counterfactuals	Explanations based on scenarios	Practical,patient-specif ic information	High-dimensional data is difficult to handle.

6 Benefits of XAI in Sleep Disorder Diagnostics

Transparency:
By improving the interpretability of AI models, XAI helps making their outputs understandable to clinicians and patients.
Clinical Validation:
By coordinating AI predictions with accepted medical knowledge, XAI enhances the reliability of diagnostic tools
Patient Trust:
Transparent models-built trust among patients, encouraging increased use of AI-based diagnostics.
Personalized Treatment:
Due to the development of tailored treatment plans, XAI makes it easier to create individualized treatment programs.

7 Challenges and Limitations

Despite its advantages, XAI faces several challenges in sleep disorder diagnostics:
Complexity of Sleep Data:
To extract valuable insights from sleep data, advanced XAI methods are required due to its high complexity and unpredictability.
Computational Overhead:
Real-time analysis is challenging one because of the computational complexity of many XAI algorithms.
Interpretability vs. Accuracy Trade-off:
Finding a balance between diagnostic accuracy and model interpretability remains a crucial task.
Limited Standardization:
The use of XAI is imited due to the lack of established procedures for integrating it into clinical processes.

8 Future Directions

By integrating AI-based methods with conventional statistical models to make use of their respective advantages. For real-time diagnostics, creating effective and lightweight XAI algorithms in wearable device will be considered. For comprehensive diagnostics, the use of XAI to combine data from multiple sources, such as PSG, wearables, and patient-reported outcomes are proposed. An addition, one has to consider standard guidelines for the safe and effective clinical application of XAI. Finally, creating XAI models that not only aid clinicians but also directly give patients intelligible information while simultaneously assisting physicians.

9 Conclusion

Explainable AI shows a transformative approach to increase the reliability and transparency of AI-based sleep disorder diagnostics. Using the bridge gap between complex AI models and clinical applicability, XAI build trust, increases diagnostic accuracy, and makes personalized treatment. Future research should concentrate on resolving the challenges of XAI integration and developing patient-centered solutions to unlock its difficulties in sleep medicine.

References

1. Xu, S., Faust, O., Seoni, S., Chakraborty, S., Barua, P.D., Loh, H.W., Elphick, H., Molinari, F., Acharya, U.R.: A review of automated sleep disorder detection. *Computers in Biology and Medicine* **150**, 106100 (2022)
2. Vaquerizo-Villar, F., Gutiérrez-Tobal, G.C., Calvo, E., Álvarez, D., Kheirandish-Gozal, L., Del Campo, F., Gozal, D., Hornero, R.: An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea. *Computers in Biology and Medicine* **165**, 107419 (2023)
3. Ingle, M., Sharma, M., Verma, S., Sharma, N., Bhurane, A., Acharya, U.R.: Automated explainable wavelet-based sleep scoring system for a population suspected with insomnia, apnea and periodic leg movement. *Medical Engineering & Physics* **130**, 104208 (2024)
4. Parbat, D., Chakraborty, M.: Multiscale entropy analysis of single lead ECG and ECG derived respiration for AI-based prediction of sleep apnea events. *Biomedical Signal Processing and Control* **87**, 105444 (2024)
5. Sharma, M., Lodhi, H., Yadav, R., Sampathila, N., Swathi, K.S., Acharya, U.R.: Automated explainable detection of cyclic alternating pattern (CAP) phases and sub-phases using wavelet-based single-channel EEG signals. *IEEE Access* **11**, 50946–50961 (2023)
6. Yook, S., Kim, D., Gupte, C., Joo, E.Y., Kim, H.: Deep learning of sleep apnea-hypopnea events for accurate classification of obstructive sleep apnea and determination of clinical severity. *Sleep Medicine* **114**, 211–219 (2024)
7. Jiménez-García, J., García, M., Gutiérrez-Tobal, G.C., Kheirandish-Gozal, L., Vaquerizo-Villar, F., Álvarez, D., Del Campo, F., Gozal, D., Hornero, R.: An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals. *Biomedical Signal Processing and Control* **87**, 105490 (2024)
8. Rossi, M., Sala, D., Bovio, D., Salito, C., Alessandrelli, G., Lombardi, C., Mainardi, L., Cerveri, P.: Sleep-see-through: Explainable deep learning for sleep event detection and quantification from wearable somnography. *IEEE Journal of Biomedical and Health Informatics* **27**(7), 3129–3140 (2023)
9. Baek, S., Baek, J., Yu, H., Lee, C., Park, C.: Explainable sleep staging algorithm using a single-channel electroencephalogram. *IEIE Transactions on Smart Processing & Computing* **11**(1), 8–13 (2022)

10. Singh, K., Mehta, A., Chaudhary, A.: AI-driven sleep apnea detection and prediction. In: *Proceedings of the 2024 International Conference on Computational Intelligence and Computing Applications (ICCICA 2024)*, vol. 1, pp. 237–241. IEEE, New York (2024)
11. Kaur, A., Neeru, N.: Automatic detection and classification of sleep disorders using AI-learning models. In: *Proceedings of the 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE 2024)*, pp. 1017–1022. IEEE, New York (2024)
12. La Fiscal, L., Jennebauffe, C., Bruyneel, M., Ris, L., Lefebvre, L., Siebert, X., Gosselin, B.: Explainable AI for EEG biomarkers identification in obstructive sleep apnea severity scoring task. In: *Proceedings of the 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER 2023)*, pp. 1–6. IEEE, San Diego (2023)