Enhancing Trust in Brain-Computer Interfaces: Interpretable Deep Learning for EEG Signal Classification

A.H.M.T.C. Bakmeedeniya
NSBM Green University
Sri Lanka
thilini.b@nsbm.ac.lk

Lashika Chamini
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
lashika.c@sliit.lk

Krishantha Ranaweera
NSBM Green University
Sri Lanka
krishantha.r@nsbm.ac.lk

***Abstract*— Brain signal is one of the methods in healthcare to identify brain activities. Mathematical modeling to machine learning and modern deep learning architectures are used to analyze these signals. These methods have been tested with EEG signals to capture meaningful insights and uncover patterns. Deep learning often outperforms conventional machine learning. But the black box nature of these signals has become a concern, especially in the healthcare sector. Hence, this study focused on explaining the Deep Learning architecture by using methods such as Permutation Feature Importance, Saliency Maps, and Gradient-weighted Class Activation Mapping. To test this, a model for sleep-stage classification using Electroencephalography (EEG) has been developed, as EEG is easier to interpret. The model was created by fusing different types of deep learning architecture. To capture sleep-related frequency activity, Power Spectral Density (PSD) features were extracted. Further, jitter ing was applied. Accuracy and F1-score used to measure the performance of the model. An average accuracy of 0.85 and the weighted F1-score of 0.87 were given on the test set. The results display that combining deep learning with explanation methods and careful info handling can develop EEG-based systems extra transparent and trustworthy.**

***Keywords*— Data Augmentation, Electroencephalography, Explainable AI, Saliency Maps, Grad-CAM**

I. INTRODUCTION

EEG signal analysis has become a growing area as it allows the extraction of meaningful insights and brain pat terns. Controlling prosthetic limbs, supporting people with severe movement disabilities [1], and providing neurofeedback therapy for conditions like Attention-Deficit Hyperactivity Disorder (ADHD) and recovery after stroke [2] are some application areas of EEG based analysis. Machine Learning is a major technique used in the past two decades, as it efficiently captures the hidden patterns. More recently, with the Deep Learning, researchers have experimented various Deep Learning architectures. It allows for capturing patterns even from the raw data without manual feature extraction. But these models are not interpretable and difficult to understand. When in the clinical setup, clinicians are required to understand how the models made the relevant results to trust the model. This helps them to make informed decisions [3]. Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are some techniques that can provide insight into which brain regions are involved in a task [4], for example in deception detection, and why a model gave a particular output [5]. Still, most of these techniques are not designed specifically for EEG or other physiological data [6] in sleep stage classification. To bridge the gap, researchers have started to look at explanation methods that take the nature of EEG signals into account. This study builds a model to classify sleep stages from EEG signals using deep learning techniques and uses explanation methods to justify the way the model gets decisions. The research covers model design and applies interpretability tools to highlight important signal features. Different explanation methods are compared to see how they affect both model interpretability and performance.

II. LITERATURE REVIEW

EEG helps in decoding human cognitive states such as attention, emotion, and mental workload. Non-invasive, high temporal resolutions and portability are some characteristics offered by EEG. These make them matches for real-time mental health analysis [7]. Support Vector Machines, and k nearest Neighbours are some commonly discussed machines learning models in the studies. These models have shown greater success in past research [4] . Similarly, deep learning models like Convolutional Neural Networks (CNNs) can also learn effectively and capture human cognitive state. CNNs, particularly the DeepConvNet and ShallowConvNet models introduced by, showed significant improvement in EEG decoding by learning frequency-specific spatial filters. Long Short Term Memory (LSTM) networks [8] have been used to capture the temporal dependencies, especially in motor imagery tasks. However, these models are having several limitations. High computational cost and a lack of generalizability across sub jects are two among those. To address these issues, [9] study has proposed EEGNet, a compact and generalizable CNN architecture designed specifically for BCIs that uses EEG. The objective was to create a model that works across multiple motor imagery and P300 with minimal preprocessing and high efficiency. Similarly, another study [3] aimed to optimize both performance and efficiency by introducing EEG-ITNet. EEG-ITNet is a deep learning model that integrates inception

modules with causal dilated convolutions. This architecture efficiently extracts multi-scale spectral, spatial, and temporal features from EEG data. Also maintains a significantly lower parameter count than models like EEG-Inception and EEG TCNet Public datasets like DEAP (for emotion analysis) and SEED (for cross-subject affective computing) have standard ized benchmarking, enabling improvements in deep learning architectures. CNNs [10] and their variants are highly used by the studies. Most recent studies have discussed Transformers [10] as an efficient model in achieving state-of-the-art accu racy. These techniques models spatiotemporal dependencies in raw signals. These techniques have improved the decoding accuracy of cognitive states vastly. Despite advances in deep learning, many models still do not show how they reach their predictions. To tackle this, researchers have applied explanation methods to EEG and other brain signals [3]. These methods offer some understanding of how the models make their decisions. However, deep learning models are still hard to interpret. This can reduce transparency and reduce trust in their results and make careful use of challenges [10], especially in sensitive fields like the medical domain.

III. METHODOLOGY

A. *Dataset*

For this study, the Sleep-EDF Expanded dataset from PhysioNet was used. This dataset includes polysomnography (PSG) recordings with corresponding hypnogram (HYP) sleep stage annotations. 153 PSG recordings and their associated hypno gram files were used. The PSG recordings consist of multiple channels, including EEG and an event marker. The annotations of sleep stages relative to the PSG data are included the hypnogram file. These annotations include sleep stages for Waken, 1, 2, 3, 4, Movement time, and? (Not scored). This study focused only on the stages W, 1, 2, 3, and 4

B. *Preprocessing*

Signals between 0.5–30 Hz were filtered using Band pass filter. This removes noise and preserve relevant sleep-related EEG frequencies. Data was resampled to a 126 Hz sam pling rate to standardize above all recordings. Signals were segmented into epochs based on the event markers in the hypnogram annotations. 30-second epochs aligned with standard sleep scoring guidelines were extracted. Then compute the Power Spectral Density for each epoch. The resulting PSD features were used as the input representation for the model rather than raw time series, capturing the frequency domain characteristics associated with different sleep stages. A separate test portion of the original dataset, approximately 20%, was reserved for testing. The remaining data was used to train the model. Previous studies have demonstrated that the Synthetic Minority Over-sampling Technique (SMOTE) outperforms one more sampling method in balancing lesson distributions for sleep stages, thus, SMOTE was applied to the training arrangement. Additionally, the Jittering method was implemented to augment the training info by introducing tiny random noise to the EEG epochs. Research indicates that jittering improves model robustness to the two multiplicative and additive noise, leading to enhanced overall performance and generalizability. The augmented data were combined with the original dataset and corresponding labels were duplicated to preserve label integrity.

C. *Proposed Deep Architecture*

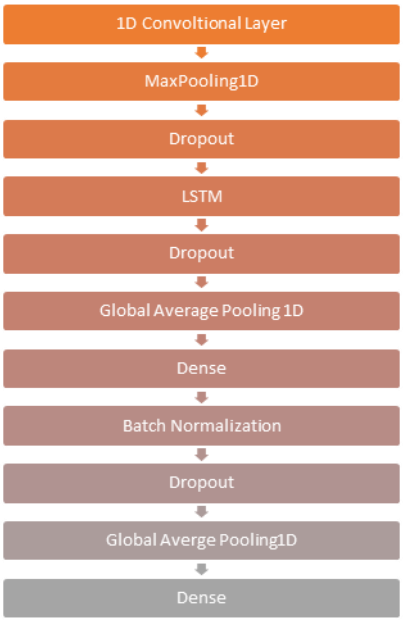

Fig 1: Proposed Deep Learning architecture for sleep stage classification

A deep neural network fusing CNNs and Long Short-Term Memory (LSTM) networks were designed for sleep stage classification. Power Spectral Density (PSD) features extracted from each EEG epoch served as the model input. Each input sample was reformed to match the required dimensions of the convolutional layer. The CNN Layer with 64 neurons was used to retrieve spatial features from the input. ReLU function was applied to capture the nonlinear behavior. Max Pooling Layer was used to down sample the feature maps. Two LSTM layers were included to model the temporal dependencies across the EEG epochs.

- LSTMlayer-128 units and returned sequences to the next LSTM layer.
- LSTM layer-64 units and returned sequences.

After each LSTM layer, a Dropout layer using 0.3 dropout rate was applied. Dense layers with 128 and 64 neurons are stacked. Further, Batch Normalization and Dropout (0.3) layers combines into the models as shown in Fig1. The final output layer was added fully connected layer with 4 neurons and SoftMax activation to perform the classification for multiple classes. “Adam optimizer” with “Categorical Cross entropy” was during the model training. Early Stopping was used during training, monitoring the validation to prevent model learning the training data overly. This executed for 50 epochs using 32 batch size.

D. *Evaluation*

The performance of the model was tested based on predictions generated from the test dataset. Accuracy indicates the number of predictions that correctly predicted out of all the predictions made.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The amount of true positive predictions compared to every positive prediction, as a percentage is measured in precision.

$$Precision = \frac{\text{TP}}{\text{TP+ FP}} \quad (2)$$

Recall is a measure that calculates the true positive rate that were accurately labeled. This indicates the model's sensitivity to detecting correctly predicted stages.

$$Recall = \frac{\text{TP}}{\text{TP+ FN}} \quad (3)$$

F1-Score provides a balanced representation between precision and recall by accounting both Type I and Type II error.

$$F1 - Score = \frac{\text{2 x Precision x Recall}}{\text{Precision+Recall}} \quad (4)$$

E. Explainable Artificial Intelligence

The following section outlines the explainability techniques applied to the model, which facilitate the identification of specific features influencing sleep stage classification.

1) Permutation Feature Importance (PFI): The labels of the test set were predicted by using the trained CNN and LSTM. The baseline accuracy was computed by comparing the model's predictions with the labels using the accuracy. A copy of the test data was made, and the values of the selected feature across all samples were randomly shuffled. This process disrupts the original information carried by that feature while preserving the distribution of other features. The model then made predictions on the shuffled dataset, and the new accuracy was calculated. The importance score for each feature was determined as the difference between the baseline accuracy and the accuracy after shuffling.

2) Saliency Map Analysis: To improve the interpretability of the CNN and LSTM hybrid model, saliency maps were employed. These maps highlight the areas of the input data that have the highest effect on the model's predictions. Saliency maps are computed by calculating the gradient of the predicted class score y^c with respect to the input features X [11].

$$S=\frac{\partial y^c}{\partial X} \quad (5)$$

For selected input samples, saliency map was computed by performing the following step. A sample input from the test set selected. Then computed the gradients of the predicted class score respect to the input feature sample selected. Those gradient values represent the impact of each input to the class prediction. The absolute value of these gradients was averaged over the time steps and produce the saliency map. The saliency maps provided an explanation of the model's behaviour. It showed which specific temporal regions of the EEG signals were most critical for accurate predictions. It highlights the time steps with the highest saliency.

3) Grad-CAM Analysis: Gradient-weighted Class Activation Mapping (Grad-CAM) was implemented to further enhance the explainability of the proposed model. Grad-CAM identifies the regions of the input that have the greatest influence on a specific prediction by computing the gradients of the class score y^c with respect to the activation maps Ak of a selected convolutional layer [11].

$$\frac{\partial y^c}{\partial A^k} \quad (6)$$

A new model is created that outputs both activations of the last convolutional layer and the final model output. The gradients were computed with respect to the class prediction and the activations in the final convolutional layer. These gradients represent the importance of each activation in the convolutional layer for the predicted class. The gradients are pooled across the spatial dimensions, and the activations are weighted by the pooled gradients a_k^c , where i and j index the spatial locations and Z is the total number of spatial locations [11].

$$a_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

The weighted sum of the feature maps generates the class discriminative heatmap. This heatmap shows the spatial localities of the original signal that contributed largely to the prediction given by the model. The normalized heatmap to a [0,1] scale has scale.

$$L_{Grad-CAM}^c = ReLU(\sum_k a_k^c \, A^k) \quad (8)$$

IV. RESULTS AND DISCUSSION

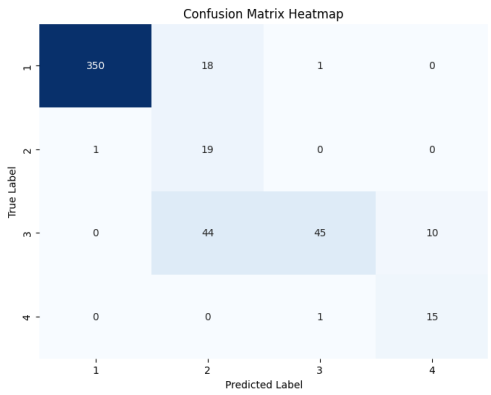

Fig 2: Confusion Matrix of the Test dataset

The Fig 3 represents that, the model's score of accuracy increases during training, as the loss decreases. The model achieves an accuracy of 0.85 for the quiz arrange. As presented in Table 1, the performance scores are too tall. The model demonstrates powerful F1-scores above all classes, with an overall average F1-score close to 80%. The confusion matrix in 2 further presents that only a few data samples were misclassified, while the majority were correctly predicted.

Fig 4 shows the permutation scores across time steps. In the first 150 ms, the permutation feature importance exhibited both positive and negative peaks. This indicates that shuffling certain time steps sometimes improved and sometimes degraded model performance.

This pattern suggests that during the early time window, the EEG features carried inconsistent or mixed information related to the classification task. Some features within this period were informative, while others might have introduced noise. After 150 ms, the permutation importance values remained stable and close to zero.

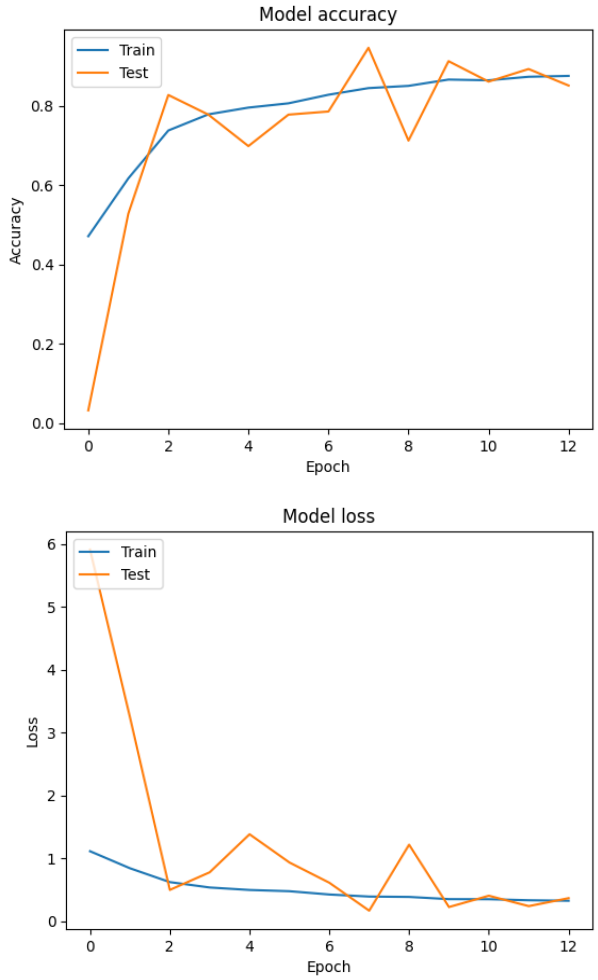

Fig 3: Accuracy and Loss during Training and the Testing

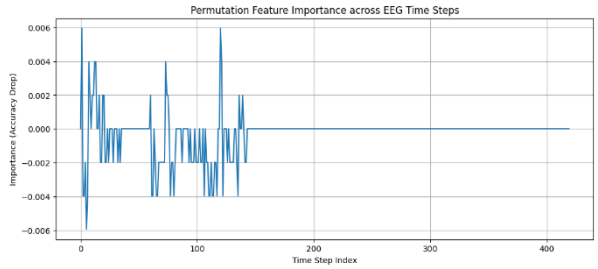

Fig 4: Results of the Permutation Feature Importance

This shows that later time steps contributed little to the model's choices. The analysis presents the temporal sensitivity of the trained CNN LSTM model and provides insights into which parts of the EEG signals were most informative for the classification. In the 6, for Sleep Stage 1, the model focuses on the latter part of the signal. Whereas, Sleep Stage 2, the most critical region appears to be the middle time steps as given in the Fig 7. At the same time, In Sleep Stage 3 represent in Fig 8, the regions of importance are similar those seen in Sleep Stage 2. This shows that both stages have a similar temporal focus.

However, for Deep Sleep, the model identifies regions with higher amplitudes (Fig 9) indicating that these areas are most influential in determining the prediction. Grad-CAM may identify important regions that impact to the model's decision over a longer temporal range, even if those regions do not have pronounced peaks. Saliency maps show the parts of the input that have the major influence on the model's prediction. In EEG signals, they usually highlight areas with quick changes or sharp peaks, which could represent short events like sleep spindles. Together, these interpretability techniques provided how the CNN-LSTM model use the EEG signal to classify different sleep stages.

These findings provide commentaries on the model's decision-making and provide an understanding of the underlying physiological features that are relevant for sleep stage prediction.

TABLE 1: *CLASSIFICATION PERFORMANCE*

Stage	Precision(P)	Recall(R)	F1-score(F1)
W	1.00	0.95	0.97
Stage 1	0.60	0.95	0.73
Stage 2	0.96	0.55	0.69
Stage 3/4	0.60	0.94	0.73
Macro Average Score	0.70	0.82	0.75
Weighted Average Score	0.95	0.85	0.87

V. CONCLUSION

In this study, several model interpretability techniques were applied, including Saliency Maps, Grad-CAM, and Permutation Feature Importance, to explain how the CNN-LSTM model learned to classify sleep stages using EEG signals. These methods helped identify the key time index in the EEG input data that the model relied on for its predictions.

The Saliency Maps revealed the importance of specific time steps in the EEG signal, while Grad-CAM confirmed these observations by highlighting the areas that most strongly contributed. to the model's output, particularly the high-amplitude regions linked to deep sleep. The Permutation Feature Importance analysis further showed that the model's accuracy was highly dependent on early signal fluctuations, emphasizing the impact of disrupting key features on overall performance.

These results provide explanations on the model's behaviour and reinforce the importance of physiological features in sleep stage prediction. In the future other varieties of XAI techniques such as model-agnostic XAI methods can be investigated. Further, this can be extended for multimodal data classification such as integrating EEG signals with other physiological measures. Moreover, this can be extended across diverse datasets and architectures

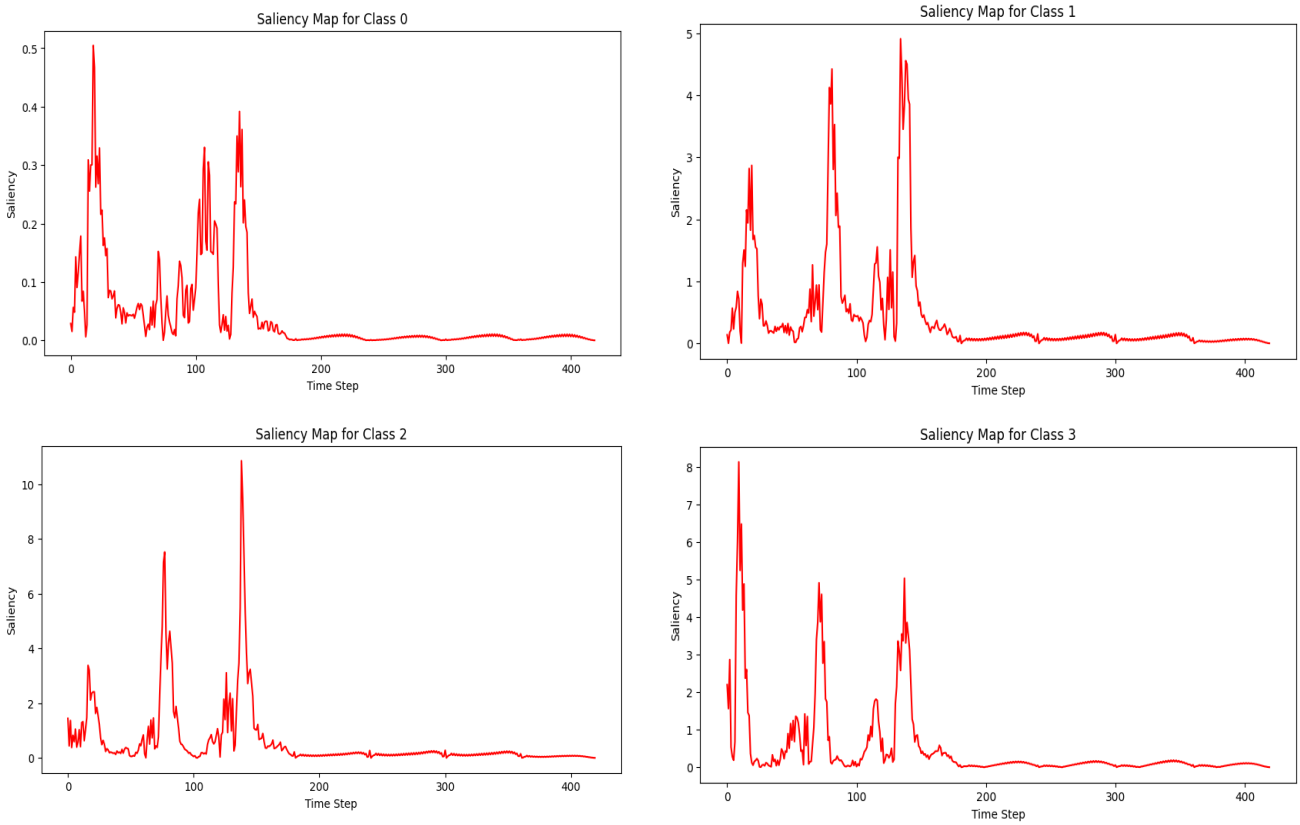

Fig 5: Results of the Saliency Maps for four samples for four different classes

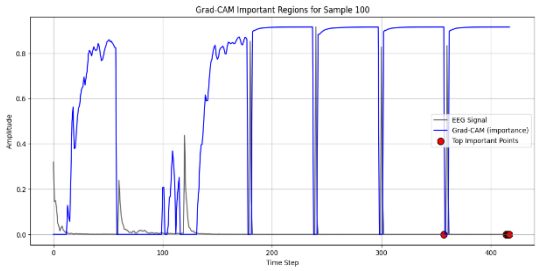

Fig 6: Results of the Grad-CAM

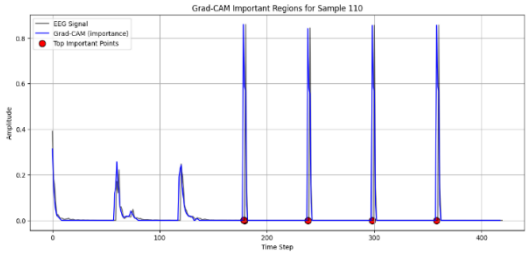

Fig 7: Results of the Grad-CAM

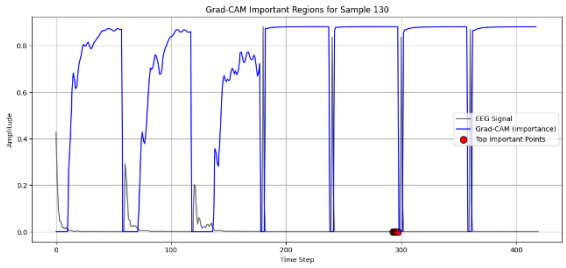

Fig 8: Results of the Grad-CAM

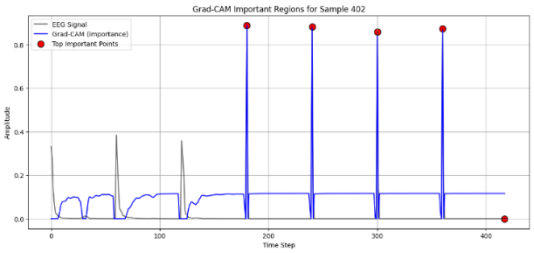

Fig 9: Results of the Grad-CAM

REFERENCES

[1] R. Chatterjee, A. Datta, and D. K. Sanyal, "Ensemble Learning Approach to Motor Imagery EEG Signal Classification," *Elsevier eBooks*, pp. 183–208, Jan. 2019, doi: https://doi.org/10.1016/b978-0-12-816086-2.00008-4.

[2] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, "Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal," *Sensors*, vol. 22, no. 24, p. 9859, Dec. 2022, doi: https://doi.org/10.3390/s22249859.

[3] A. Salami, J. Andreu-Perez, and H. Gillmeister, "EEG-ITNet: An Explainable Inception Temporal Convolutional Network for Motor Imagery Classification," *IEEE Access*, vol. 10, pp. 36672–36685, 2022

[4] M. Nauta *et al.*, "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," Feb. 2023, doi: https://doi.org/10.1145/3583558.

[5] N. Noor and I. Prova, "Explainable AI-Powered Multimodal Fusion Framework for EEG-Based Autism Spectrum Disorder Classification, *SSRN.com*, 2025. https://ssrn.com/abstract=5114986 (accessed Nov. 17, 2025).

[6] D. Raab, A. Theissler, and M. Spiliopoulou, "XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in

EEG time series," *Neural Computing and Applications*, vol. 35, no. 14, pp. 10051–10068, Sep. 2022.

[7] A. Asif, M. Majid, and S. M. Anwar, "Human stress classification using EEG signals in response to music tracks," *Computers in Biology and Medicine*, vol. 107, pp. 182–196, Apr. 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.02.015.

[8] C. Sotirakis *et al.*, "Identification of motor progression in Parkinson's disease using wearable sensors and machine learning," *npj Parkinson's Disease*, vol. 9, no. 1, pp. 1–8, Oct. 2023, doi: https://doi.org/10.1038/s41531-023-00581-2.

[9] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul. 2018, doi: https://doi.org/10.1088/1741-2552/aace8c.

[10] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, "EEGformer: A transformer–based brain activity classification method using EEG signal," *Frontiers in Neuroscience*, vol. 17, p. 1148855, Mar. 2023.

[11] Christoph Molnar, "Interpretable Machine Learning," *Github.io*, Aug. 27, 2019.