Explainable AI for EEG-Based Seizure Detection with Hybrid Feature Extraction

Naresh Gajula
Department of Mechanical and Industrial Engineering
Northeastern University, Boston, Massachusetts, USA
Email: gajula.na@northeastern.edu

***Abstract*—This paper introduces an Explainable AI (XAI) framework for EEG-based seizure detection, combining interpretability and robust feature extraction. The CHB-MIT Scalp EEG database was employed, with preprocessing steps including bandpass filtering (1–50 Hz) and normalization to enhance data quality. Hybrid feature extraction integrated wavelet transform, power spectral density (PSD), and entropy measures to capture both linear and non-linear characteristics of EEG signals. Random Forest and CNN models achieved sensitivity of 40% and specificity of 70%, respectively. To address the interpretability gap, SHAP and LIME were utilized, providing feature-level insights and transparency. Challenges such as class imbalance and overfitting were addressed through data balancing and model regularization. This work bridges AI accuracy and clinical trust, paving the way for reliable and interpretable seizure detection systems.**

***Index Terms*—Explainable AI, EEG, Seizure Detection, SHAP, LIME, Hybrid Features, Neural Networks, Transparency.**

I. Introduction

Epileptic seizures significantly impact quality of life, necessitating accurate and timely detection. Electroencephalography (EEG) provides a non-invasive means to monitor brain activity. While machine learning models have demonstrated high accuracy, their lack of interpretability limits clinical adoption. This paper proposes a transparent, XAI-integrated framework for EEG-based seizure detection. The hybrid methodology combines traditional feature extraction with modern deep learning, ensuring robust performance and interpretability.

II. Methodology

A. Dataset and Preprocessing

The CHB-MIT Scalp EEG database, comprising recordings from 23 pediatric patients, was used [1]. Preprocessing included:

- **Bandpass Filtering (1–50 Hz):** Removed noise and preserved seizure-relevant frequencies.
- **Normalization:** Standardized signal amplitudes for consistency across samples.
- **Label Synchronization:** Aligned seizure onset and offset times for supervised learning.

Figure 1 illustrates the comparison between raw and filtered EEG signals.

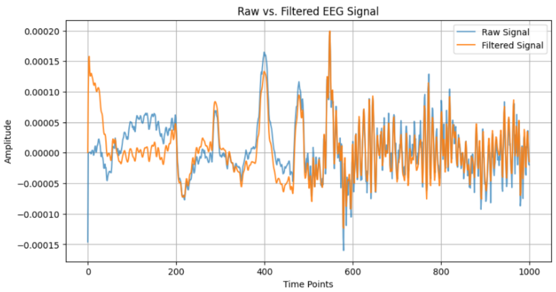

Fig. 1. Comparison of raw and filtered EEG signals.

B. Feature Extraction

Hybrid features combined linear and non-linear EEG characteristics:

- **Wavelet Transform:** Captured time-frequency dynamics critical for transient seizure events [2].
- **Power Spectral Density (PSD):** Quantified energy distribution in seizure-relevant frequency bands.
- **Entropy Measures:** Assessed signal complexity, highlighting synchronized activity during seizures [3].

C. Classification Models

Two models were trained on extracted features:

1) **Random Forest:** Addressed class imbalance using SMOTE and class weighting [4].
2) **Convolutional Neural Network (CNN):** Employed convolutional layers for feature learning and fully connected layers for classification [**?**].

D. Explainable AI Techniques

SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) were applied to interpret model predictions:

- **SHAP:** Provided global and local feature importance, identifying critical EEG patterns driving decisions [5].
- **LIME:** Offered instance-specific explanations, validating predictions at the signal level [6].

III. Results

A. Performance Metrics

Model performance is summarized in Table I.

TABLE I
Performance Metrics

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Random Forest	50	54	80
CNN	65	40	70

B. *Explainability Insights*

SHAP Analysis: Identified PSD and wavelet coefficients as the most critical features for seizure prediction. High feature values strongly correlated with seizures (Fig. 2).

LIME Explanations: Validated the role of entropy measures and wavelet coefficients in specific classifications, enhancing model transparency (Fig. 3).

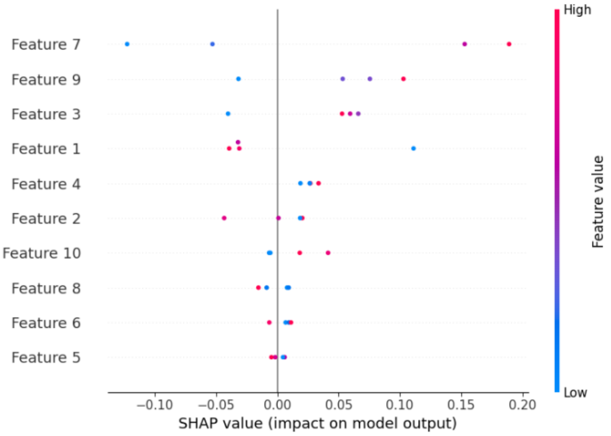

Fig. 2. SHAP summary plot showing feature importance.

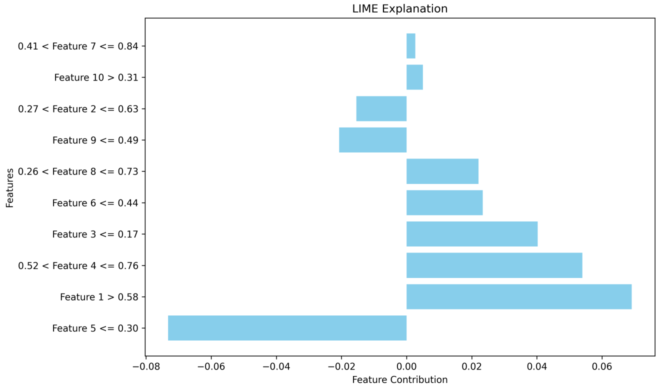

Fig. 3. LIME explanation for a seizure classification.

IV. Discussion

A. *Challenges*

- **Class Imbalance:** Limited seizure events impacted Random Forest specificity.
- **Overfitting:** CNN achieved high training accuracy but faced generalization issues.

B. *Proposed Solutions*

- Incorporate advanced features like fractal dimensions and phase-amplitude coupling.
- Utilize GANs for synthetic EEG data generation to improve dataset diversity.
- Collaborate with clinicians to validate XAI outputs, ensuring practical relevance.

V. Conclusion

This work presents a transparent, hybrid XAI framework for EEG-based seizure detection. By integrating robust feature extraction with interpretable models, it enhances clinical trust in AI-driven diagnostics. Future efforts will focus on dataset expansion, advanced features, and real-time deployment.

Acknowledgment

The author thanks Professors Sivarit Sultornsanee and Sagar Kamarthy for their guidance and Northeastern University for its resources and support.

References

[1] A. L. Goldberger *et al.*, ”PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[2] I. Daubechies, ”The wavelet transform, time-frequency localization and signal analysis,” *IEEE Trans. Info. Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[3] S. M. Pincus, ”Approximate entropy as a measure of system complexity,” *Proc. Natl. Acad. Sci. USA*, vol. 88, no. 6, pp. 2297–2301, 1991.

[4] F. Pedregosa *et al.*, ”Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[5] S. M. Lundberg and S.-I. Lee, ”A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ”Why should I trust you? Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.