Neural Condition Classification Project

1. Introduction

1.1 Task Description

The primary goal of this project is to analyze and classify Electroencephalography (EEG) data collected during a brain activity experiment. The classification focuses on differentiating between three conditions:

CONSISTENT, MISLEADING, and CONTROL.

The participants were shown stimuli corresponding to these conditions, and their brain activity was recorded across 32 electrodes (excluding Fp1) over multiple time windows and trials. The final objective is to train a machine learning model to classify these conditions based on the recorded EEG metrics.

1.2 Data Description

The dataset is organized into two primary types of files:

1. ALLTRIALS Files: These files measure the Area Under the Curve (AUC) of brain activity across specified time windows:

- Time Windows: 150-275 ms, 150-350 ms, 250-400 ms, 300-500 ms, and 550-800 ms.

- Each file combines the AUC values for the 8 trials per condition for all 32 electrodes.

2. PEAK Files: These files provide peak values and latencies for different components:

- Components:

- P2: 150-275 ms

- FN400: 300-500 ms

- P3: 250-400 ms

- N2: 150-350 ms

Dataset Statistics

Participants: 25 participants

Conditions: 3 (CONSISTENT, MISLEADING, CONTROL)

Electrodes: 32 electrodes (excluding Fp1)

Total Data Points: 75 (25 participants × 3 conditions)

The key research question guiding this analysis is:

How does brain activity differ in misleading trials compared to consistent and control trials?

2. Data Preparation

2.1 Data Exploration

The following steps were conducted to explore and clean the dataset:

1. Loaded all the ALLTRIALS and PEAK files.

2. Verified the structure of each file and confirmed the exclusion of the Fp1 electrode.

3. Merged all data files into a single combined DataFrame for further analysis.

4. Identified missing values and cleaned the data by removing or imputing irrelevant columns.

2.2 Data Visualization

To better understand the structure and variability of the data, the following visualizations were generated:

1. Boxplots: Plotted the distribution of EEG values for each condition to observe differences visually.

2. Heatmap: Generated a heatmap of mean electrode values for each condition to identify patterns.

2.3 Preprocessing

The preprocessing steps included:

1. Normalization: Standardized all EEG values using StandardScaler to ensure consistency across features.

2. Reshaping: Transformed the data into a long format, including columns:

File, Electrode, Condition, and Value.

3. Feature Extraction: Extracted meaningful electrode-condition features by grouping the data.

3. Feature Generation, Selection, and Transformation

3.1 Feature Selection

Removed low-variance features using VarianceThreshold (threshold=0.01) to reduce redundant features.

3.2 Dimensionality Reduction

Applied Principal Component Analysis (PCA) to retain 95% of the variance while reducing dimensionality.

3.3 Final Dataset Formation

The final dataset contains:

Principal components representing EEG features and Condition labels as the target variable.

4. Model Development

4.1 Model Selection

A Random Forest Classifier was chosen for its robustness and ability to handle high-dimensional data.

4.2 Hyperparameter Tuning

Hyperparameters were optimized using GridSearchCV with 5-fold cross-validation:

- Parameters Tuned:

n_estimators: [50, 100, 200]

max_depth: [None, 10, 20, 30]

min_samples_split: [2, 5, 10]

min_samples_leaf: [1, 2, 4]

4.3 Model Evaluation

The model was evaluated using the following metrics:

Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

5. Results and Conclusion

5.1 Results

Best Parameters (after tuning):

- n_estimators: 100

- max_depth: 20

- min_samples_split: 5

- min_samples_leaf: 2

Performance on Test Set:

- Accuracy: 85%
- Precision: High precision for all conditions
- Recall: Misleading condition had slightly lower recall compared to others.

5.2 Feature Importance

The most influential electrodes for condition classification:

- Fz
- Cz
- Pz

These results highlight the role of specific brain regions in differentiating between conditions.

5.3 Conclusion

The Random Forest model successfully classified EEG data into three conditions with high accuracy. Significant differences in brain activity were observed between misleading trials and other conditions.

6. Final Deliverables

1. Final Report
2. Cleaned and Processed Dataset
3. Model Code and Results
4. Visualizations: Boxplots, Heatmaps, Confusion Matrix, Feature Importance

7. Acknowledgments

Special thanks to Professor Dhanuska Bandara for providing the resources and guidance for this project.