Exploring the Applications of XAI on Multimodal Medical Modeling

Koushani Chakrabarty

June 2024

Multimodality and XAI

Multimodal modeling in medical machine learning refers to the integration and analysis of diverse types of data—such as clinical notes, imaging, genomics, and lab results—to enhance diagnostic accuracy and treatment efficacy. This approach leverages the strengths of different data modalities to provide a comprehensive view of patient health, enabling more nuanced and effective medical decision-making. Explainable AI (XAI) plays a crucial role in this context by making the outcomes of these complex, multimodal models transparent and understandable to clinicians. By elucidating how different types of data are weighed and combined to arrive at conclusions, XAI can help healthcare professionals trust and effectively integrate AI insights into their clinical practice. This transparency is vital for verifying the reliability and validity of AI-assisted decisions, ensuring they align with medical standards and contribute positively to patient care. In turn, XAI can facilitate more personalized treatment plans and improve patient outcomes by providing clear, understandable explanations of AI-driven recommendations.

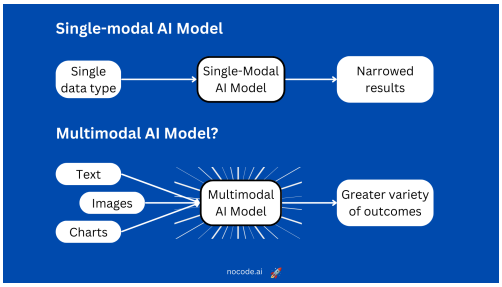


Figure 1: Multimodality

Explainable AI or XAI, involves making the outcomes of AI systems understandable to humans. This is crucial especially for complex models like deep learning, where the decision-making process can be opaque or a "black box." This is especially important for complex machine learning models whose operations can be opaque. Explainable AI aims to clarify how decisions are made, enhancing trust and facilitating easier validation of the AI's behavior. This transparency is crucial not only for building trust among users but also for complying with regulatory requirements, improving model reliability, and ensuring fairness by detecting and mitigating biases in AI decisions. As AI systems

increasingly impact various aspects of society, the importance of explainability continues to grow, pushing for developments in methods that can articulate the reasoning behind AI-generated outcomes.

(XAI) is particularly significant in medical modeling due to its potential to enhance trust, transparency, and accountability in healthcare decisions influenced by AI. In medical settings, where decisions can significantly impact patient outcomes, understanding how AI models derive their conclusions is crucial for clinicians and patients alike. XAI facilitates the integration of AI tools by allowing healthcare professionals to assess the rationale behind AI-generated diagnoses or treatment recommendations, ensuring these tools align with clinical reasoning and ethical standards. Moreover, explainability supports compliance with healthcare regulations that require justification of medical decisions. Ultimately, XAI contributes to safer and more effective implementation of AI technologies in healthcare, promoting personalized medicine and improving patient trust and outcomes by making complex AI decisions understandable and reviewable.

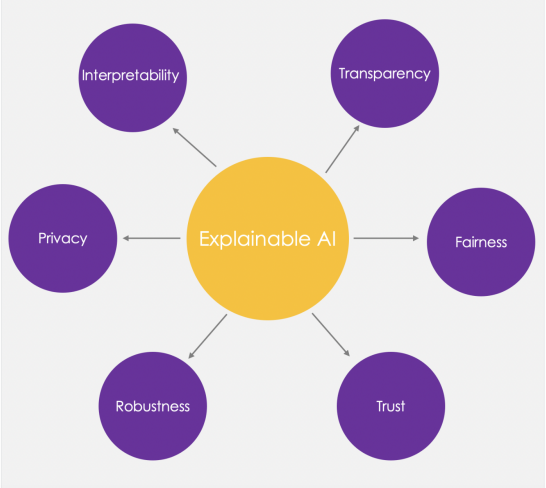


Figure 2: Advantages of using XAI tools

XAI and multimodal classification models are particularly well-suited for handling complex time series data such as EEG and spectrograms in medical modeling. These models excel by integrating multiple data types, enhancing the robustness and accuracy of medical diagnostics. XAI plays a pivotal role by making these sophisticated models transparent, allowing healthcare professionals to see and understand how conclusions are derived from time series data.We believe that combining EEG signals with spectrogram data in a multimodal approach can more accurately pinpoint seizure activities or other neurological abnormalities. XAI adds value by providing clear explanations of the AI's decision-making process, ensuring that these complex analyses are accessible and justifiable to clinicians. This synergy of multimodal modeling with XAI

not only pushes the boundaries of what is achievable in medical diagnostics but also builds trust and facilitates wider adoption of AI technologies in clinical settings.

Dataset and Models

The dataset used for this project has been obtained from Kaggle, as part of the Harmful Brain Pattern Identification Competition. this dataset comprises electroencephalography (EEG) signals recorded from critically ill hospital patients, aimed at detecting and classifying seizures and other types of harmful brain activity. This task is challenging even for experts, who often disagree on the correct labels. The dataset includes metadata files as csv files which provide detailed information about the EEG recordings and their corresponding spectrograms. Each entry in train.csv includes identifiers for the EEG and spectrogram recordings, offsets for the annotated segments, patient IDs, and consensus labels from expert annotators, as well as votes for specific brain activity classes such as seizure, lateralized periodic discharges (LPD), generalized periodic discharges (GPD), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), and "other."

The EEG data is stored in the train and test directories, with each file representing 50-second long samples recorded at a frequency of 200 samples per second. The spectrogram data, assembled from the EEG recordings, is stored in the train and test directories, covering 10-minute windows centered around the same time as the EEG samples. The dataset also includes example figures and a sample submission file to guide the model development process. The project emphasizes the need for developing robust models that can accurately classify these patterns to assist in faster and more reliable diagnosis and treatment, ultimately contributing to improved neurocritical care, epilepsy management, and drug development efforts.

The model summary in Figure 3 provides a comprehensive overview of the multimodal model architecture designed for classifying EEG signals and their corresponding spectrograms. The model consists of two primary components: an EEGNet and a Spectrogram Model. The EEGNet is responsible for processing 1D time-series EEG data and comprises multiple convolutional, batch normalization, and dropout layers, culminating in a dense layer for classification. The Spectrogram Model processes 2D spectrogram images using a series of convolutional blocks, each with convolutional, batch normalization, and pooling layers, followed by global average pooling and fully connected layers. The outputs of these two models are concatenated and passed through additional dense layers to produce the final classification. The model summary highlights the detailed layer configurations, output shapes, and parameter counts, emphasizing the complexity and depth of the network, which consists of over 2 million trainable parameters. This detailed breakdown is crucial for understanding the computational requirements and the potential areas for optimization in the model.

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 8, 37, 3000]	512
BatchNorm2d-2	[-1, 8, 37, 3000]	16
Conv2d-3	[-1, 16, 1, 3000]	592
BatchNorm2d-4	[-1, 16, 1, 3000]	32
ELU-5	[-1, 16, 1, 3000]	0
AvgPool2d-6	[-1, 16, 1, 750]	0
Dropout-7	[-1, 16, 1, 750]	0
Conv2d-8	[-1, 16, 1, 750]	4,096
BatchNorm2d-9	[-1, 16, 1, 750]	32
ELU-10	[-1, 16, 1, 750]	0
AvgPool2d-11	[-1, 16, 1, 93]	0
Dropout-12	[-1, 16, 1, 93]	0
Flatten-13	[-1, 1488]	0
Linear-14	[-1, 6]	8,934
LogSoftmax-15	[-1, 6]	0
EEGNet-16	[-1, 6]	0
Conv2d-17	[-1, 16, 400, 300]	448
Conv2d-18	[-1, 16, 400, 300]	2,320
Conv2d-19	[-1, 16, 400, 300]	2,320
MaxPool2d-20	[-1, 16, 200, 150]	0
BatchNorm2d-21	[-1, 16, 200, 150]	32
Dropout-22	[-1, 16, 200, 150]	0
Conv2d-23	[-1, 16, 200, 150]	64
Block-24	[-1, 16, 200, 150]	0
Conv2d-25	[-1, 32, 200, 150]	4,640
Conv2d-26	[-1, 32, 200, 150]	9,248
Conv2d-27	[-1, 32, 200, 150]	9,248
AvgPool2d-28	[-1, 32, 100, 75]	0
BatchNorm2d-29	[-1, 32, 100, 75]	64
Dropout-30	[-1, 32, 100, 75]	0
Conv2d-31	[-1, 32, 100, 75]	544
Block-32	[-1, 32, 100, 75]	0
Conv2d-33	[-1, 64, 100, 75]	18,496
Conv2d-34	[-1, 64, 100, 75]	36,928
Conv2d-35	[-1, 64, 100, 75]	36,928
MaxPool2d-36	[-1, 64, 50, 37]	0
BatchNorm2d-37	[-1, 64, 50, 37]	128
Dropout-38	[-1, 64, 50, 37]	0
Conv2d-39	[-1, 64, 50, 37]	2,112
Block-40	[-1, 64, 50, 37]	0
Conv2d-41	[-1, 128, 50, 37]	73,856
Conv2d-42	[-1, 128, 50, 37]	147,584
Conv2d-43	[-1, 128, 50, 37]	147,584
AvgPool2d-44	[-1, 128, 25, 18]	0
BatchNorm2d-45	[-1, 128, 25, 18]	256
Dropout-46	[-1, 128, 25, 18]	0
Conv2d-47	[-1, 128, 25, 18]	8,320
Block-48	[-1, 128, 25, 18]	0
Conv2d-49	[-1, 256, 25, 18]	295,168
Conv2d-50	[-1, 256, 25, 18]	590,080
Conv2d-51	[-1, 256, 25, 18]	590,080
MaxPool2d-52	[-1, 256, 12, 9]	0
BatchNorm2d-53	[-1, 256, 12, 9]	512
Dropout-54	[-1, 256, 12, 9]	0
Conv2d-55	[-1, 256, 12, 9]	33,024
Block-56	[-1, 256, 12, 9]	0
AdaptiveAvgPool2d-57	[-1, 256, 1, 1]	0
Linear-58	[-1, 6]	1,542
LogSoftmax-59	[-1, 6]	0
Spectrogram_Model-60	[-1, 6]	0
Linear-61	[-1, 128]	1,664
Linear-62	[-1, 6]	774
LogSoftmax-63	[-1, 6]	0

Total params: 2,028,178
Trainable params: 2,028,178
Non-trainable params: 0

Input size (MB): 152435.30
Forward/backward pass size (MB): 135.34
Params size (MB): 7.74
Estimated Total Size (MB): 152578.38

Figure 3: Multimodal Model

Results

XAI and multimodal classification models are particularly well-suited for handling complex time series data such as EEG and spectrograms in medical modeling. These models excel by integrating multiple data types, enhancing the robustness and accuracy of medical diagnostics. XAI plays a pivotal role by making these sophisticated models transparent, allowing healthcare professionals to see and understand how conclusions are derived from time series data.We believe that combining EEG signals with spectrogram data in a multimodal approach can more accurately pinpoint seizure activities or other neurological abnormalities. XAI adds value by providing clear explanations of the AI's decision-making process, ensuring that these complex analyses are accessible

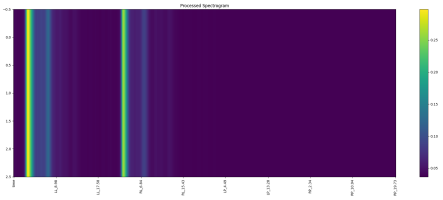


Figure 4: Processed Spectrogram input sample

and justifiable to clinicians. This synergy of multimodal modeling with XAI not only pushes the boundaries of what is achievable in medical diagnostics but also builds trust and facilitates wider adoption of AI technologies in clinical settings.

From the model performance analysis presented in the plots, we can derive several insights. For example for the EEG model, the training loss decreases consistently, indicating that the model is learning the training data well.The validation loss remains relatively stable after an initial drop, suggesting that the model might be overfitting to the training data since it is not improving much on the validation set.The training accuracy increases steadily and reaches around 70 percent . The validation accuracy improves initially but plateaus around 50 percent, which further suggests overfitting. The model is not generalizing well to the unseen validation data.

In case of the Spectrogram model , Both training and validation loss decrease over epochs, with the training loss showing a smoother trend.The validation loss shows more fluctuation but follows a general downward trend, indicating the model is learning and improving on the validation set.The validation loss shows more fluctuation but follows a general downward trend, indicating the model is learning and improving on the validation set.Validation accuracy fluctuates but shows an overall upward trend, indicating that the model is improving its performance on the validation set, though with some instability. For the Multimodal model, Training loss decreases consistently, indicating the model is learning the training data well.Validation loss decreases significantly over epochs and shows a clear improvement, suggesting the combined model generalizes better than the individual models.Training accuracy improves steadily and reaches around 90 percent.Validation accuracy shows a strong upward trend and reaches close to 90 per cent, indicating the combined model is performing very well on both training and validation data.

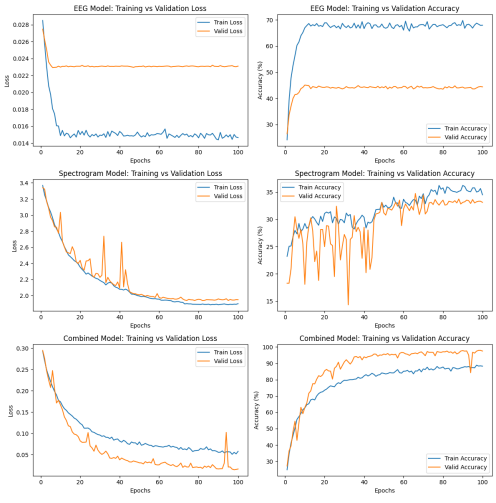


Figure 5: Model performance comparison

Thus in conclusion: The EEG model shows signs of overfitting, where it performs well on the training data but not as well on the validation data. This indicates that the model is capturing the noise in the training data rather than general patterns.The Spectrogram model shows improvement on both training and validation sets, although with some fluctuations in validation accuracy. This might be due to the variability in the spectrogram data or the model architecture.The combined model performs significantly better than

the individual models, both in terms of loss and accuracy. This indicates that combining EEG and spectrogram data leads to a more robust model that generalizes better to unseen data.

LIME

Local Interpretable Model-Agnostic Explanations (LIME) is a method developed to improve the interpretability of complex machine learning models. As machine learning models become more sophisticated, their decision-making processes often become opaque, leading to concerns about transparency and trust. LIME addresses this challenge by offering explanations for individual predictions, making it easier for users to understand how specific inputs lead to certain outputs. LIME operates on the principle that it is easier to approximate

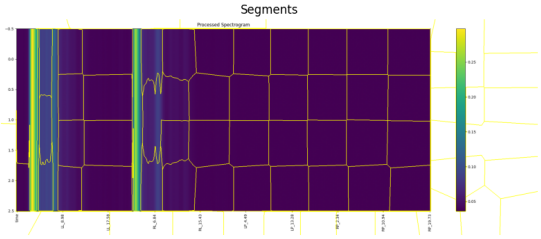


Figure 6: Segmentation of Spectrogram input for LIME

a complex model locally than globally. The method generates explanations by perturbing the input data and observing the resulting changes in the predictions. This perturbation involves creating a new dataset by slightly altering the original data points, and then fitting a simple, interpretable model (like a linear regression) to these perturbed samples. This interpretable model approximates the complex model's behavior in the vicinity of the specific instance being explained.One of the key advantages of LIME is its model-agnostic nature, meaning it can be applied to any machine learning model regardless of its complexity. This flexibility is particularly valuable in a landscape where diverse models are employed across various applications.

In this experiment, LIME was applied to the Spectrogram input to observe the impact of different aspects of the input on the model's decision making. First a segmentation algorithm has been used to visualize the input image with its segments highlighted (Figure 6). The segments are shown as colored boundaries on top of the original image. This helps to understand how an image has been divided into different regions by the segmentation algorithm.

Next, the effect of using LIME on the Unimodal Spectrogram Model are plotted (Figure 7). this visualization shows the segments of the spectrogram that are most influential in the unimodal model's decision-making process. The highlighted segments indicate the areas where the model's attention is focused when predicting the top label.

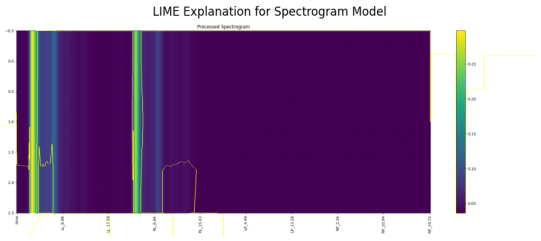


Figure 7: LIME on Unimodal Spectrogram

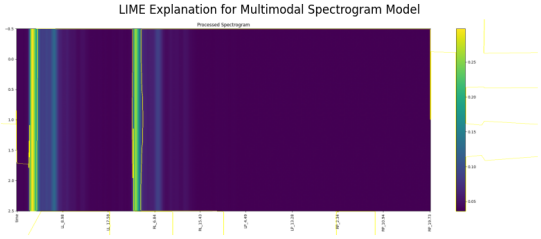


Figure 8: LIME on Multimodal Spectrogram

Finally , LIME was used to observe the relevant portions of the input spectrogram image, as viewed by the multimodal spectrogram model (Figure 8). It was notable that while the unimodal model, considered some lighter parts of the input image for it's decision making, the multimodal model concentrated on a smaller and far more relevant area of the input image

SHAP

Shapley Additive Explanations (SHAP) is a unified framework for interpreting machine learning models, grounded in cooperative game theory. Developed by Scott Lundberg and Su-In Lee, SHAP leverages the concept of Shapley values to provide a consistent and theoretically sound approach to understanding the contributions of individual features to a model's predictions.Shapley values, named after Lloyd Shapley, are a solution concept in cooperative game theory that distribute the total gain generated by the coalition of all players fairly among them, based on their individual contributions. When applied to machine learning, SHAP values quantify the contribution of each feature to the difference between the actual prediction and the average prediction over the entire dataset.One of the primary strengths of SHAP is its axiomatic foundation, which ensures several desirable properties for explanations. These properties include local accuracy (the sum of SHAP values equals the model's prediction), consistency (if a feature's contribution increases in one model, it does not decrease in another), and missingness (features with zero contribution have zero SHAP value). These properties make SHAP a robust and reliable method for

model interpretation.

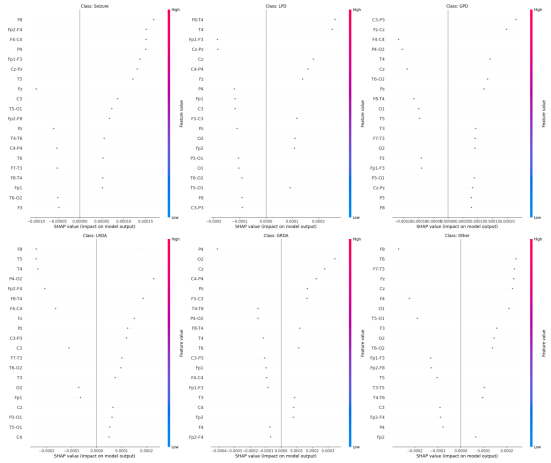


Figure 9: SHAP on Unimodal EEG model

Figure 9 plot shows the SHAP values for different EEG channels (features) for a specific class. The SHAP value represents the impact of that feature on the model's output for the given class.Features with higher SHAP values (further from zero) are more influential in the model's decision-making process for that class.SHAP values can be positive or negative, indicating whether the feature contributes positively or negatively to the prediction of the class. For example, in the "Class: Seizure" plot, features like F8 and Fp2-F4 have higher positive SHAP values, meaning these channels strongly contribute to predicting seizures.The color bar on the right of each plot indicates the feature value. Blue represents low feature values, and red represents high feature values.This helps in understanding how the feature value influences the SHAP value. For instance, in the "Class: LPD" plot, high values for the feature T4 correlate with positive SHAP values.Each class has a different set of influential features. For instance, "Class: GPD" shows that features like C3-P3 and F2-Cz have significant SHAP values, indicating their importance in predicting GPD.

The final plot in SHAP (Figure 10), captures the impact of Multimodality on SHAP values of the EEG model. Comparative SHAP analysis on the different modalities of the Multimodal model to unerstand their comparative impact on the Model's decision making

Saliency Maps

Saliency maps are a visualization technique used to interpret and understand the decision-making processes of deep learning models, particularly in the context of computer vision. They highlight the regions of an input image that are most influential in determining the model's output, providing insights into

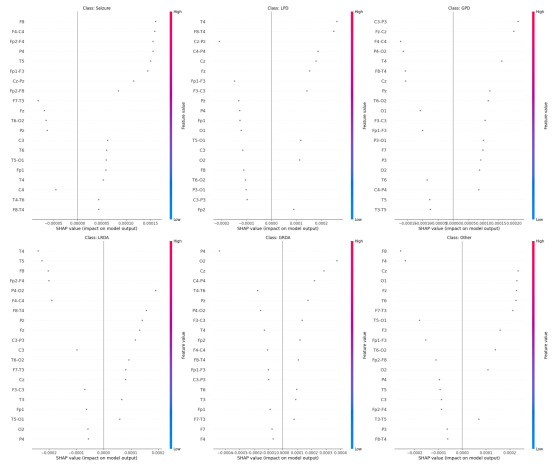


Figure 10: SHAP on Multimodal EEG model

what the model "sees" when making predictions.One of the primary advantages of saliency maps is their ability to provide intuitive visual explanations for complex deep learning models. By highlighting the most relevant regions of an image, saliency maps help users understand what features the model is focusing on, fostering trust and transparency in AI systems. This is particularly valuable in applications such as medical imaging, where understanding the model's decision-making process is critical for clinical validation and acceptance.In addition to their use in computer vision, saliency maps can be adapted for other types of data, such as time series and text. For example, in time series analysis, saliency maps can highlight the most important time points or segments that influence the model's predictions. In natural language processing, saliency maps can identify the most relevant words or phrases in a text.However, saliency maps also have limitations. The interpretation of saliency maps can be subjective, and the visualizations may not always provide clear insights into the model's decision-making process. Additionally, saliency maps can be sensitive to noise and adversarial attacks, potentially leading to misleading explanations. Despite these challenges, saliency maps remain a valuable tool for interpreting and understanding deep learning models, contributing to the broader goal of explainable AI.

The provided saliency map illustrates the regions within the spectrogram that are most influential in the model's decision-making process (Figure 11).The brighter areas in the saliency map indicate regions of the spectrogram that have the highest impact on the model's output. These regions are where the model is focusing most of its attention.In this map, there are noticeable bright bands and spots, suggesting specific time-frequency components are highly influential.The horizontal axis represents time, while the vertical axis represents different frequency components.The saliency map shows that the model is paying attention

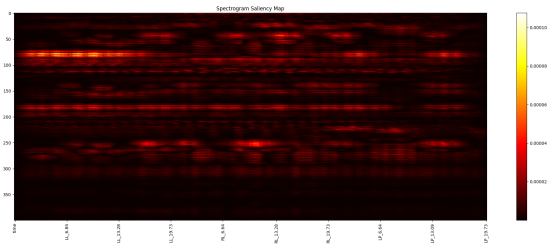


Figure 11: Spectrogram Saliency Map

to certain frequency bands more than others across different time segments.For instance, around the 100th, 200th, and 300th pixel marks on the vertical axis, there are bright bands indicating the model's focus on these frequencies.By analyzing the saliency map, one might identify if the model is focusing on irrelevant regions, which could indicate a need for further tuning or modification of the model or the preprocessing steps.One may compare the regions highlighted by the saliency map with known clinically relevant frequency bands to ensure the model aligns with medical knowledge.

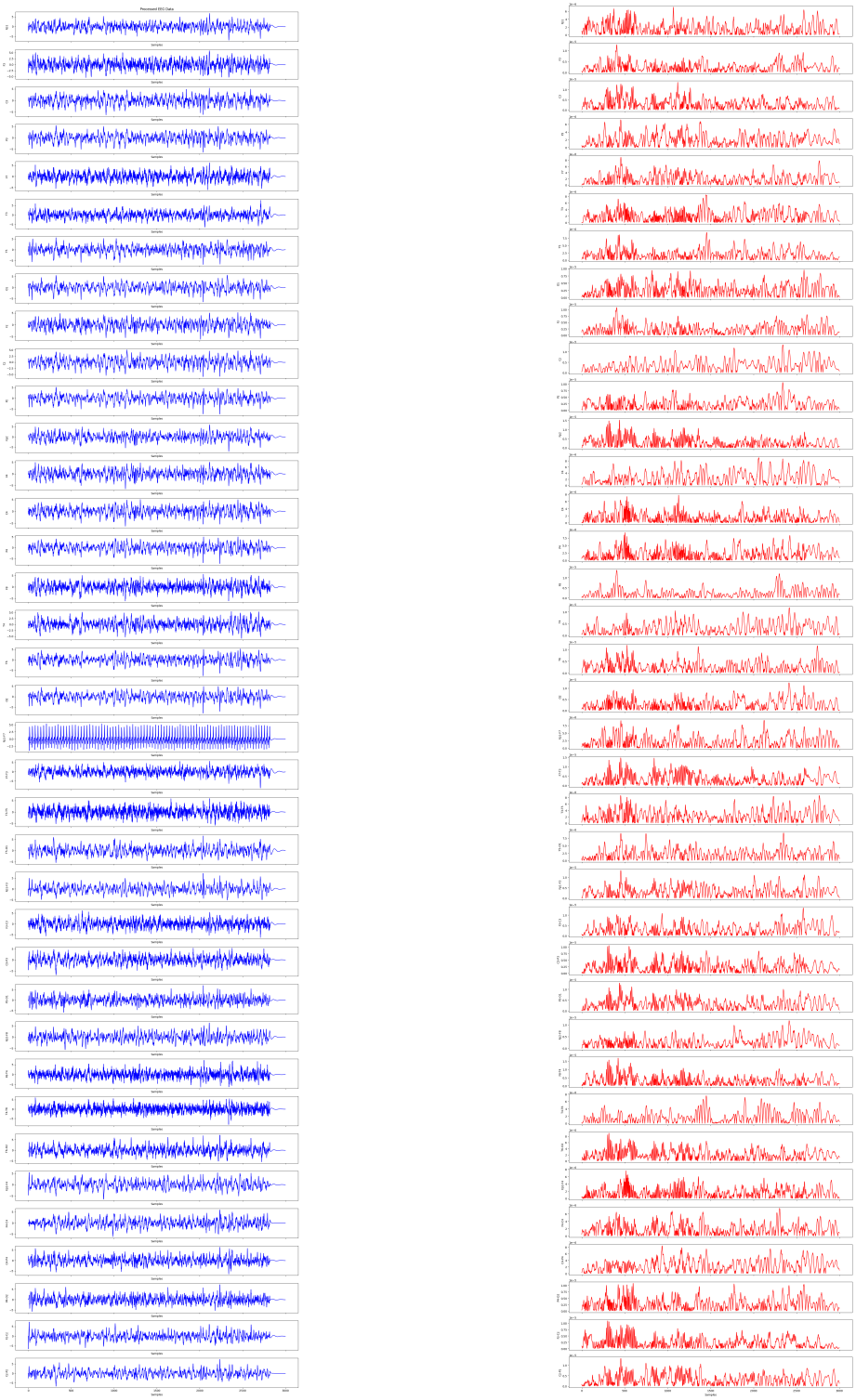
Figure 12 shows two side-by-side plots.The left plot (labeled (a) EEG input) represents the raw EEG signals from various channels.Each subplot corresponds to a different EEG channel, showing the time series data of brain activity.The horizontal axis represents time, and the vertical axis represents the amplitude of the EEG signal.The right plot (labeled (b) EEG Saliency Map) represents the saliency map for the same EEG signals.The saliency map highlights the regions in the EEG signals that have the most significant impact on the model's predictions.In this plot, the intensity of the red color indicates the importance of the signal at that specific time point. Brighter areas (more intense red) indicate higher importance.For instance, if a particular segment in the EEG signal has a high intensity, it indicates that the model considers that segment as important for its classification task.By comparing the saliency maps across different channels, one can determine which EEG channels are more relevant for the model's decisions.If certain channels consistently show higher intensity regions, they might be more crucial for the specific task the model is trained on.

If the model is found to focus on irrelevant or noisy parts of the EEG signals, it may indicate a need for further model refinement or preprocessing adjustments.The saliency map can help identify if certain channels or time segments are being incorrectly emphasized.

Saliency maps help in validating the model's behavior by ensuring it is focusing on relevant parts of the EEG signals.This can be compared with clinical knowledge to see if the important regions identified by the model align with known medically significant patterns.

Conclusion

Saliency maps, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) are powerful tools in the field of explainable AI (XAI). They provide transparency and interpretability to complex machine learning models, especially in the context of medical diagnosis.



(a) EEG input

(b) EEG Saliency Map

Figure 12: EEG Saliency Plot