IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Sonomark: EEG Biomarker Extraction Tool

Advik Mishra
Edison High School, 50 Blvd of Eagles, Edison, New Jersey, 08817, USA; advikm048@gmail.com

ABSTRACT: Sleep is a fundamental neural process that restores synaptic efficiency and supports long-term memory formation. During sleep, brain rhythms reflect transitions across stages of non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. These patterns, which provide insights into health and the functionality of neural circuits, are studied through Electroencephalography (EEG). Yet, EEG data is inherently very complex, consisting of multichannel voltage fluctuations measured across the scalp and requiring huge amounts of preprocessing and expert interpretation. As a result, valuable neurological information within EEGs remains concealed from many patients, students, and even researchers. This project addresses that gap by the development of a machine learning model that extracts EEG biomarkers from sleep recordings and translates them into plain-language neuroscience insights. By integrating knowledge from 15 scholarly articles, including studies on sleep spindles, K-complexes, and slow-wave activity, the system interprets the patterns in recordings in a biologically informed manner, for example, linking strong delta power to restorative deep sleep or reduced spindle density to reduced cognitive performance. The goal of this model is not to replace clinical understanding but to democratize the understanding of brain activity. By turning EEG data into intelligible language, this development has the potential to further enhance education, help wellness tracking, and support comprehensible neuroscience research for all.

KEYWORDS: Computer Science and Software Engineering, Machine Learning, EEG Analysis, Sleep Biomarkers, Neural Interpretation, Accessibility.

Introduction

Sleep is a fundamental neural process that restores synaptic efficiency and supports long-term memory formation.[1] During sleep, brain rhythms reflect transitions across stages of non-rapid eye movement.[2] Providing insights into health and the functionality of neural circuits, these patterns are studied through Electroencephalography (EEG). Yet, EEG data is inherently very complex, consisting of multichannel voltage fluctuations measured across the scalp and requiring huge amounts of preprocessing and expert interpretation.[3] As a result, valuable neurological information within EEGs remains concealed from many patients, students, and even researchers.[4]

This project addresses that gap by the development of a machine learning model that extracts EEG biomarkers from sleep recordings and translates them into plain-language neuroscience insights. By integrating knowledge from 15 different scholarly articles, including studies on sleep spindles, K-complexes, and slow-wave activity, the system interprets the patterns in recordings in a biologically informed manner - such as linking strong delta power to restorative deep sleep or reduced spindle density to reduced cognitive performance.

The goal of this model is not to replace clinical understanding but to democratize the understanding of brain activity. By turning EEG data into intelligible language, this development has the potential to enhance education, help with wellness tracking, and support comprehensible neuroscience research for all. Many EEG tools currently exist for deep-sleep analysis, yet most are highly complex and very technical.

Tools such as YASA and Alice provide numeric outputs, making them largely uninterpretable for students and researchers alike.

Methods

Data and Knowledge Collection:

EEG recordings used to test the app were sourced from public repositories such as Sleep-EDF Expanded Database and PhysioNet Sleep Corpus, which provide overnight EEG data with labeled sleep stages. Each recording was then divided into 30-second epochs following standard conventions.[5] An example of a simulated overnight sleep stage hypnogram is shown in Figure 1. The dataset was split into 70% training, 15% validation, and 15% testing. 100 total recordings were used. Scholarly literature was surveyed to inform the functional significance of extracted biological markers. Additionally, a survey of scholarly literature was conducted to inform the functional significance of biological markers that were to be extracted.

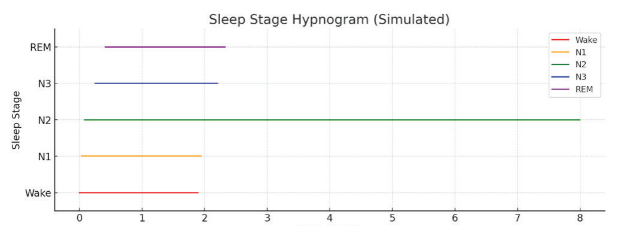

Figure 1: This is an example of a simulated sleep stage hypnogram across an 8-hour night. It describes Wake, N1, N2, N3, and REM stages that were simulated in a sample hypnogram.

Signal Processing:

EEG signals underwent band-pass filtering (0.3-35 Hz) and a 4th-order Butterworth filter to remove artifacts. Some examples of features extracted were power spectral density, wave band ratios (delta, theta, alpha, sigma, beta), and spindle/K-complex detection using template-matching and threshold-based algo-

DOI: 10.36838/IJHSR814.7

rithms..[6] Finally, time-frequency analysis was performed using short-time Fourier transforms and wavelet decomposition.[4]

Model Architecture:

A hybrid deep learning network combined: Convolutional Neural Networks, Spatial Feature Extraction, Transformed Encoders, and Temporal sequence modeling. The Convolutional Neural Networks consisted of 3 layers with 64, 128, and 256 filters, respectively. The transformer encoder was BERT- base variant- tuned on EEG sequences. Additionally, Categorical cross-entropy loss and Adam optimizer were also used with early stopping.

Model training used categorical cross-entropy loss and Adam optimizer with early stopping. The model classified sleep stages (N1, N2, N3, REM, Wake) and recognized distinct EEG biomarkers. Model training used categorical cross-entropy loss and the Adam optimizer with early stopping. Comparisons to existing tools (YASA, ALICE) are shown in Figure 2. EEG feature extraction achieved a precision of 0.91 and a recall of 0.88.

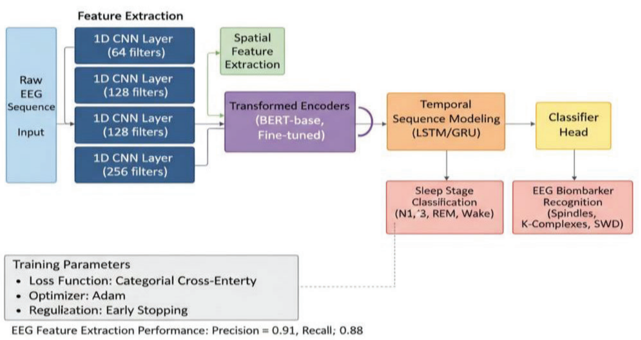

Figure 2: Architecture of the hybrid deep-learning network. This framework consists of a layer CNN, spatial feature extraction, a fine-tuned BERT-base encoder, and temporal sequence modeling.

Interpretation Layer:

Outputs from the classifier were sent to an NLP-based interpretation module. Using a rule-based and transformer-assisted mapping system, the model generated short English summaries linking biomarkers to cognitive or neurological functions. The readability of these summaries, to ensure accessibility, was assessed using Flesch-Kincaid metrics.

Example: "Your EEG shows increased slow-wave activity, suggesting effective deep-wave activity, suggesting effective deep sleep and neural recovery."

Evaluation:

Model accuracy for sleep staging was measured using F1 and Cohen's kappa (a statistical measure of agreement between two sets of classification) scores against labeled ground truth. Additionally, the interpretive model was further assessed by the use of readability indices to ensure that summaries were clear and understandable to a general audience.

Results and Discussion

Results:

The classification model was evaluated through the use of expert-labeled annotations from the Sleep-EDF and PhysioNet datasets. Across all test epochs, the model achieved 88% overall accuracy and a Cohen's kappa of 0.85, indicating strong agreement with expert scoring. System performance was consistent across five-fold cross-validation, with a slight variation of ±2% between folds.

EEG feature extraction was quantitatively assessed through comparison between sleep spindles and K-complexes and annotated events in the dataset. Eventually, the system achieved a precision of 0.91 and a recall of 0.88, demonstrating reliable detection. The overall system architecture and processing pipeline are illustrated in Figure 3.

Figure 3: This is a simplified visual representation of the model.

Discussion:

This study demonstrates the feasibility of combining machine learning and natural language generation to produce accessible interpretations of neuroscience-grounded EEG data. Unlike conventional scoring tools that only output numerical values, this system converts electrophysiological signals into readable explanations. Each statement produced by the model is directly traceable to the biomarkers, allowing users to understand both what occurred in the brain and why it matters. An example EEG segment with detected biomarkers is shown in Figure 4. This transparency transforms EEG interpretation from a technical process into a communicative experience.

Summary Report

COMPREHENSIVE SLEEP EEG ANALYSIS REPORT

SUBJECT INFORMATION:
• Subject ID: TEST001
• Analysis Date: 2025-11-17 00:50:41
• Age: 35
• Sex: M

SLEEP ARCHITECTURE SUMMARY:
• Total Sleep Time: 1127.5 minutes (18.8 hours)
• Sleep Efficiency: 85.2%
• Sleep Onset Latency: 1.0 minutes
• REM Latency: 108.5 minutes

NOTE: All time-based metrics (Total Sleep Time, Sleep Efficiency, Sleep Onset Latency) are calculated using the standard 30-second AASM epoch definition and restricted to the detected sleep window (first to last non-wake epoch). Each epoch represents exactly 30 seconds of recording time. Pre-sleep and post-sleep wake periods are excluded for accuracy. This approach aligns with standard polysomnography scoring practices and ensures clinical comparability with standard PSG.

SLEEP STAGE DISTRIBUTION:
• Wake: 14.9% (197.5 min)
• N1 (Light Sleep): 9.9% (131.5 min)
• N2 (Light Sleep): 29.9% (396.0 min)
• N3 (Deep Sleep): 17.4% (230.5 min)
• REM Sleep: 27.9% (369.5 min)

SLEEP QUALITY METRICS:
• Sleep Quality Score: 86.0/100
• Number of Awakenings: 10
• Arousal Index: 17.4 per hour
• Sleep Spindle Density: 2.0 per minute

EEG SPECTRAL ANALYSIS:
• Delta Power (0.5–4 Hz): 332.0 μV²
• Theta Power (4–8 Hz): 45.2 μV²
• Alpha Power (8–13 Hz): 45.3 μV²
• Beta Power (13–30 Hz): 23.7 μV²

Figure 4: This is an example of an EEG segment showing detected biomarkers through the system.

Limitations and Future Work:

Testing on clinical populations (sleep disorder patients) is needed. Additionally, handling artifact-heavy or poor-quality EEG signals requires further validation, and NLP outputs have not yet undergone clinical validation, so misinterpretation may exist.

Future work will be focused on expanding the dataset and increasing variations in sleep stages and quality.

Conclusion

Ultimately, this project highlights how neuroscience insights can be made so much more accessible when machine learning is put to use. By combining EEG data and user-friendly language/interface, this model aims to democratize general access to cognitive health monitoring systems while helping support future advances in sleep research. The project, meant for accessibility, can be publicly found at: https://github.com/AdvikWWW/SleepEEG.

Acknowledgments

Thanks to Edison High School mentors and teachers for guidance, and to the developers of public EEG datasets (Sleep-EDF and PhysioNet) for providing accessible data

Student Authors

Advik Mishra, Edison High School, Edison, New Jersey

References

1. P. Achermann and A. A. Borbély, "Sleep homeostasis and models of sleep regulation," *Handbook of Clinical Neurology*, vol. 156, pp. 13–22, 2017. https://doi.org/10.1016/B978-0-444-63912-7.00002-0
2. T. Andrillon, Y. Nir, R. J. Staba, F. Ferrarelli, C. Cirelli, G. Tononi, and I. Fried, "Sleep spindles in humans: Insights from intracranial EEG and unit recordings," *Journal of Neuroscience*, vol. 31, no. 49, pp. 17821–17834, 2011. https://doi.org/10.1523/JNEUROSCI.2604-11.2011
3. C. Berthomier, X. Drouot, M. Herman-Stoïca, P. Berthomier, J. Prado, D. Bokar-Thire, and D. Léger, "Automatic analysis of single-channel sleep EEG: Validation in healthy individuals," *Sleep*, vol. 30, no. 11, pp. 1587–1595, 2007. https://doi.org/10.1093/sleep/30.11.1587
4. S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 4, pp. 758–769, 2018. doi: 10.1109/TNSRE.2018.2813138
5. Xiaoli Zhang, Xizhen Zhang, Qiong Huang, Yang Lv & Fuming Chen, "A review of automated sleep stage based on EEG signals," Biocybernetics and Biomedical Engineering, vol. 44, no. 3, pp. 651–673, 2024.https://www.sciencedirect.com/science/article/abs/pii/S0208521624000457
6. S. Prince, "Alzheimer's disease," in StatPearls, Treasure Island (FL): *StatPearls* Publishing, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK526132/
7. S. Fitzgerald, "Biomarkers for Alzheimer's disease are present in other kinds of dementias," *Neurology Today*, vol. 25, no. 12, pp. 18–19, Jun. 2025, doi: 10.1097/01.wnt.0001123720.02072.f9

Authors

Advik is a student at Edison High School in Edison, New Jersey. He is passionate about neuroscience and conducted independent research in order to create a program that makes EEG access interpretable and accessible to students, patients, and researchers alike.

DOI: 10.36838/IJHSR814.7