INDEX

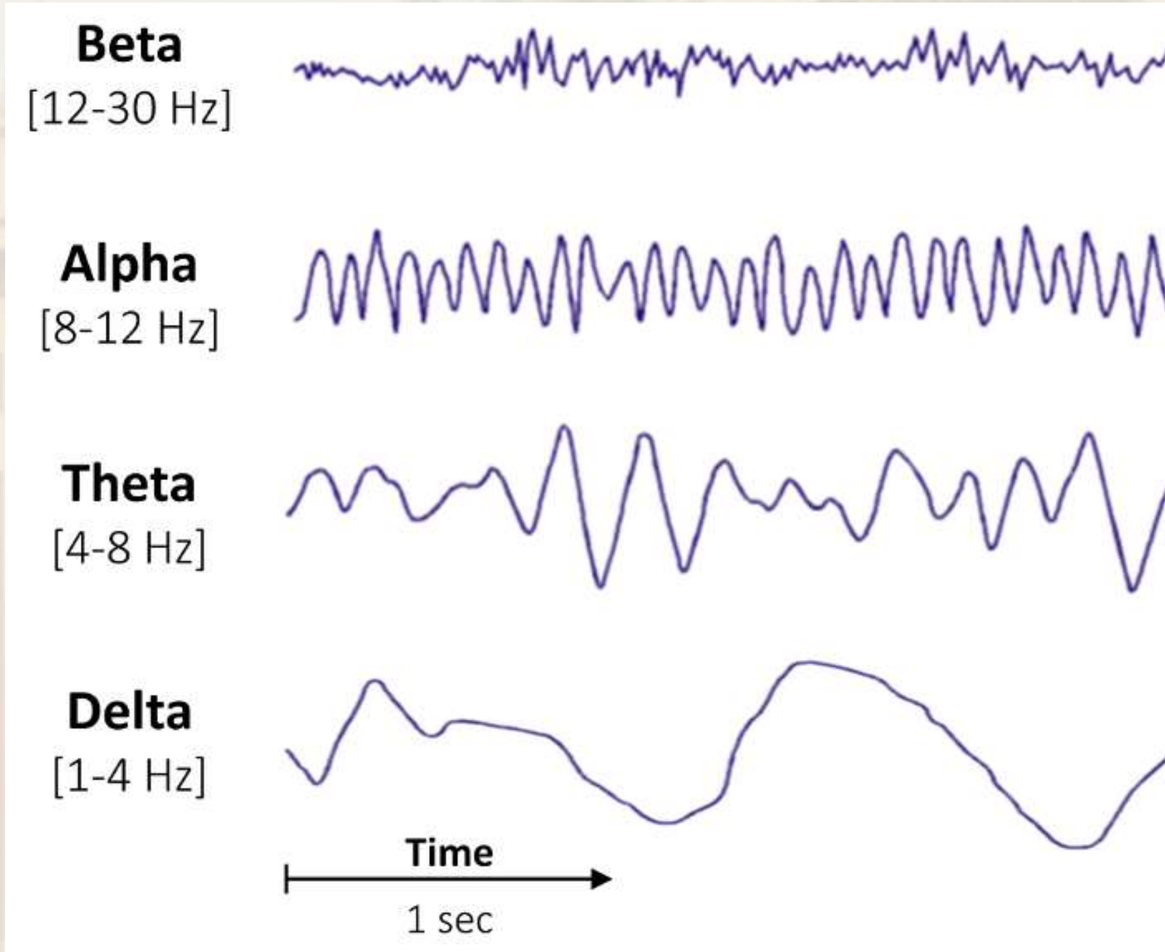
Topic Overview

Neurodisorders like schizophrenia and Alzheimer's require early, accurate diagnosis, but traditional methods are often subjective and inconsistent. AI, especially machine learning, helps classify these disorders by analyzing complex data such as EEG signals. However, most AI models act as black boxes, offering little insight into how decisions are made.

Explainable AI (XAI) solves this by making AI decisions interpretable through tools, which show what features influence predictions. This builds trust with clinicians and supports better decision-making. For example, in classifying schizophrenia, XAI highlights specific EEG patterns responsible for the diagnosis, combining high accuracy with transparency.

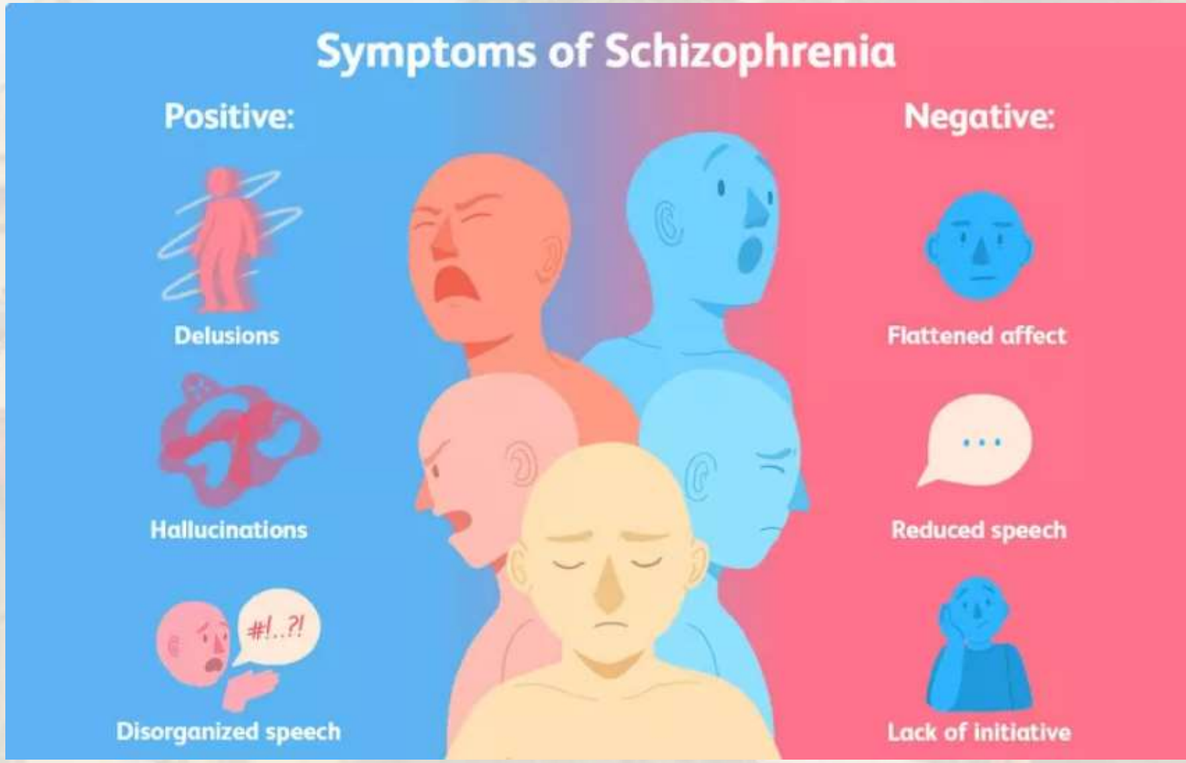
EEG (Electroencephalogram)

EEG (Electroencephalography) is a non-invasive method used to record electrical activity of the brain using sensors placed on the scalp. It helps detect abnormalities in brain function and is commonly used in diagnosing neurological disorders like epilepsy, schizophrenia, and sleep disorders.



Schizophrenia Disease

Schizophrenia is a serious mental health condition that affects how people think, feel and behave. It may result in a mix of hallucinations, delusions, and disorganized thinking and behavior. Hallucinations involve seeing things or hearing voices that aren't observed by others



Objective

The objective of this project is to apply Explainable AI (XAI) models for the classification of neurodisorders, ensuring transparency and interpretability in predictions.

The project aims to:

- Develop and validate different XAI models to enhance the interpretability of machine learning predictions in neurodisorder classification.
- Identify and analyze key contributing features that influence model decisions, providing deeper insights into the neurological patterns associated with the disorder.
- Improve trust and reliability in AI-driven diagnosis by making model predictions more understandable to clinicians and researchers.

Dataset

- The dataset comprised 14 patients with disease and 14 healthy controls.
- Channels (chs): 19 EEG electrodes were used to record brain activity. The listed channels (e.g., Fp2, F8, T4, etc.) correspond to specific scalp locations based on the standard 10-20 EEG system.
- Sampling Frequency (sfreq): 250.0 Hz — meaning 250 data points were recorded per second per channel.
- Total Samples: 231,250 data points were recorded for each of the 19 channels.
- Recording Duration: ~925 seconds (about 15.4 minutes).
- Frequency Range:Highpass: 0.0 Hz ,Lowpass: 125.0 Hz .

```
Extracting EDF parameters from /content/h01.edf...
EDF file detected
Setting channel info structure...
Creating raw.info structure...
Reading 0 ... 231249  =      0.000 ...   924.996 secs...
(<Info | 8 non-empty values
  bads: []
  ch_names: Fp2, F8, T4, T6, O2, Fp1, F7, T3, T5, O1, F4, C4, P4, F3,
  chs: 19 EEG
  custom_ref_applied: False
  highpass: 0.0 Hz
  lowpass: 125.0 Hz
  meas_date: 2003-06-23 16:14:37 UTC
  nchan: 19
  projs: []
  sfreq: 250.0 Hz
  subject_info: 1 item (dict)
 >,
 ['Fp2',
  'F8',
  'T4',
  'T6',
  'O2',
  'Fp1',
  'F7',
  'T3',
  'T5',
  'O1',
  'F4',
  'C4',
  'P4',
  'F3',
  'C3',
  'P3',
  'Fz',
  'Cz',
  'Pz'],
 (19, 231250),
 (231250,))
```

Preprocessing

1.Channel Selection

- we can select only the EEG channels (excluding ECG, EOG, etc., if present).
- In my case, all channels are EEG.

2.Filtering

- High-pass filter (e.g., >0.5 Hz): Removes slow drifts and DC offset.
- Low-pass filter (e.g., <40-50 Hz): Removes muscle noise .
- My current settings are:
  - highpass: 0.0 Hz (you may want to raise this to ~0.5 Hz)
  - lowpass: 125.0 Hz (can reduce to 50 Hz for cleaner signals)Artifact Removal

3.Use methods like ICA (Independent Component Analysis) to remove artifacts such as:

- Eye blinks,Muscle movements,Line noise

4.Epoching (Optional)

- Divide the continuous signal into fixed-length segments (e.g., 25 seconds) for easier processing.

5.Normalization / Standardization

- Normalize the signal values for each channel so that ML models perform better.

Machine Learning Models

1. Logistic Regression
   - Type: Linear classifier.
   - Use: Binary or multi-class classification.
   - How it works: Predicts the probability that a data point belongs to a class using a logistic function.
   - Good for: Simple, linearly separable data.
   - Example use: Classify EEG as seizure vs. non-seizure

2. Support Vector Machine (SVM)
   - Type: Linear or non-linear classifier.
   - Use: Classification and regression.
   - How it works: Finds the best hyperplane that separates classes by the maximum margin.
   - Good for: High-dimensional data and smaller datasets.
   - Kernel Trick: Allows it to work well with non-linear data.

3. Random Forest
   - Type: Ensemble method (bagging).
   - Use: Classification and regression.
   - How it works: Builds multiple decision trees and takes a majority vote or average.
   - Good for: Handling missing values, reducing overfitting, and working with mixed data types.

4. AdaBoost (Adaptive Boosting)
   - Type: Ensemble method (boosting).
   - Use: Mainly classification.
   - How it works: Combines weak learners (usually decision stumps) into a strong classifier by focusing more on previously misclassified samples.
   - Good for: Improving accuracy on challenging data.

DEEP LEARNING MODEL

5.. XGBoost (Extreme Gradient Boosting)

- Type: Advanced boosting algorithm.
- Use: Classification and regression.
- How it works: Builds trees sequentially, with each tree correcting the errors of the previous one.
- Features:
  - Regularization to reduce overfitting
  - Handles missing values
  - Very fast and efficient

LSTM (Long Short-Term Memory)

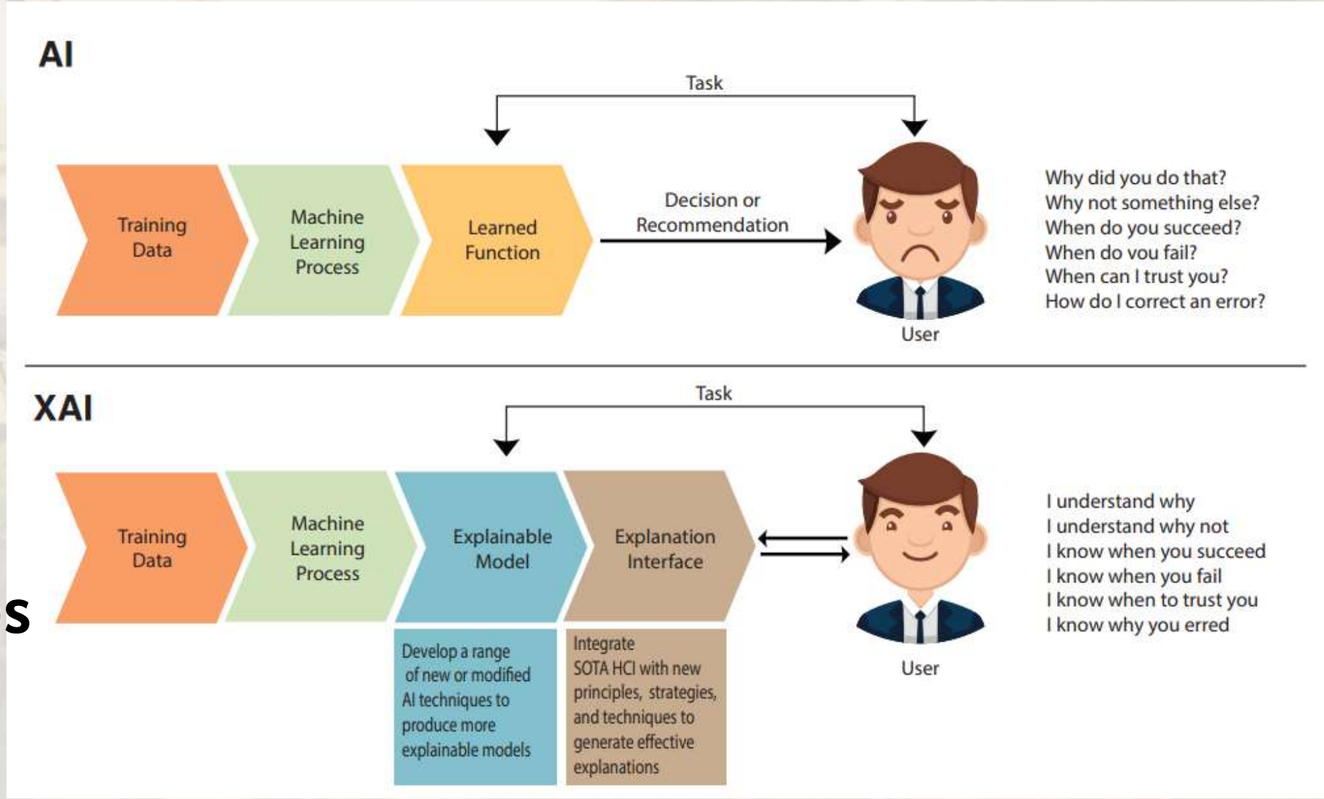
- Type: Special kind of RNN.
- Use: Same as RNN but better for long sequences.
- How it works: Uses memory cells and gates (input, forget, output) to retain long-term dependencies.
- Best for: EEG signal classification, language processing, anomaly detection in time-series.

COMPARISON TABLE:

Model	Non-linear	Handles Time Series	Fast	Accuracy	Overfitting Risk
Logistic Reg.	✗	✗	✓	Moderate	Low
SVM	✓	✗	✗	High	Moderate
Random Forest	✓	✗	✓	High	Low
AdaBoost	✓	✗	✓	High	Medium
XGBoost	✓	✗	✓✓	Very High	Low
RNN	✓	✓	✗	High	High
LSTM	✓✓	✓✓	✗	Very High	Medium

Explainable AI(XAI)

- Explainable AI is set of process and method that allows human user to understand and trust the result and output created by machine learning.
- XAI can explain the ML output and contribution of features in disease predication model
- XAI is used to describe AI model . It helps categorise model accuracy fairness, transparency and outcome in AI power decision making.



XAI Models

1. LIME (Local Interpretable Model-agnostic Explanations)

- Goal: Explain individual predictions.
- How: Approximates the model locally using an interpretable model (like linear regression).
- Use Case: Helps understand why a specific prediction was made (e.g., why a patient was labeled "epileptic").
- Pro: Model-agnostic (works with any ML/DL model).

2. SHAP (SHapley Additive exPlanations)

- Goal: Explain the output of any ML model using game theory.
- How: Calculates each feature's contribution to the final prediction.
- Use Case: Provides both global and local interpretability (important features overall and for one prediction).
- Pro: Theoretically strong and consistent.

3. ELI5

- Goal: Debug and visualize ML models and their predictions.
- How: Works mainly with scikit-learn and XGBoost; provides weights, contributions, and feature importances.
- Use Case: Quick insights into model internals (e.g., weights in logistic regression).
- Pro: Easy to use and integrates well with scikit-learn pipelines.

4. PDP (Partial Dependence Plots)

- Goal: Show how a feature affects the prediction on average.
- How: Marginalizes the output over the dataset by changing one feature at a time.
- Use Case: Global interpretability — understand feature influence on model behavior.
- Pro: Good for visualizing interactions in tree-based models.

5. Integrated Gradients

- Goal: Attribute predictions in deep neural networks (esp. for images and sequences).
- How: Measures the gradients along the path from a baseline to the input.
- Use Case: Interpreting deep models like CNNs or RNNs on EEG, image, or text data.
- Pro: More accurate than vanilla gradients; works well for deep learning.

Comparison Table

Method	Type	Local/Global	Works With	Strength
LIME	Model-agnostic	Local	Any ML/DL	Simple loc
SHAP	Model-agnostic	Both	Any ML/DL	Fair, consi

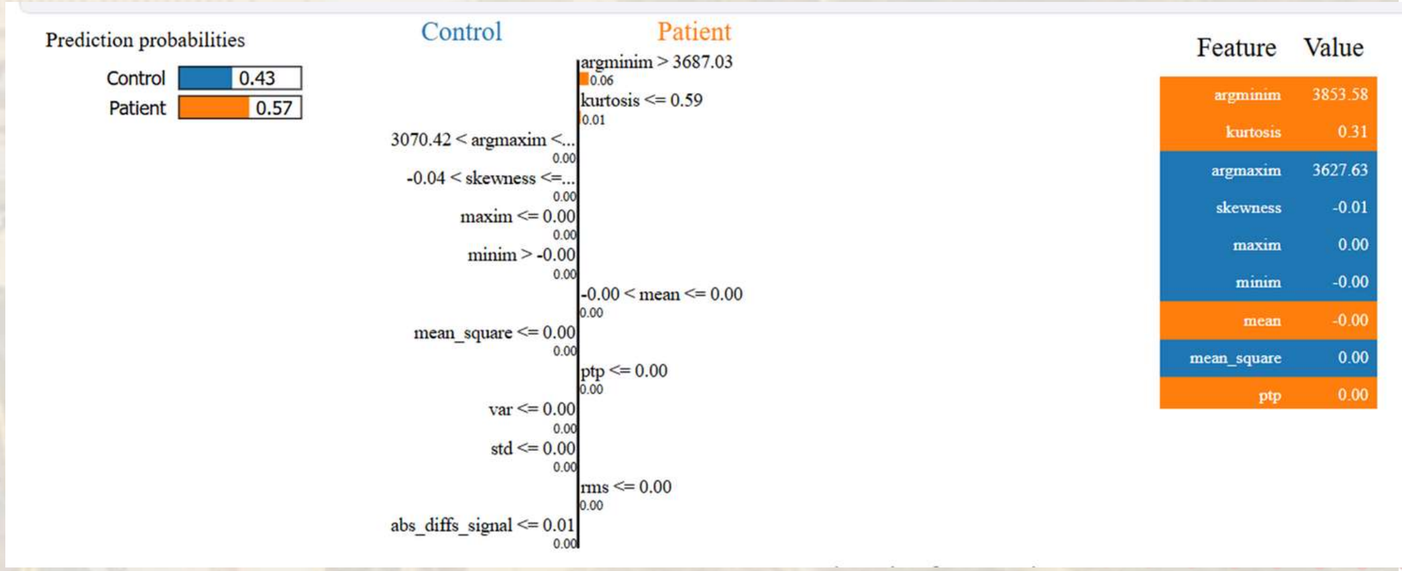
Results

I.Applied explainable AI model over logistic regression

I.Lime

- The image displays a LIME explanation for a single instance.
- The prediction probabilities show a 57% chance that the sample belongs to the Patient class.
- Key contributing features include:
- argminim (Time of Minimum Value in Signal) – This positively impacts the prediction.
- kurtosis (Sharpness of the Signal Distribution) – A lower kurtosis value contributes to the Patient class.

argmaxim (Time of Maximum Value in Signal) – This feature also played a role in distinguishing between classes



II. Shap

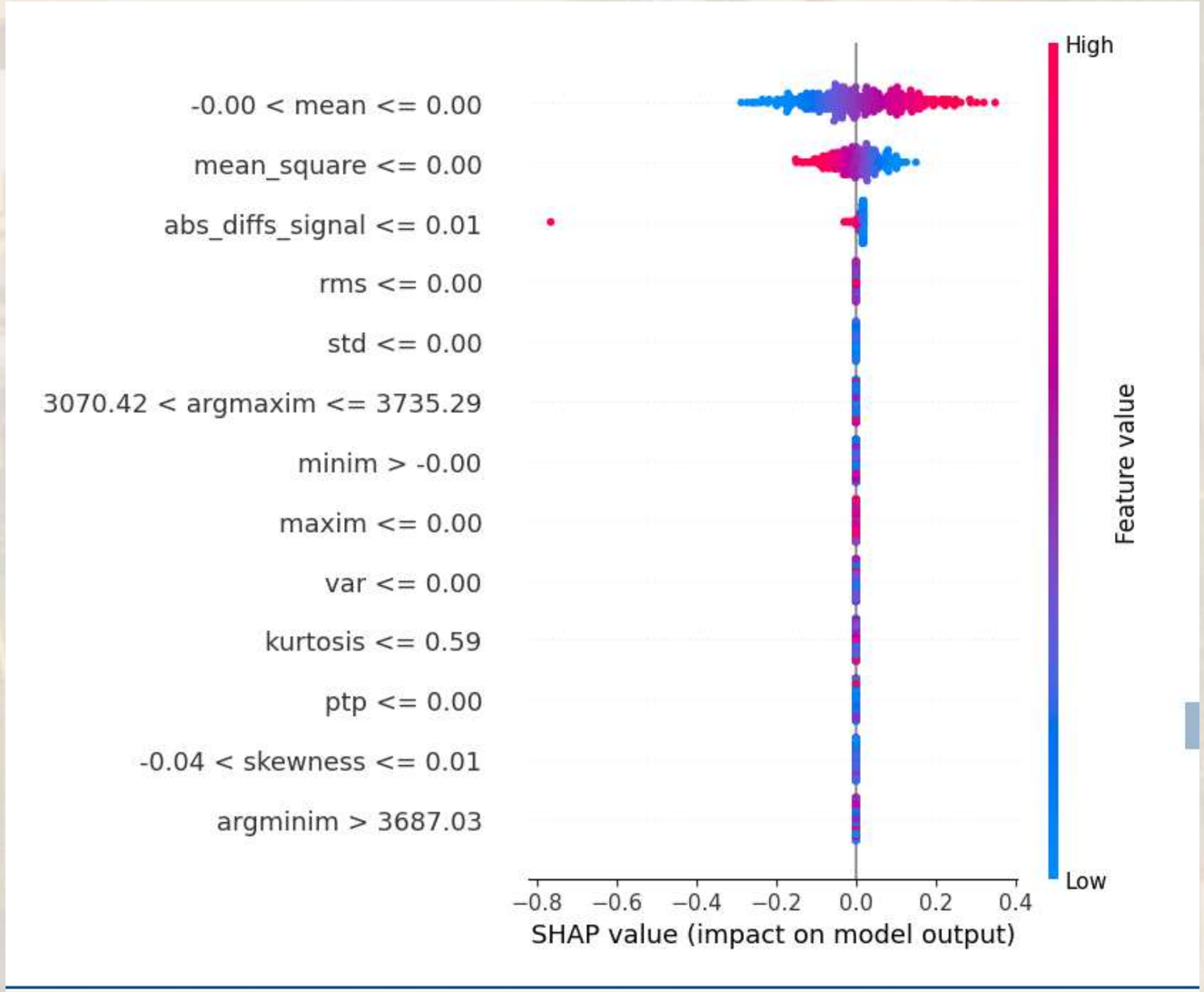
SHAP Analysis

- Features with a higher absolute SHAP value contribute significantly to predictions.

- Key takeaways : argmaxim, argminim, and kurtosis are among the most influential features.

- Blue indicates lower feature values, while red indicates higher values.

- Features like mean, skewness, and peak-to-peak amplitude have relatively lower impact.



III.
Eli5

Eli5 Analysis

•The top contributing features are:
argmax (Weight: 0.0117)
argmin (Weight: 0.0064)
kurtosis (Weight: 0.0012)
Other features like mean square, variance, and peak-to-peak signal have negligible importance.

Weight	Feature
0.0117 ± 0.0165	ArgMax
0.0064 ± 0.0301	ArgMin
0.0012 ± 0.0047	Kurtosis
0 ± 0.0000	Skewness
0 ± 0.0000	Abs Diff Signal
0 ± 0.0000	Mean Square
0 ± 0.0000	RMS
0 ± 0.0000	Max
0 ± 0.0000	Min
0 ± 0.0000	Variance
0 ± 0.0000	Peak-to-Peak
0 ± 0.0000	Std
0 ± 0.0000	Mean

2.Applied Explainable AI model over Random Forest

I. Lime

LIME Explanation for a Single Prediction :

Explanation: This visualization shows how individual feature values contribute to predicting whether a subject is healthy or schizophrenic.

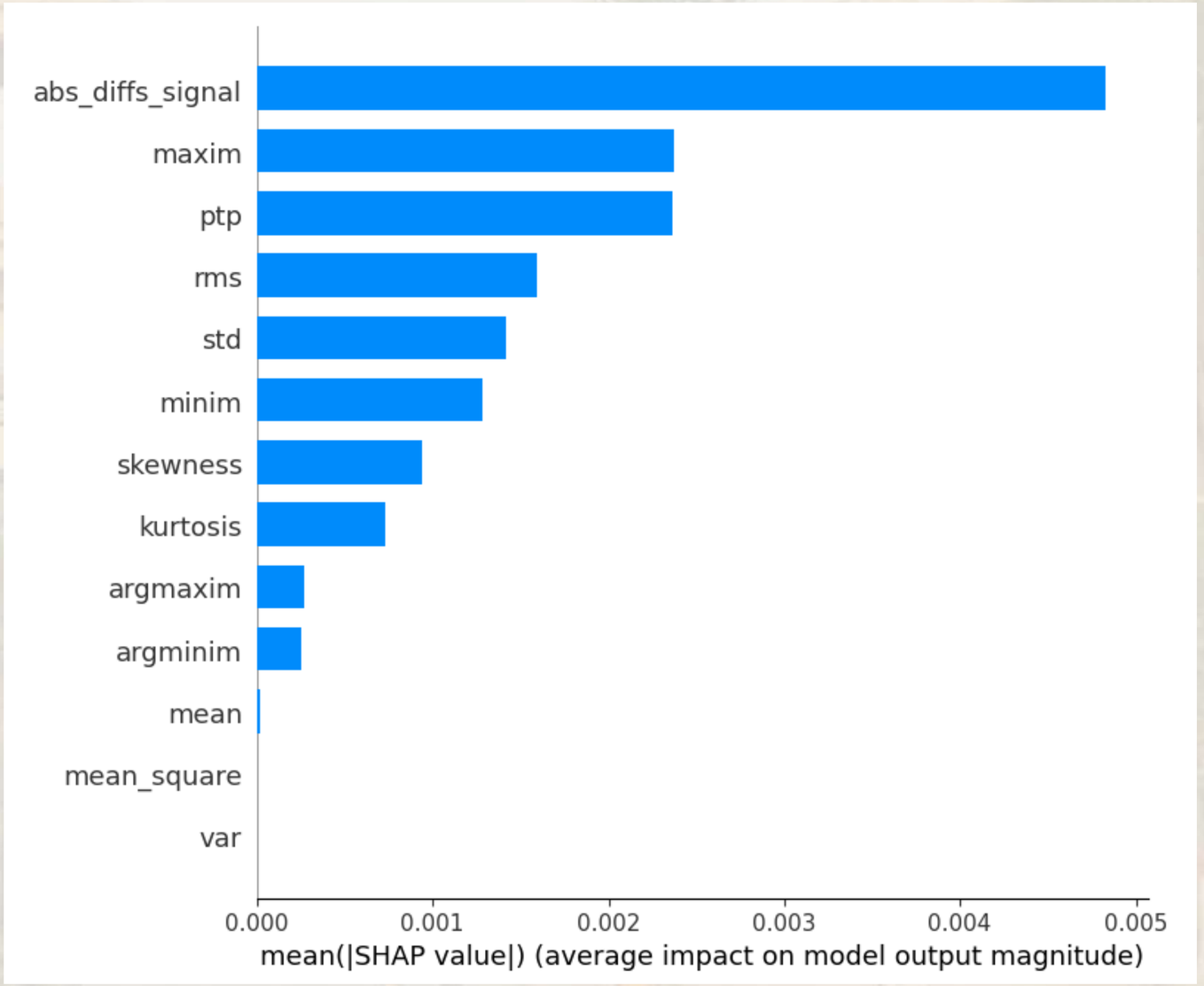
Key Observations : Abs Diff Signal (0.01) and Skewness (-0.08) are the dominant features influencing the prediction.

The prediction probability is 60% Healthy and 40% Schizophrenic.

II. Shap

SHAP-Based Feature Importance Ranking

- Explanation: This bar chart represents the mean absolute SHAP values for different features used in the Random Forest model. The higher the SHAP value, the more impact the feature has on the model's predictions.
- Key Observations:
  - The most influential feature is abs_diffs_signal, followed by maxim, ptp (peak-to-peak amplitude), and rms.
  - Features like argmaxim, argminim, mean, and variance have minimal impact on the predictions.



III.
Eli5

ELI5 Feature Importance Table :

Explanation: This table shows feature importance scores derived from ELI5, another explainability tool.

Key Observations:

The Abs Diff Signal is again identified as the most crucial feature.

Unlike SHAP, Skewness and Standard Deviation have higher importance scores here.

Negative importance scores indicate features that decrease the model's confidence in classification.

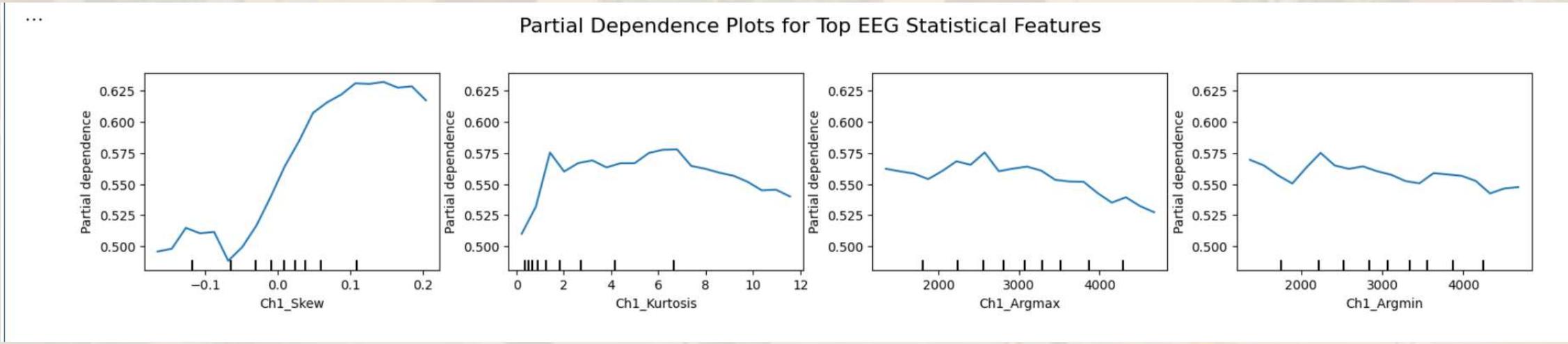
	Feature	Importance
10	Abs Diff Signal	0.001473
11	Skewness	0.000706
1	Std Dev	0.000675
9	RMS	0.000675
6	Argmin	0.000246
0	Mean	0.000000
3	Variance	0.000000
8	Mean Square	0.000000
2	Peak-to-Peak	-0.000031
4	Min	-0.000031
12	Kurtosis	-0.000061
5	Max	-0.000123
7	Argmax	-0.000246

IV. PDP

This plot shows how four top EEG features affect a model's prediction:

- Ch1_Skew: Strong positive impact — higher skew increases the predicted value.
- Ch1_Kurtosis: Moderate impact — peak influence around kurtosis = 2–3.
- Ch1_Argmax: Minimal effect — slight fluctuations, mostly flat.
- Ch1_Argmin: Slight negative impact — higher values reduce predictions.

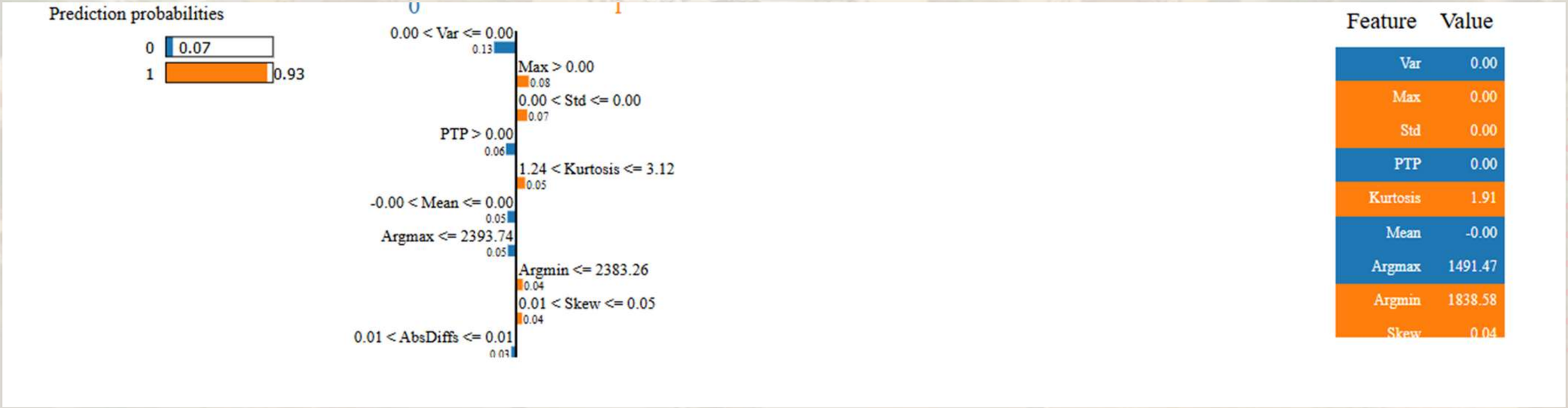
Overall, Ch1_Skew is the most influential feature.



3.Applied Explainable AI model over XGBoost

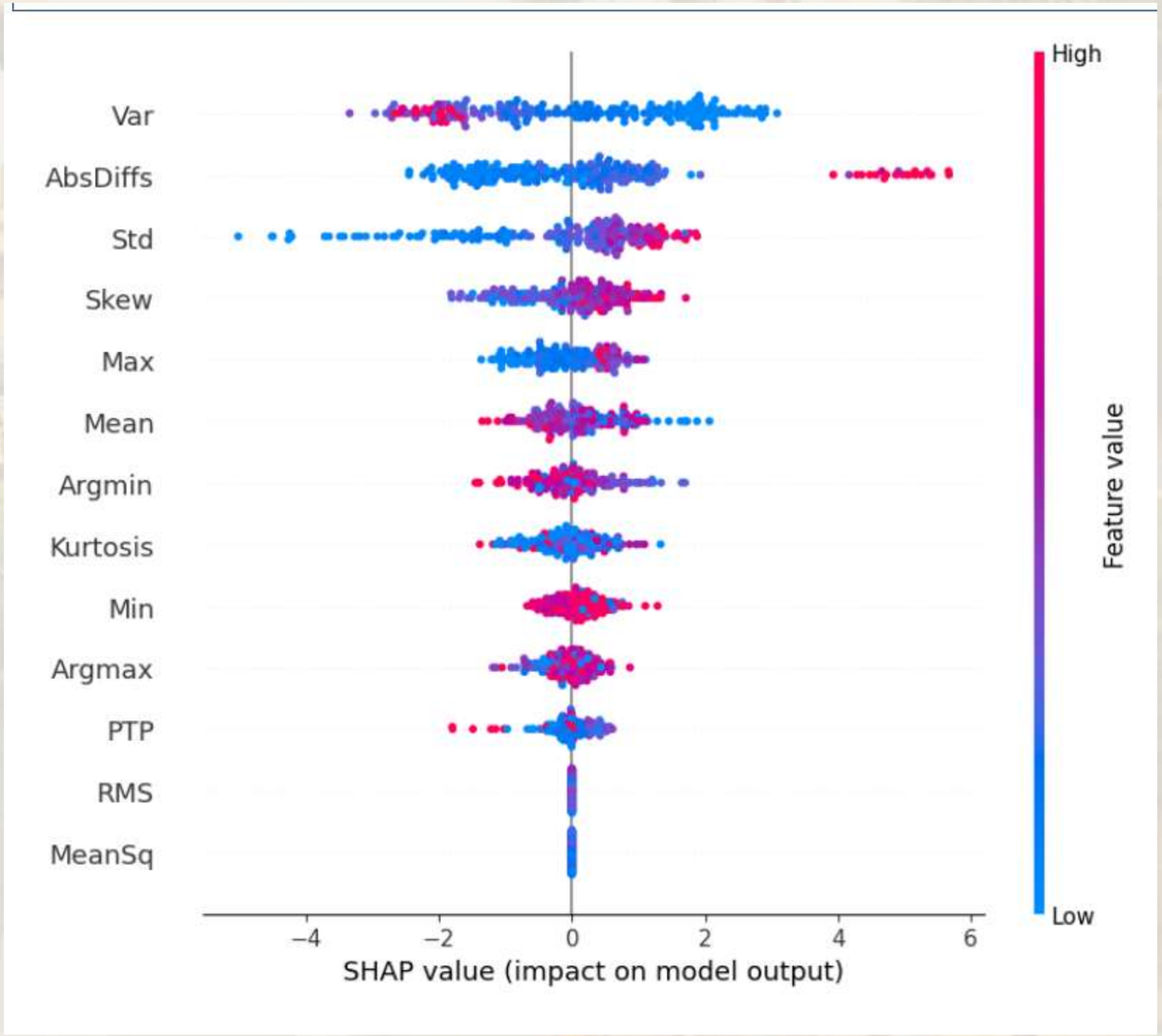
I. Lime

The LIME explanation visualizations show how a particular instance is classified.
The probability distribution bars indicate the likelihood of a patient being schizophrenic or healthy.
Decision thresholds for key features are displayed, such as Max > 0, Kurtosis > I.24, and PTP > 0.



II. Shap

•SHAP summary plots showing how each feature influences the model's prediction.
•Key Features:
•AbsDiffSignal, Max, PTP (Peak-to-Peak), RMS, and Std Dev are the most influential features in both models.
•SHAP values show how these features impact the model's decision, with red indicating higher feature values and blue indicating lower values.



III.Eli5

- The top features (AbsDiffSignal, Std, Var, Max) have high positive weights.

- Some features (like Argmax, MeanSq) have near-zero or negative importance.

...

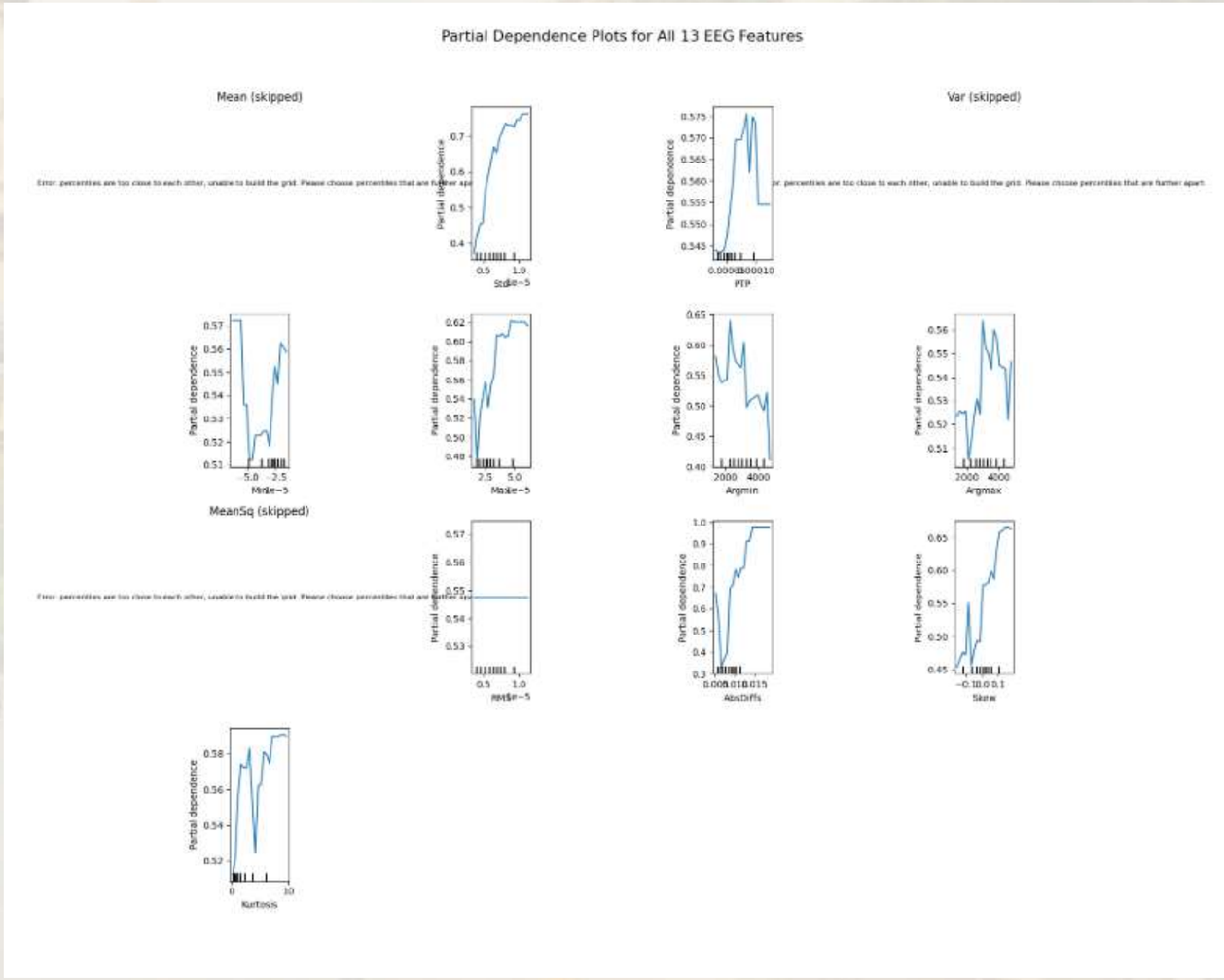
Weight	Feature
0.1747 ± 0.0331	AbsDiffs
0.1485 ± 0.0292	Std
0.0812 ± 0.0334	Var
0.0367 ± 0.0196	Max
0.0183 ± 0.0443	Kurtosis
0.0183 ± 0.0300	Skew
0.0148 ± 0.0171	Min
0.0070 ± 0.0180	PTP
0.0035 ± 0.0360	Mean
0 ± 0.0000	RMS
0 ± 0.0000	MeanSq
-0.0009 ± 0.0102	Argmax
-0.0009 ± 0.0261	Argmin

IV. PDP

This plot shows partial dependence of the model output on 13 EEG features. Some features like Mean, MeanSq, and Var were skipped due to insufficient variation.

Key takeaways:

- Std, Maxim, Kurtosis, Skewness, and AbsDiffs have a strong positive influence on prediction.
- Argmin shows a negative impact at higher values.
- Features like RMS, PTP, and Minim have relatively flat or minimal influence.



4. Applied Explainable AI model over AdaBoost

I.Lime

LIME Explanation (Local Interpretation)

- The LIME image shows Healthy classification with 99% probability.
- Key thresholds that influenced the classification include Std, Var, Min, Skewness, and ArgMax.
- The values of these features determined why the model strongly classified the sample as Healthy.

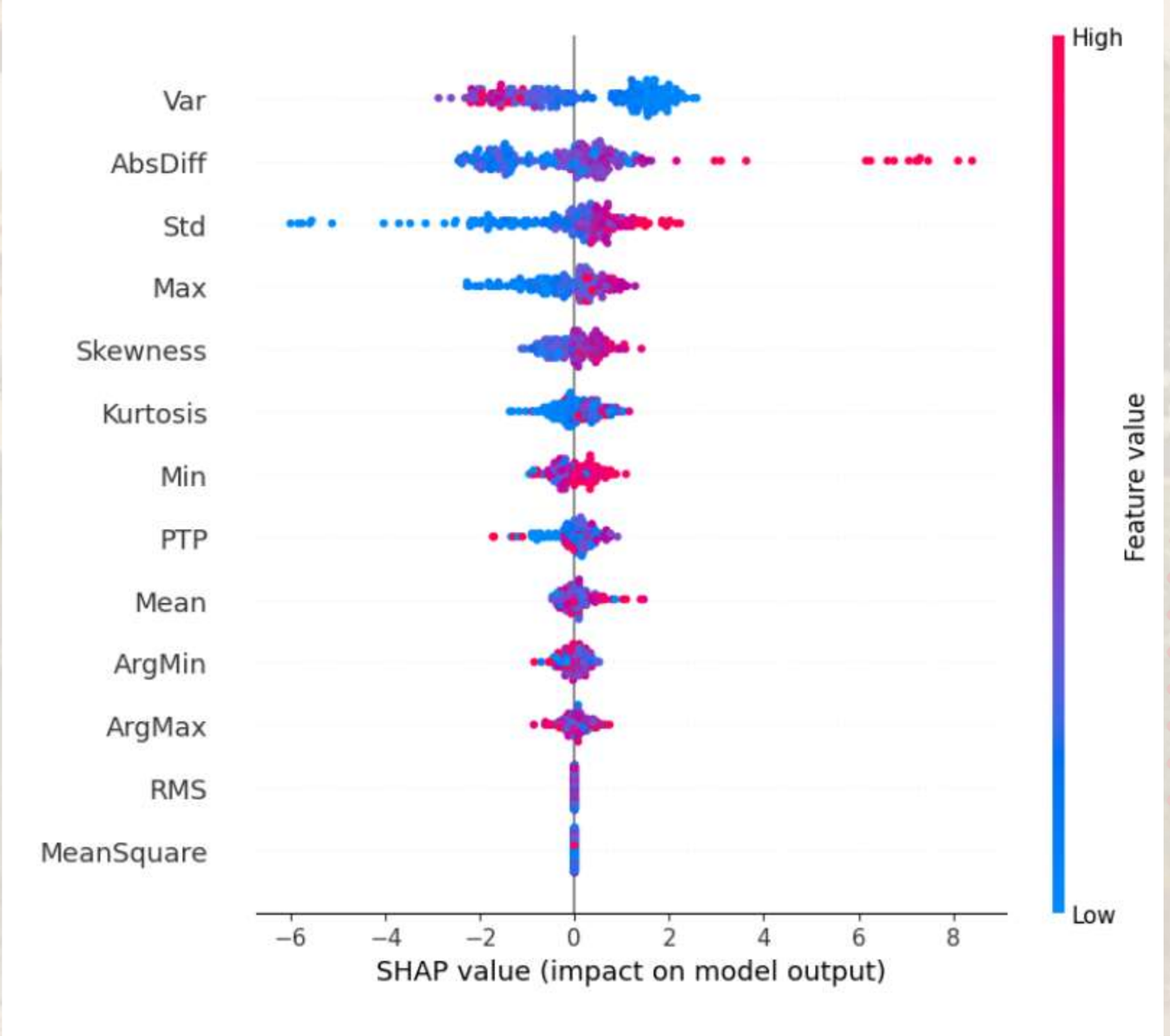
II. Shap

SHAP Summary Plot (Global Importance)

•Similar to XGBoost, the SHAP summary plot shows key features influencing predictions.

•Variance, AbsDiffs, and Std are the most significant features.

•The spread of SHAP values shows that some features have a stronger influence in AdaBoost compared to XGBoost.



III.Eli5

ELI5 Feature Importance (Global Importance)

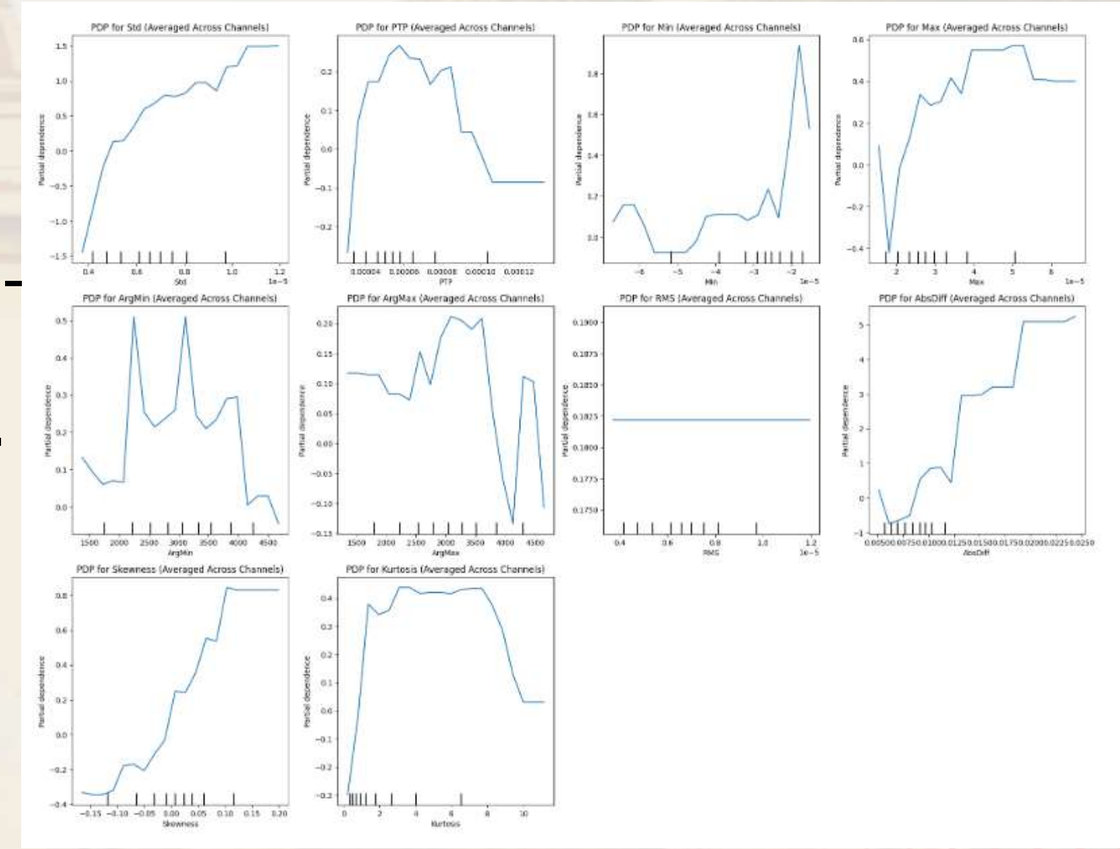
- AbsDiffs (0.1231), Std (0.0865), and Variance (0.0769) were the most impactful features.
- Unlike XGBoost, in AdaBoost, ArgMin had a slightly negative impact (-0.0087).
- Some features such as MeanSquare and RMS had no weight, meaning they did not contribute significantly.

Weight	Feature
0.1231 ± 0.0256	AbsDiff
0.0865 ± 0.0355	Std
0.0769 ± 0.0570	Var
0.0157 ± 0.0361	Skewness
0.0157 ± 0.0118	Kurtosis
0.0105 ± 0.0279	Max
0.0096 ± 0.0160	Mean
0.0061 ± 0.0251	PTP
0.0052 ± 0.0102	ArgMax
0.0026 ± 0.0225	Min
0 ± 0.0000	RMS
0 ± 0.0000	MeanSquare
-0.0087 ± 0.0228	ArgMin

IV. PDP

This plot shows PDPs for various statistical EEG features averaged across all channels. Here's a short interpretation for each:

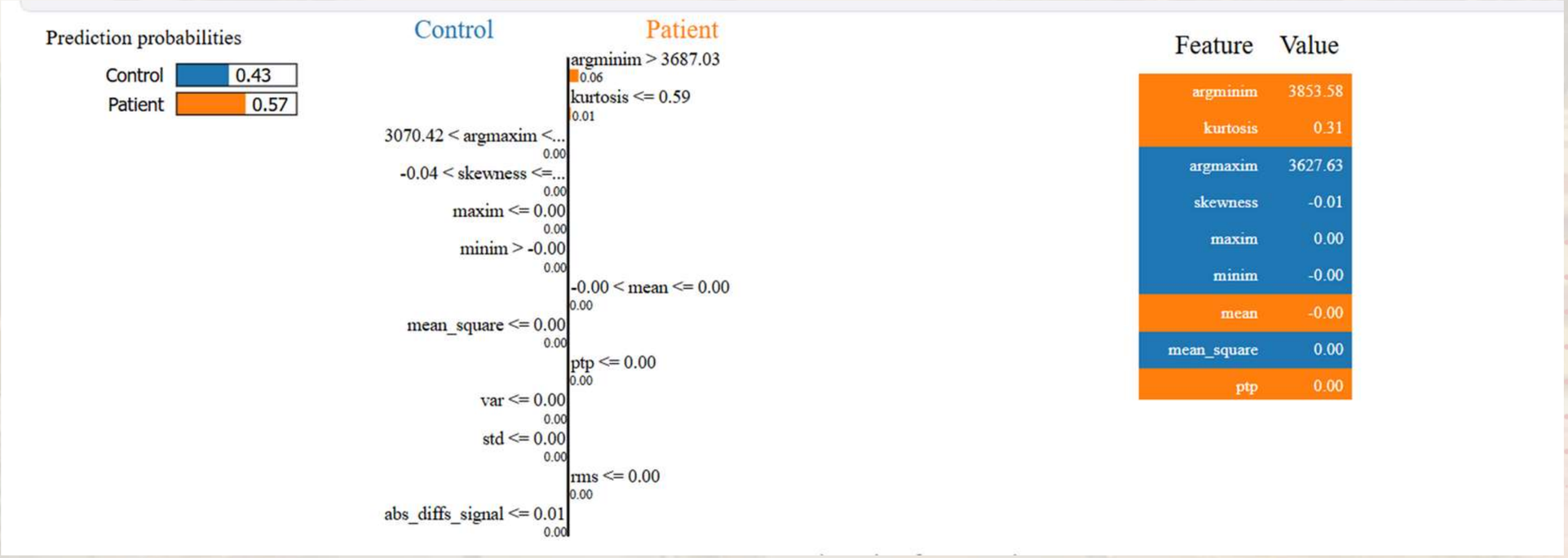
- Std: Strong positive trend — higher standard deviation increases prediction.
- PTP (Peak-to-Peak): Peak influence at mid-range; flattens at extremes.
- Min: Sharp increase at extreme values — indicates strong non-linear effect.
- Max: Moderate positive influence; peak effect at upper values.
- ArgMin / ArgMax: Noisy with small peaks — weak or inconsistent influence.
- RMS: Flat line — negligible or no effect on prediction.
- AbsDiff: Strong positive trend — larger differences boost predictions.
- Skewness: Clear positive trend — more skewed signals increase prediction.
- Kurtosis: Bell-shaped — moderate kurtosis most influential.



5. Applied Explainable AI model over SVM

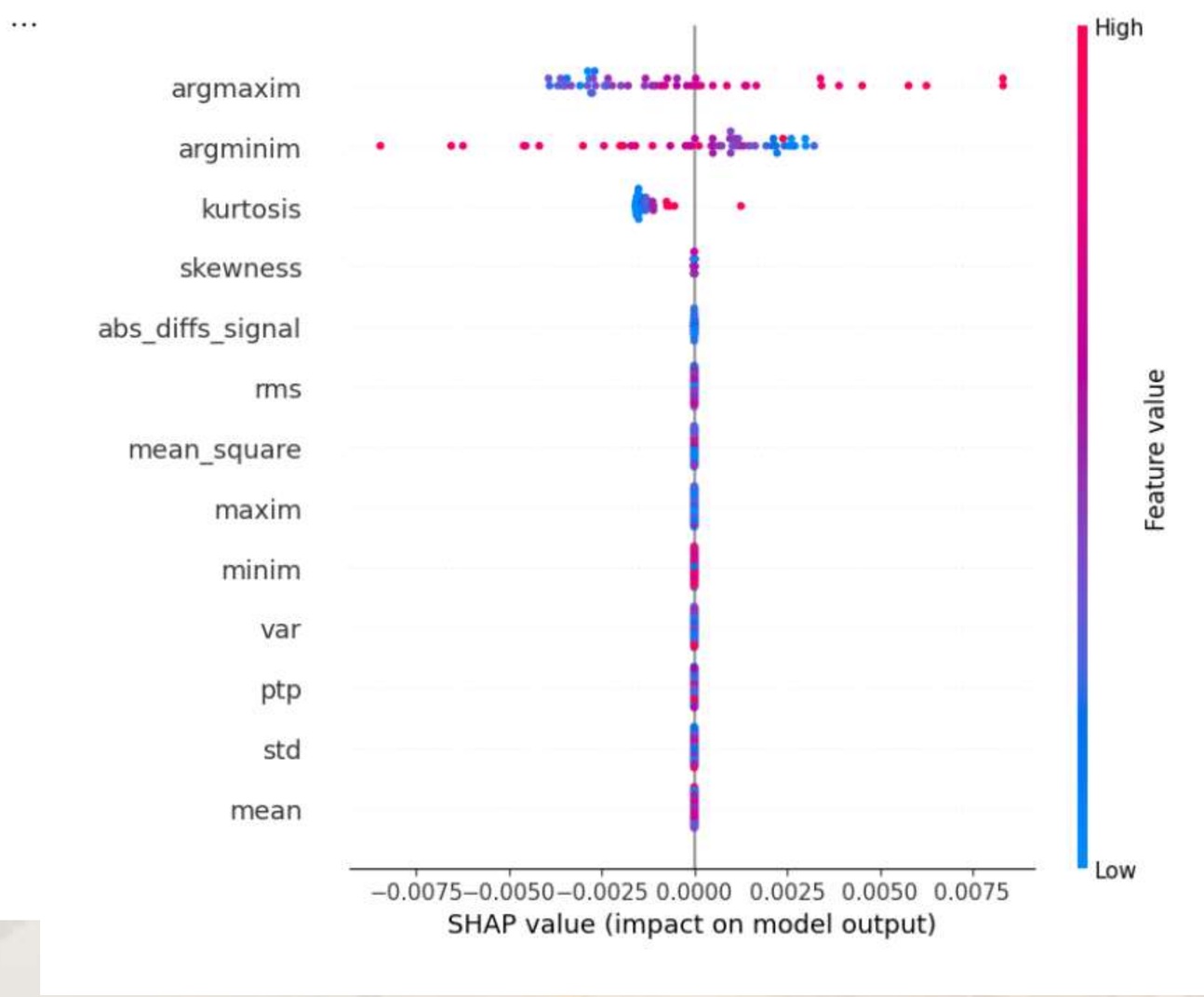
I.Lime

The SVM model (53% accuracy) predicts the sample as Patient (55%), mainly influenced by features like maxim, argmaxim, and argminim.



II. Shap

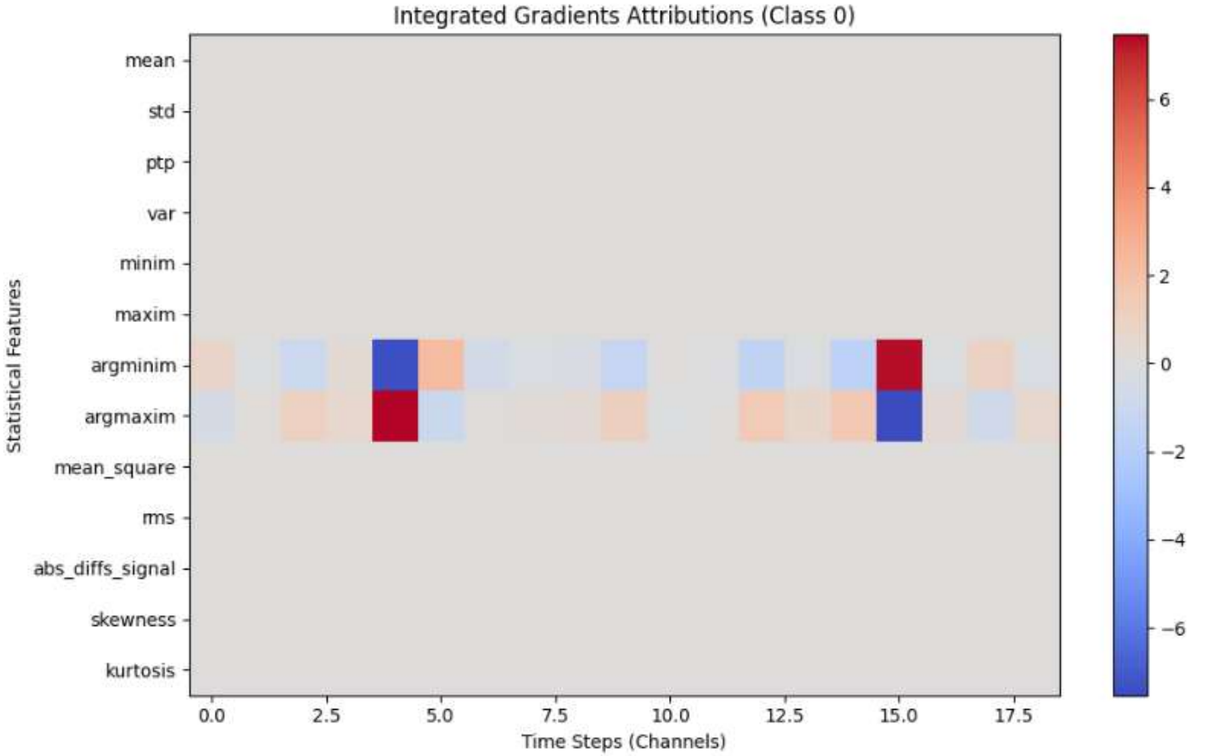
The model is most influenced by argmaxim, argminim, and kurtosis. Other features have minimal impact. High and low values of these key features significantly affect predictions.

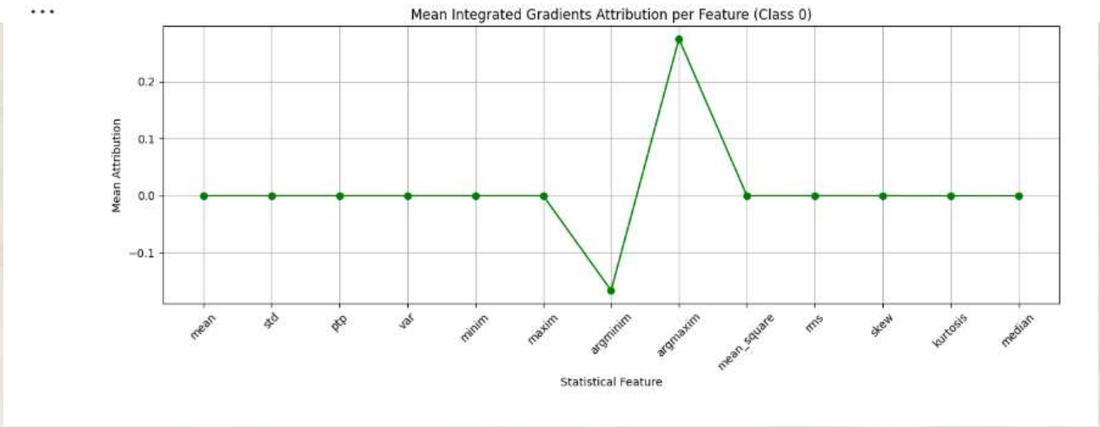


5. Applied Explainable AI model over LSTM

I. Integrated Gradients

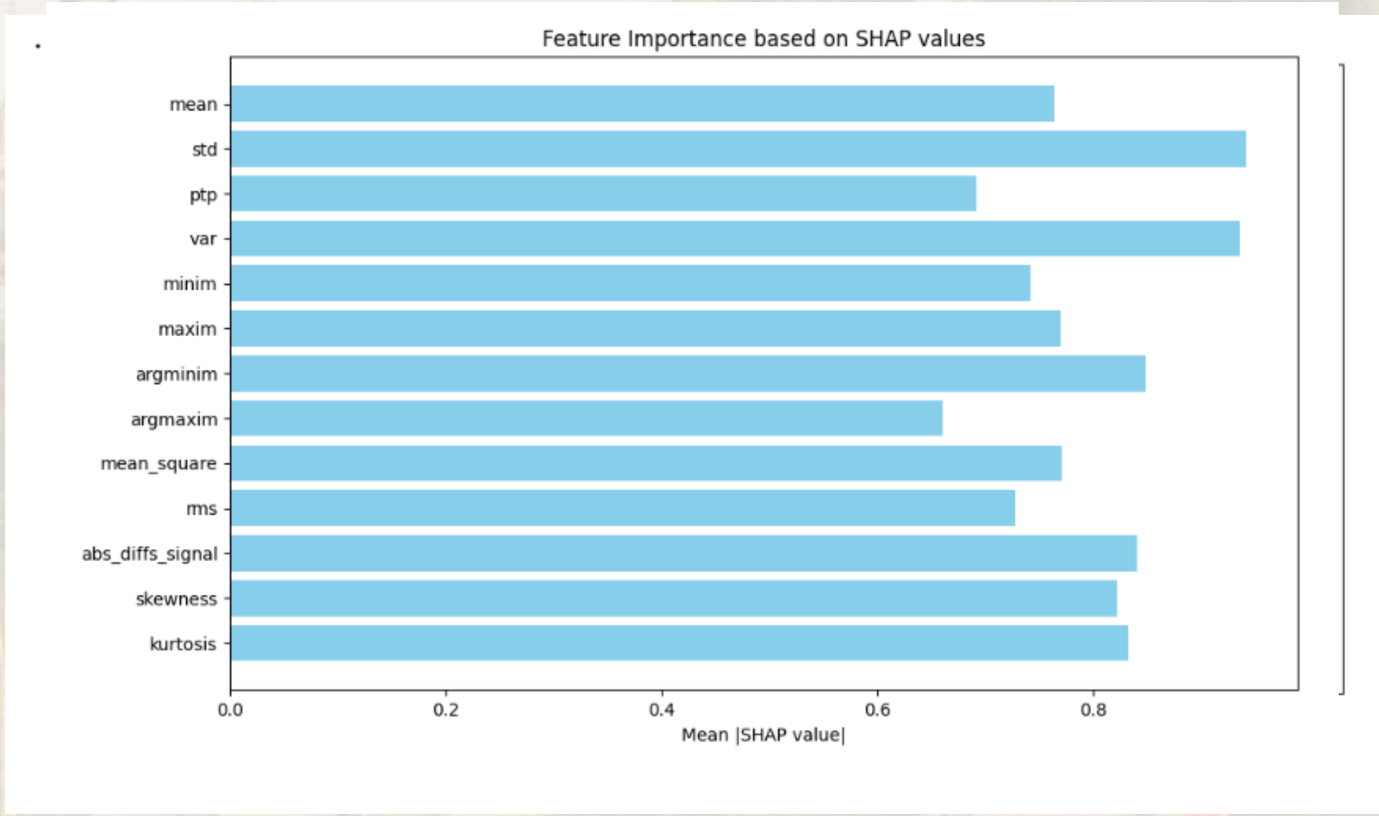
- Line Plot: argmaxim positively impacts Class 0 prediction; argminim negatively impacts it. Other features have minimal effect.
- Heatmap: Key contributions from argmaxim and argminim at specific time steps; rest show low influence.





II. Shap

The SHAP bar chart shows which EEG features most influenced the model's predictions. Features like standard deviation, variance, and argmin index are the most important. Others like kurtosis, skewness, and abs_diffs_signal also contribute significantly. This helps explain and interpret the model's decisions.



Conculsion

The application of XAI techniques—LIME, SHAP, ELI5,PDP and Integrated Gradients—over the Five machine learning models (Random Forest, AdaBoost, etc.) and One Deep learning model(RNN(LSTM))provided valuable insights into model behavior, feature importance, and decision-making processes.

Key conclusions are:

- Feature Importance Across Models
- Across all models, Abs Diff, Standard Deviation (Std), Variance (Var), and Skewness emerged as critical features in distinguishing between healthy and schizophrenic subjects.
- Min ,ArgMin,ArgMax,AbsDiff values were sometimes pushing classification towards schizophrenia, indicating their potential role in capturing unique patterns in EEG signals.

Future Work

- Expanding Data Volume: Focus on acquiring and incorporating more diverse and extensive EEG datasets. Working with larger datasets can significantly improve model robustness and performance, enabling the identification of more nuanced patterns associated with schizophrenia and other disorders.
- Data Augmentation: Investigate advanced data augmentation techniques specific to EEG signals to increase the dataset size artificially, improving the model's generalization ability without the need for additional data collection.
- We applied bandpass filtering to extract key EEG bands: Delta, Theta, Alpha, Beta, and Gamma. For enhanced signal quality, we plan to explore additional filters: Notch , High-pass , Low-pass , Adaptive filtering, and ICA-based filtering
- We can try to use CNN as blackbox model and use Grad-cam XAI tool for better interpretation

References

1. Github repositories for code :https://github.com/Ayushi10-kumari/Explainable-AI-In-Classifying-Nurodisorder
2. Dataset for this dissertation :https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188629
3. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signalby Mohammed Saidul Islam1,Iqram Hussain2,3,,Md Mezbaur Rahman1,Se Jin Park4 andMd Azam Hossain
4. HTTPS://WWW.DATACAMP.COM/TUTORIAL/EXPLAINABLE-AI-UNDERSTANDING-AND-TRUSTING-MACHINE-LEARNING-MODELS
5. SMITH K. KHAREA, U. RAJENDRA ACHARYA ,https://www.sciencedirect.com/science/article/pii/S0010482523001415
6. HTTPS://WWW.YOUTUBE.COM/WATCH?V=OZJIIGSGP9E&LIST=PLV8YXWGOXVVOVP-J6ZTXHF3QCKXT6VORU
7. Ahmad ,Chaddad,1,2,* Yihang Wu,1 Reem Kateb,3 and Ahmed Bouridane4 Chang-Hwan Im, Academic Editor and Yvonne Tran, Academic Editor, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10385593/
8. Smith K. Khare , U. Rajendra Acharya, https://www.sciencedirect.com/science/article/pii/S0950705123006081
9. Graph-based analysis of brain connectivity in schizophrenia
10. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal by Mohammed Saidul Islam1,Iqram Hussain2,3,,Md Mezbaur Rahman1,Se Jin Park4 and Md Azam Hossain
11. https://shap.readthedocs.io/en/latest/index.html
12. https://github.com/talhaanwarch/youtube-tutorials
13. https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441 https://www.mdpi.com/1424-8220/22/24/9859