EEG-based BCI Control System using Machine Learning: Leveraging Spectral and Canonical Correlation Features

Anas Taha, Mohamed Ahmed, Yousef Mohamed.

Faculty of Science, Helwan University - Faculty of Computers and Artificial Intelligence, Cairo University

Objective:

The objective of this project is to design and implement a machine learning pipeline capable of classifying motor imagery (MI) and Steady State Visually Evoked Potential (SSVEP) EEG signals into real-time control applications in a cross-subject setting. The system aims to generalize well across different subjects by leveraging preprocessing, informative feature extraction, and supervised classification techniques.

Dataset Description and Challenges:

The dataset comprises EEG recordings from 30 subjects performing motor imagery (MI) tasks. Each subject completed 8 sessions, with each session consisting of multiple trials sampled at 250 Hz. The MI trials lasted for 9 seconds, while SSVEP trials lasted for 7 seconds. The validation set contains data from only 5 subjects, each contributing 1 session with 10 trials, resulting in a total of 50 trials for validation.

This limited quantity posed challenges in accurately evaluating model performance. To address this, we adopted a Leave-One-Subject-Out Cross-Validation (LOSO-CV) strategy, which offers a more reliable estimate of inter-subject generalization — a critical aspect in brain-computer interface (BCI) applications.

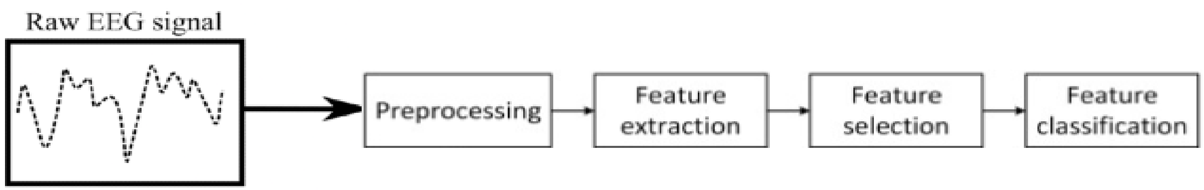
Due to the inherently noisy nature of EEG signals, the system requires a careful preprocessing pipeline. Furthermore, the absence of test labels prevented any form of subject-specific adaptation or fine-tuning. As a result, the preprocessing and feature extraction stages were designed to maximize generalization across unseen subjects by focusing on robust, discriminative features rather than subject-dependent patterns.

In the next few pages, we will be showcasing the methodology for both MI and SSVEP tasks Separately.

1. Motor Imagery Classification (MI)

In this task, the objective is to classify EEG signals into two classes: left-hand and right-hand motor imagery. While Common Spatial Patterns (CSP) is a well-established technique for extracting discriminative features in binary MI classification, its performance deteriorated significantly in our cross-subject setup.

This decline is attributed to CSP's sensitivity to subject-specific spatial filters, which do not generalize well across individuals. As a result, we adopted a more robust pipeline, focusing on preprocessing and feature extraction methods designed to improve inter-subject generalization. The following sections detail the steps taken to prepare the data and construct meaningful features for classification.



1.1 Signal Acquisition and Preparation

The `EEGPSDDataset` class is responsible for loading, preprocessing, and feature extraction of EEG trial data for classification. It reads raw EEG signals from CSV files based on trial metadata, extracts a predefined set of channels, and applies a preprocessing pipeline to clean and segment the data

To improve the quality of the EEG data and reduce the influence of motion artifacts, we implemented a filtering mechanism based on accelerometer and gyroscope data embedded within each trial.

Each trial's motion is quantified using the motion_artifact_score function, which calculates two deviation metrics

- Accelerometer deviation: The mean of the 3-axis accelerometer (AccX, AccY, AccZ) is compared to the expected resting vector [0, 0, 1]. A high deviation indicates non-stationary head/body movement.

- Gyroscope deviation: The mean of the 3-axis gyroscope (Gyro1, Gyro2, Gyro3) is compared to the stationary baseline [0, 0, 0]. A nonzero average suggests rotational motion during the trial.

The final motion artifact score is a weighted combination of these two deviations.

Once scores are computed for all trials, we normalize them using Z-score normalization and apply a threshold-based filtering strategy. Trials with Z-scores exceeding a specified threshold (e.g., |z| > 2.5) are considered outliers and are removed from the dataset.

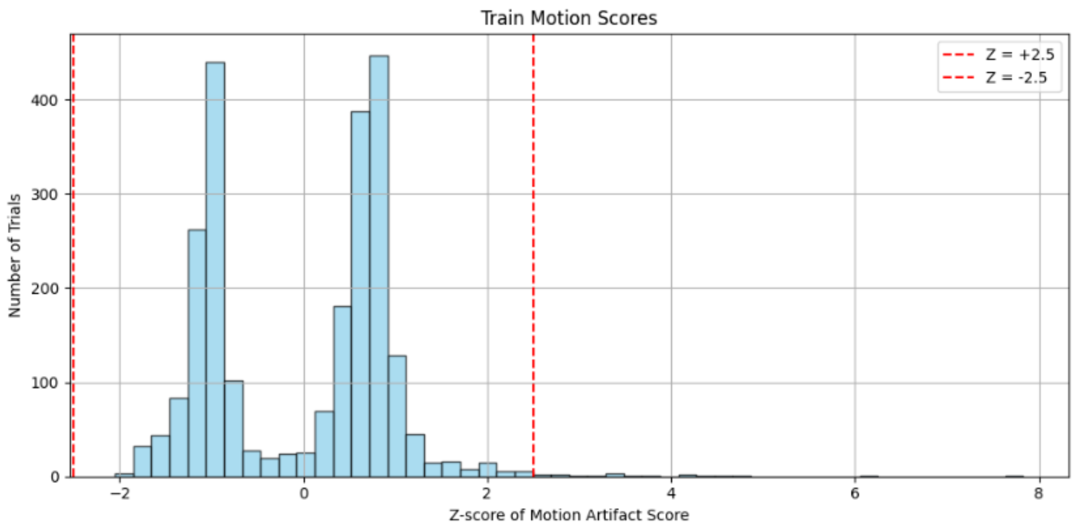


Figure 1 - Z-score distribution for MAS

1.2 Signal Preprocessing

The EEG signal preprocessing pipeline is implemented in the `EEGPreprocessor` class and is designed to clean, standardize, and extract informative task-relevant segments from each trial. The preprocessing is crucial for removing artifacts, reducing subject variability, and enhancing the generalizability of learned features.

- **1.2.1 Bandpass Filter**

  A 4th-order Butterworth bandpass filter is applied with cutoff frequencies set to 8–30 Hz, targeting the Mu (8–13 Hz) and Beta (13–30 Hz) bands. These bands are known to contain discriminative information for motor imagery tasks

- **1.2.2 Notch Filter**

Upon inspecting the power spectral density (PSD) distribution across the EEG channels, a prominent spike was observed at 50 Hz, indicating a significant presence of power line interference. This is a common artifact in EEG recordings, particularly in countries where the electrical grid operates at 50 Hz, such as Egypt.

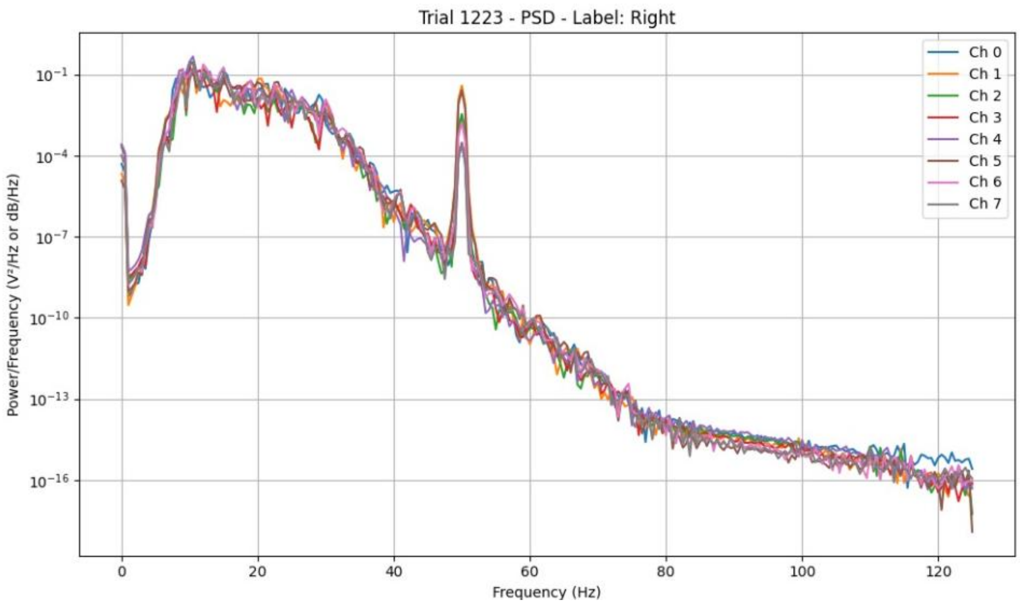


Figure 2 - PSD for a trial without notch filtering

Removing this interference prevents it from dominating or distorting the PSD and other frequency-domain features.

- **1.2.3 Independent Component Analysis (ICA)**

ICA is applied using the MNE library to identify and suppress artifact-related components, such as eye movements (EOG), muscle artifacts, or other stereotypical noise patterns. Although automated exclusion is possible, a manual exclusion of components [0, 1, 2] was used empirically based on observations.

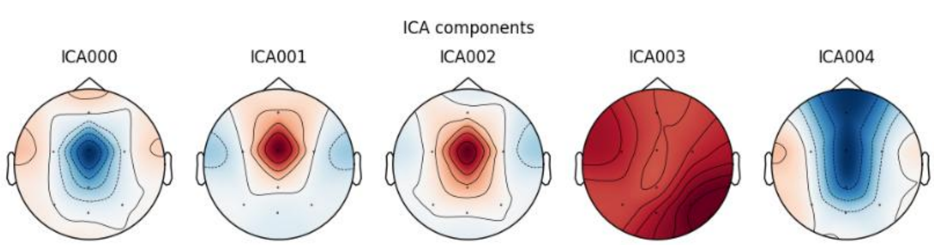


Figure 3 - ICA Analysis showing components such as ICA000 and ICA001 display spatial topographies characteristic of artifacts

Also, upon inspecting the temporal waveforms in the second image, we notice that IC000 and IC001 display characteristic large-amplitude, sharp deflections, definitively confirming their nature as prominent eye blink artifacts.

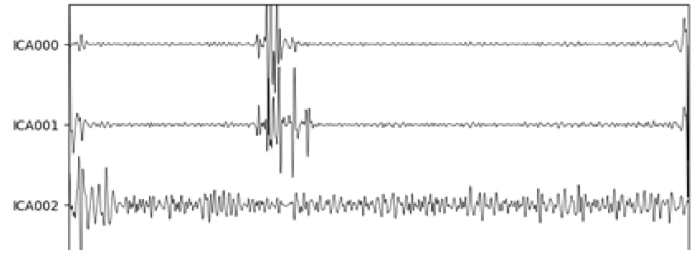


Figure 4 - Eye blink artifacts in ICA000 and ICA001

- **1.2.4 Baseline Correction**

Baseline correction is performed using the first 2 seconds (0.0–2.0 sec) of the trial as the reference window. The mean activity in this interval is subtracted from the entire signal to account for subject-specific baseline offsets. This centers the signal and minimizes inter-subject variability.

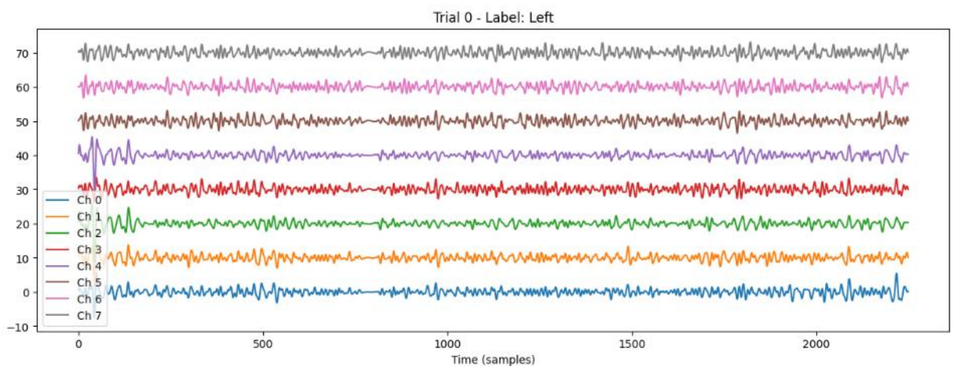


Figure 5 - EEG signals after baseline correction

- **1.2.5 Cropping Trial**

The task-relevant EEG segment is extracted between **2.5 and 6.0 seconds** after the trial onset. This window was selected empirically and after many expirments to focus on the period where motor imagery activity is expected to be most prominent.

- **1.2.6 Z-score Standardization**

The data is standardized per channel using z-score normalization. This ensures that all features are on the same scale and prevents channels with higher amplitude from dominating the learning process.

1.3 Feature Extraction

For each trial, the Power Spectral Density (PSD) is computed using Welch's method. The PSD is computed for each selected EEG channel (FZ, C3, CZ, C4, PZ) independently using the following parameters. The resulting power spectra are then filtered to retain only the frequency bins within the **8–30 Hz range**, corresponding to the **Mu (8–13 Hz)** and **Beta (13–30 Hz)** bands.

After extraction, all feature vectors are **standardized using z-score normalization** across the training set to ensure uniform scaling.

1.4 Classification

Multiple machine learning classifiers were evaluated using grid search cross-validation, with performance assessed through Leave-One-Subject-Out Cross-Validation (LOSO-CV) across all 35 training subjects. The AdaBoost classifier achieved **62%** accuracy on the validation data and **49.8%** accuracy on the LOSO test set.

```
Training: AdaBoost with Standard
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.71      0.68        28
           1       0.58      0.50      0.54        22

    accuracy                           0.62        50
   macro avg       0.61      0.61      0.61        50
weighted avg       0.62      0.62      0.62        50

[Standard] AdaBoost Best Params: {'learning_rate': 1.0, 'n_estimators': 100}
[Standard] Val Accuracy: 0.6200
```

2. Steady-State Visually Evoked Potential (SSVEP)

In this task, the objective is to classify EEG signals into one of four visual stimulus directions—**Forward, Backward, Left, and Right**—based on Steady-State Visual Evoked Potential (SSVEP) responses. While traditional techniques like Canonical Correlation Analysis (CCA) and Filter Bank CCA (FBCCA) have shown strong performance in within-subject SSVEP classification, their effectiveness diminishes in cross-subject scenarios due to variability in subjects' frequency responses, noise tolerance, and channel-specific sensitivity.

To address these challenges, we developed a robust and scalable pipeline that combines **advanced preprocessing (notch filtering, bandpass filtering, normalization, and epoching)** with **enhanced FBCCA feature extraction across multiple filter banks**. These features are then classified using a **linear Support Vector Machine (SVM)** with class-balanced training. This approach significantly improves generalization across subjects while maintaining interpretability and computational efficiency.

2.1 Signal Preprocessing

The EEG signal preprocessing pipeline is implemented through the advanced_preprocess_eeg function and is carefully designed to enhance the signal-to-noise ratio (SNR), remove common artifacts, and extract task-relevant SSVEP activity. Each step in the pipeline targets specific challenges of SSVEP data, especially in cross-subject settings, where robustness and consistency are critical.

2.1.1 Bandpass Filter

A **4th-order Butterworth bandpass filter** is applied to all EEG channels, with a frequency range of **5–40 Hz**. This range is selected to retain the fundamental SSVEP stimulus frequencies (7, 8, 10, and 13 Hz) and their first few harmonics while suppressing low-frequency drifts and high-frequency noise.

This broad band ensures we preserve both primary and harmonic frequency components important for FBCCA-based feature extraction.

2.1.2 Notch Filtering

Prominent 50 Hz and 60 Hz spikes were observed in the EEG's Power Spectral Density (PSD), typical of power line interference in countries like Egypt. To mitigate this:

- Notch filters are applied at 50 Hz and 60 Hz, along with their first two harmonics (100/120 Hz).
- Each filter uses a Q-factor of 30 to maintain a narrow attenuation band and avoid distortion of neighboring frequencies.

This step is essential to prevent artificial frequency spikes from contaminating frequency-based features like PSD and CCA correlation values.

2.1.3 Z-score Outlier Suppression

To reduce the influence of sudden noise bursts or muscle artifacts:

- A z-score thresholding mechanism is applied to each EEG channel.
- Samples where $|z| > 5$ are replaced with the median of the channel to preserve structure without harsh clipping.

This ensures occasional high-amplitude outliers (e.g., spikes or jerks) do not skew frequency-domain calculations.

2.1.4 Channel-wise Normalization

After filtering and artifact suppression, each EEG channel is standardized using:

Zero-mean, unit-variance normalization

Computed per channel using:

$$\boldsymbol{x_{norm}} = \frac{\boldsymbol{x - \mu}}{\boldsymbol{\sigma}}$$

This step ensures consistent scaling across all electrodes, stabilizing the downstream feature extraction process (FBCCA, PSD, etc.).

2.1.5 Common Average Referencing (CAR)

To further enhance spatial SNR:

- A Common Average Reference (CAR) is computed across all available EEG channels.
- Each channel's signal is re-referenced by subtracting the mean of all channels at each time step.

This improves SSVEP signal localization by suppressing global noise components and aligning channel amplitudes to a common baseline.

2.1.6 Trial Cropping (Epoching)

Rather than using the entire trial, only the most relevant time window is extracted:

- Trials are cropped to the 1s–5s interval, resulting in 1000 samples per trial.
- This segment empirically captures the most stable and task-relevant SSVEP response, while avoiding:
  - Transient visual onset artifacts at the beginning
  - Subject fatigue or drift at the end

This selection was based on time-frequency visualizations and validation accuracy optimization experiments.

2.2 Feature Extraction

We used Extended Filter Bank Canonical Correlation Analysis (FBCCA) to extract discriminative features from the EEG signals. This method improves upon traditional CCA by applying a series of bandpass filters before feature computation, making it more robust to individual variability in frequency responses.

2.2.1 Filter Bank Design

To capture both the fundamental stimulus frequency and its harmonics, we applied eight bandpass filters using a 4th-order Butterworth design. These filters covered overlapping frequency bands from 5 Hz to 38 Hz, such as:

- (5–15 Hz), (6–14 Hz), (10–20 Hz): capturing fundamental SSVEP responses
- (14–22 Hz), (22–30 Hz), (30–38 Hz): capturing higher harmonics
- Additional overlapping bands improve frequency resolution and generalization.

This filter bank ensures we capture relevant information even when subjects' responses are shifted slightly from the stimulus frequency due to noise or brain variability.

2.2.2 Canonical Correlation Computation

For each band, we:

1. Bandpass-filtered the EEG.
2. Generated reference signals (sine/cosine) at the 4 stimulus frequencies and their harmonics.
3. Computed the canonical correlation between the EEG and reference signals.
4. Stored the raw and squared correlation values as features.

This resulted in a 64-dimensional feature vector per trial (32 raw + 32 squared correlations), representing the alignment between EEG signals and visual stimuli.

2.3 Classification

A variety of machine learning models were initially evaluated. The **linear Support Vector Machine (SVM)**, configured with `class_weight='balanced'`, was selected as the final model due to its superior performance and efficiency in this high-dimensional feature space.

The model's ability to generalize to unseen subjects was rigorously assessed using **Leave-One-Subject-Out Cross-Validation (LOSO-CV)** across the 30 training subjects. This validation strategy ensures that the model is always tested on data from a subject it has never seen before. The final pipeline achieved a robust **mean accuracy of 62.0%** in this cross-subject validation setting.

In conclusion, our two approaches achieved a total score of **0.67437** in the competition. Although we were unable to explore neural network-based methods due to time and resource constraints, we were able to develop a solution leveraging Machine Learning for Brain-Computer Interface (BCI) control systems.

Team Name: Brain Decoders

Team Name in Kaggle: The Brain Decoders