EEG SEMANTIC CATEGORY CLASSIFICATION

OVERVIEW

This project explores the use of electroencephalographic (EEG) signals for identifying **semantic categories** in the human brain. The work was originally developed as part of my thesis, using a wearable EEG device (Emotiv EPOC).

The goal was to classify brain responses into three categories:

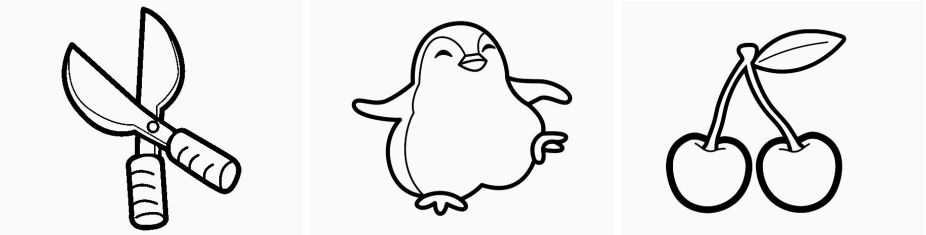
- Tools
- Animals
- Flowers/Fruits

This type of research contributes to the development of **Brain-Computer Interfaces (BCI)**, which can improve life quality for people with speech or communication disorders.

EXPERIMENT CONTROL

EEG recordings were collected with an **Emotiv EPOC** headset using **OpenViBE** scenarios and Lua scripts. These scripts are included in the `/BCI/` folder for reproducibility and can be replayed either with simulated EEG signals or real headsets connected via **CyKit** or alternative libraries.

Finally, the **stimuli** presented to the subjects is in the `/BCI/stimuli` folder, but here are some examples:



DATASET

With the working OpenViBE interface, **10 participants** completed the task, where 90 visual prompts (belonging to three categories) were presented every three seconds. Participants classified each stimulus using keyboard buttons: **left, down, and right,** corresponding to the **Animals, Tools, and Flowers/Fruits** classes.

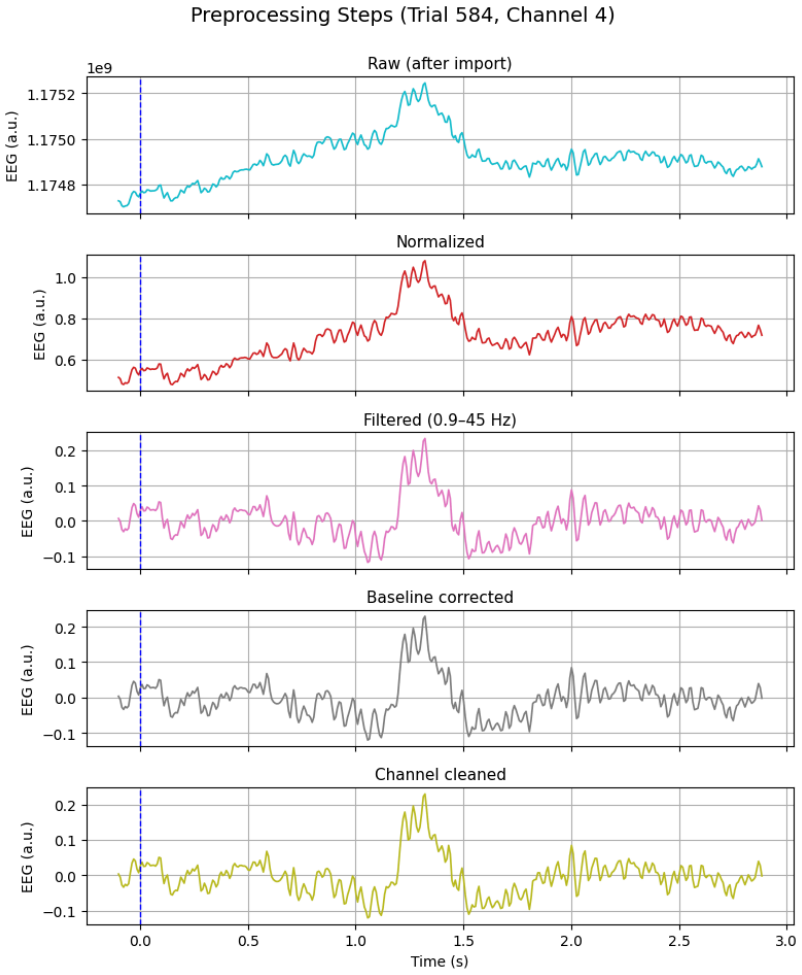
To reduce fatigue and maintain concentration, the session for each participant was split into **three trials of 30 stimuli each**.

The final dataset consists of **900 single-trial EEG records** (300 per class), organized into `/Data/user1`, `/Data/user2`, ... `/Data/user10` folders, together with the ground-truth labels (`targets`) and the participants' responses (`answers`).

METHODOLOGY & RESULTS

Signal Preprocessing

The signal preprocessing included the Epoch extraction, Normalization per user, Filtering (0.9-45 Hz), Baseline Correction, and Channel Selection per trial. The complete data loading and preprocessing methodology is in the `/Preprocessing/eeg_prep.ipynb` file. Here's an example of the signal preprocessing steps on a random trial, random channel:

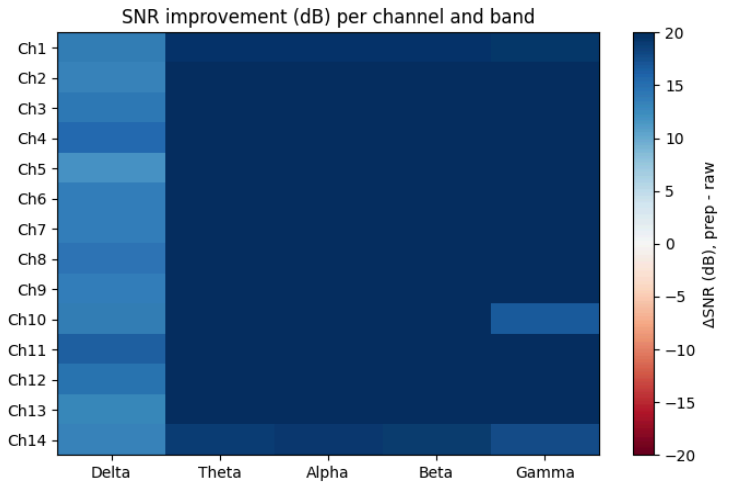


Additionally, I calculated the Signal to Noise Ratio (SNR) of the raw and preprocessed EEG signals considering the meaningful (signal) power as the EEG bands and the meaningless (noise) power as the power outside the interesting signal bands.

The interesting EEG frequency bands are:

- **Delta:** 0.5-4 Hz
- **Theta:** 4-8 Hz
- **Alpha:** 8-13 Hz
- **Beta:** 13-30 Hz
- **Gamma:** 30-45 Hz

The following figure shows the SNR improvement in dB per channel and frequency band of the raw EEG vs. the Preprocessed EEG.



Demonstrating that **the preprocessing improved the quality of the recorded EEG signals** in every channel and every frequency band.

Feature Extraction

The feature extraction included time-domain features (for example the response time) and frequency-domain features (such as the PSD in the EEG interesting bands: alpha, beta, theta, gamma and delta). Finally, statistical features of the ERP components (P300, N400) were also included.

The detailed feature extraction process is in `/Features/feature_extraction.ipynb` and the resulting `RT.csv`, `X_features.csv`, `X_final.csv`, `X.csv`, and `y.csv` files are in the `/Data/` folder.

Classification

1. Machine Learning

For the ML classification, I evaluated six classifiers (Logistic Regression, SVM, Random Forest, Naïve Bayes, MLP, and LightGBM) on two feature sets (`X_features` and `X_final`) using three **feature selection** strategies: Variance Thresholding, Correlation Filtering, and their combination.

- The best performance came from **`X_final` + Correlation Filtering**, with models **MLP** (Acc ≈ 0.39, F1 ≈ 0.38) and **Logistic Regression** (Acc≈0.38, F1≈0.38).
- **SVM, Naïve Bayes, and LightGBM** models performed close to random guessing, showing poor generalization.
- Variance Thresholding alone was ineffective, while Correlation Filtering gave small but consistent gains by reducing redundancy.
- Cross-validation confirmed that **MLP and LogReg generalize best**, though accuracy remained modest (Acc ≈ 0.3–0.4).

Overall, **simpler models (LogReg, shallow MLP)** proved most suitable, with correlation-based feature selection offering incremental benefits.

`/Classification/classification_ML.ipynb` contains the details in the classification and evaluation processes via Machine Learning.

2. Deep Learning

For the DL classification, I trained two end-to-end models directly on the EEG matrix input (`X`): a **CNN** and a **CNN+LSTM hybrid**. To improve robustness, five **data augmentation** (DA) strategies were considered:

- **None**: baseline.
- **Reversed**: time-axis reversal.
- **Shuffled**: permuting channel order.
- **Noise**: adding Gaussian noise.
- **Ch_drop**: randomly dropping one channel.

DA was applied only to the training set to prevent leakage.

- The **CNN model consistently outperformed CNN+LSTM**, which often struggled to converge.
- **Reversing and Noise augmentation** gave the best improvements, reaching **Acc = 0.40–0.41**, slightly above baseline.
- **Channel dropping** degraded performance, confirming the importance of data completeness.
- Under **Leave-One-Subject-Out CV**, CNN showed strong variability: some subjects reached Acc ≈ 0.39, others dropped near chance (0.25–0.30).

DL models showed promise, but **generalization across users remained limited**.

`/Classification/classification_DL.ipynb` contains the details in the classification and evaluation processes via Machine Learning.

LIMITATIONS

- Each subject provided only **90 trials (=4.5 min of data)**, insufficient for robust training.
- Only **10 users** were recorded, making cross-subject generalization especially challenging.
- EEG data is inherently **noisy and user-dependent**, requiring larger datasets for stable performance.

FUTURE WORK

To advance this research, I propose to:

- **Expand the dataset** (more trials, more users, more examples per class) or use another *larger* database.
- **Enrich features** with time-frequency transforms (wavelets, spectrograms).
- **Regularize and adapt** with transfer learning, domain adaptation, or adversarial strategies.
- **Combine ML & DL** (hybrid models mixing handcrafted features with CNN embeddings).
- **Subject-specific fine-tuning** after pretraining to improve personalization.

Finally, while the current results are modest, they confirm that **EEG-based semantic category classification is feasible**. With larger and more diverse data, plus stronger generalization strategies, performance can improve significantly.