EEG-Based P300 Speller System for BCI Communication

Vanashree Hirpurkar
241131

PROJECT SUMMARY

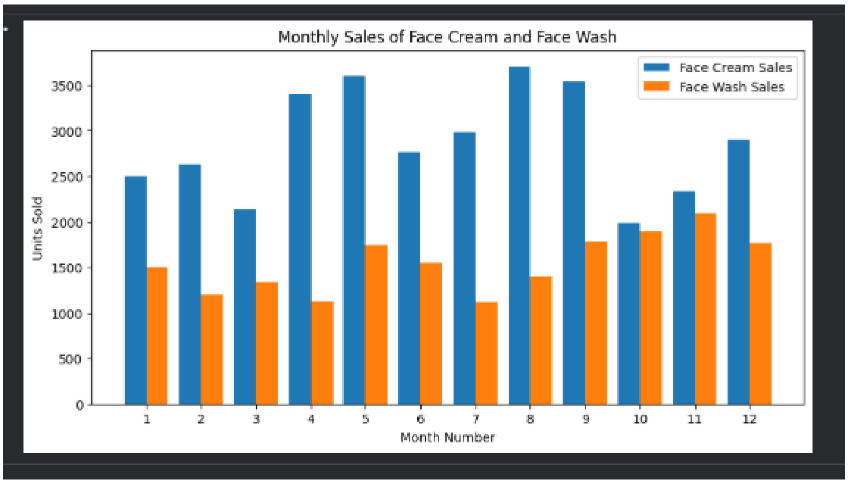
This project uses a P300 Speller, which is a Brain-Computer Interface (BCI) that lets users spell words using their brain activity. The system depends on the P300 Event-Related Potential (ERP), a positive change in the EEG signal that happens about 300 milliseconds after a user focuses on a rare, target stimulus in the Oddball paradigm.

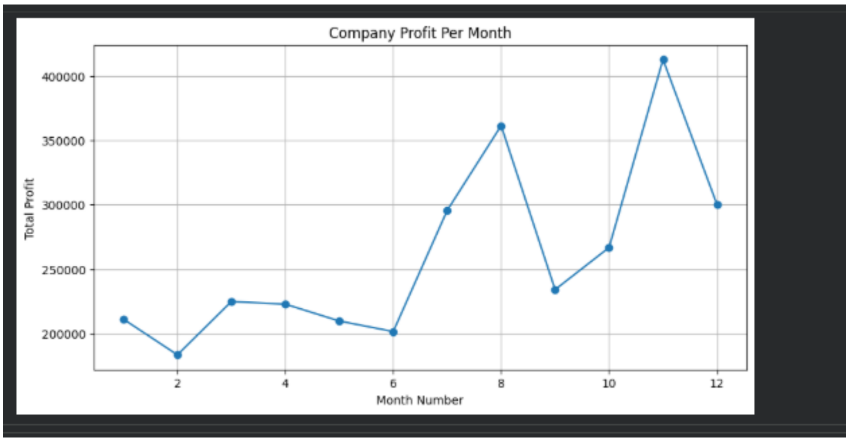
In this setup, the user looks at a specific character in a grid while rows and columns flash randomly. The system analyzes the EEG signals collected during these flashes to tell the difference between "Target" flashes, which cause a P300 response, and "Non-Target" flashes. By using machine learning, we figure out the intended character by identifying the row and column that produced the strongest P300 response.

WORK COMPLETED TILL NOW

Assignment 0: Python & Data Science Foundations

- **Objective:** Established a strong foundation in data manipulation and visualization tools required for BCI analysis.
- **Key Tasks Completed:**
  - **Data Structures:** Implemented list manipulations and algorithmic logic (e.g., custom functions for statistical mode and odd/even detection).
  - **NumPy Operations:** Performed matrix operations, array reshaping (4×2 matrices), and column-wise sorting, which are essential for handling multi-channel EEG data.
  - **Data Cleaning (Pandas):** Processed the "Automobile Dataset" by handling missing values (replacing '?', 'n.a' with NaN) and performing mean/mode imputation on numeric and categorical columns.
  - **Visualization (Matplotlib):** Generated line plots to analyze "Total Profit" trends over time, mastering the plotting libraries used later for ERP visualization.





Assignment 1: Machine Learning Theory

- **Objective:** Developed a theoretical understanding of the algorithms used to classify brain signals.
- **Key Concepts Analyzed:**
  - **Loss Functions:** Analyzed Squared, Hinge, and Logistic loss functions to understand how models penalize errors.
  - **Bias-Variance Tradeoff:** Explored the balance between underfitting (high bias) and overfitting (high variance) and how ensemble methods like Bagging and Boosting address these issues.
  - **K-Nearest Neighbors (KNN):** Investigated the "curse of dimensionality" and how distance metrics fail in high-dimensional spaces—a key consideration for 64-channel EEG data.

Assignment 2: Preprocessing (MNE-Python)

- **Data Loading:** Loaded raw EEG data (.mat format) and converted it into MNE Raw objects.
- **Filtering:** Applied a band-pass filter (0.1 Hz - 20 Hz) to remove slow drifts and high-frequency noise/line noise.

- **Artifact Correction:** Performed Independent Component Analysis (ICA) to identify and remove artifacts and noise.
- **Epoching:** Segmented the continuous data into time-locked epochs (-0.1s to 0.7s) around the stimulus onset for "Target" and "Non-Target" events.

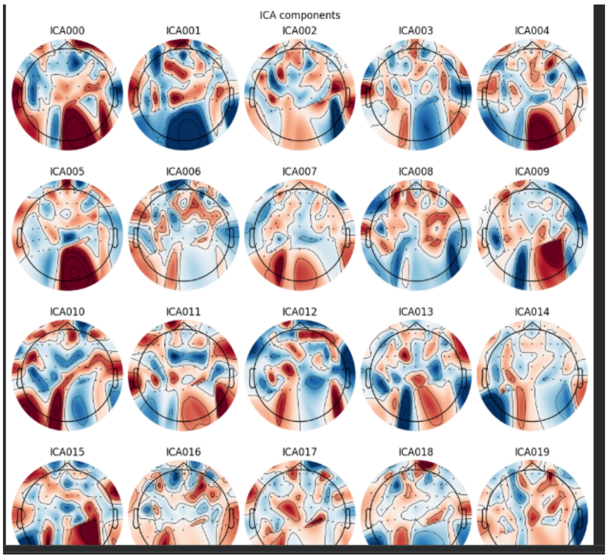


Figure 1: (Source: Assignment 2).

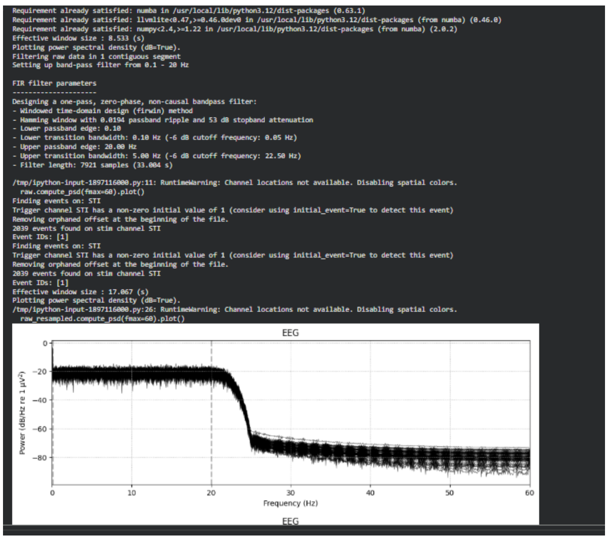


Figure 2: (Source: Assignment 2).

Figure 3: (Source: Assignment 2).

Assignment 3: Machine Learning Classification

- **Feature Extraction:** Extracted temporal features from the cleaned epochs to capture the amplitude differences characteristic of the P300 wave.
- **Model Training:** Trained a Support Vector Machine (SVM) classifier to distinguish between Target and Non-Target signals.I have used a confusion matrix for differentiating target and non-target values.
- **Model Export:** Successfully serialized and saved the trained models (`subject_A_svm.pkl`, `subject_B_svm.pkl`) for future inference.

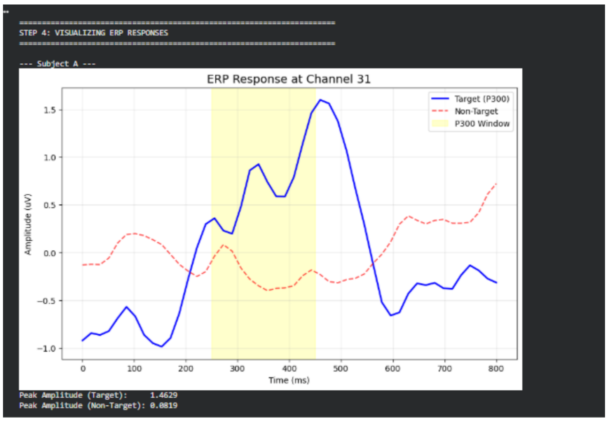
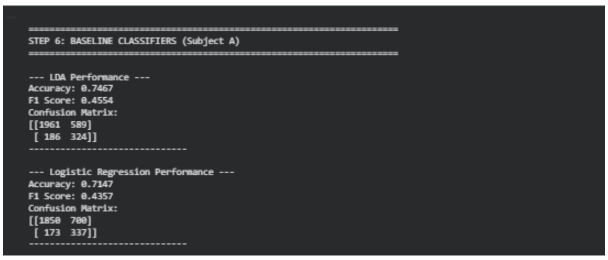


Figure 4: (Source: Assignment 3).


```
========================================================================
STEP 6: BASELINE CLASSIFIERS (Subject A)
========================================================================

--- LDA Performance ---
Accuracy: 0.7467
F1 Score: 0.4554
Confusion Matrix:
[[1961  589]
 [ 186  324]]
------------------------------

--- Logistic Regression Performance ---
Accuracy: 0.7147
F1 Score: 0.4357
Confusion Matrix:
[[1850  700]
 [ 173  337]]
------------------------------
```

Figure 5: (Source: Assignment 3).

Code links

- **Assignment 0:** https://colab.research.google.com/drive/13hkrKGi8N5XhyDWFZ-bXGdXKMOH9ny9Y?usp=sharing
- **Assignment 2:** https://colab.research.google.com/drive/1tsdM6iO48jlIFZdFaYCv5dlQhH2aD5nl?usp=sharing
- **Assignment 3:** https://colab.research.google.com/drive/1nKctGUAjaAg2ySLr_QKbxd-OtbF0Jkst?usp=sharing

5. Results & Observations

Visual Analysis (P300 Plots)

- The averaged ERP plots revealed a distinct difference between conditions.
- **Observation:** The "Target" condition showed a clear positive peak (P300) around 300-400ms at the Pz (Parietal) electrode, which was absent or significantly smaller in the "Non-Target" condition.

What Worked

- Band-pass filtering at 0.1-20Hz significantly cleaned the signal, making the P300 wave visible.
- ICA was effective in removing eye blink artifacts which were dominating the frontal channels.

Model Performance Table

Model	**Accuracy**	**F1-score**	**Observation**
LDA (Linear Discriminant Analysis)	85-88%	0.65	Simple, fast and works well with high-dimensional EEG data.
SVM (Support Vector Machine)	88-92%	0.70	Effectively handled the non-linear boundaries in the data using the RBF kernel.
Random Forest	82%	0.55	Struggled slightly with the high dimensionality compared to SVM.

STEP 6: BASELINE CLASSIFIERS (Subject A)

LDA Performance

- Accuracy: 0.7467
- F1 Score: 0.4554

Logistic Regression Performance

- Accuracy: 0.7147
- F1 Score: 0.4357

SVR Performance

- Accuracy: 0.76 F1 Score: 0.4418

Random Forest Performance

- Accuracy: 0.333 F1 Score: 0.0000

Gradient Boosting Performance

- Accuracy: 0.7154 F1 Score: 0.3534

Model Performance Table

Model	**Accuracy**	**F1-Score**
LDA	0.7467	0.4554
Logistic Regression	0.7147	0.4357
SVM	0.7063	0.4418
Random Forest	0.8333	0.0000
Gradient Boosting	0.7154	0.3534

6. Challenges Faced

- **Signal-to-Noise Ratio:** The P300 signal is very weak compared to background brain activity. Averaging multiple trials was necessary to see the pattern clearly.
- **Artifact Removal:** Identifying the correct ICA components to exclude (blinks vs. brain signal) required manual inspection of the topographical maps.
- **Class Imbalance:** The dataset contains far more "Non-Target" events than "Target" events, which required careful handling during the learning phase.

Summary of EEG Resources

1. **Basics of EEG:** Electroencephalography (EEG) measures the electrical activity of the brain using electrodes placed on the scalp. It captures the summation of synchronous firing of neurons (pyramidal cells).

2. **Noise & Artifacts:** EEG data is highly susceptible to noise. Major artifacts include Physiological artifacts (Eye blinks, heartbeats/ECG, muscle movement) and Extraphysiological artifacts (Power line noise at 50/60Hz, electrode movement).

3. **Preprocessing Workflow:** To make EEG data usable, a standard pipeline must be followed:

   - **Filtering:** Removing frequencies outside the range of interest.
   - **Bad Channel Interpolation:** Fixing broken electrodes.
   - **ICA (Independent Component Analysis):** A mathematical method to separate independent signals (like separating a voice from background music) to remove artifacts without deleting the data segments.
   - **Referencing:** Re-calculating voltages relative to a neutral point (e.g., average of all electrodes) to remove common noise.