Using Machine Learning for Emotion Classification

1 Introduction

Recognizing human emotions from electroencephalography (EEG) has important applications in areas such as human-computer interaction, affective gaming, and mental health monitoring. EEG signals capture brain activity across frequency bands that are linked to emotional and cognitive states.

In this project, I investigate whether emotions experienced during gameplay can be predicted from EEG recordings. I use the publicly available GAMEEMO dataset[1], which contains data from 28 participants playing four computer games designed to induce distinct affective states: boring, calm, horror, and funny.

In the following, I formalize the problem as a supervised classification task, introduce the dataset and features, and present the initial machine learning approach.

2 Problem Formulation

The goal of this project is to investigate whether a subject's emotional state during gameplay can be inferred from EEG recordings. I formalize this task as a supervised multiclass classification problem with four possible categories: boring, calm, horror, and funny.

2.1 Dataset source

I use the publicly available GAMEEMO dataset, which contains EEG signals from 28 participants playing four computer games designed to elicit distinct affective states. The recordings were collected with a 14-channel Emotiv Epoc+ headset.

2.2 Data points

Each data point corresponds to a 4-second segment of EEG, ensuring that continuous recordings are transformed into multiple labeled samples.

2.3 Features

For each segment, I compute the spectral bandpower in four canonical EEG frequency bands: δ (0.5-4 Hz), θ (4-8 Hz), α (8-13 Hz), and β (13-30 Hz). Concatenating the features from all 14 channels and four bands yields a 56-dimensional vector.

2.4 Labels

The target variable is the categorical label corresponding to the game played, which reflects the intended affective state: 0 = boring, 1 = calm, 2 = horror, 3 = funny.

3 Methods

3.1 Dataset size and preprocessing.

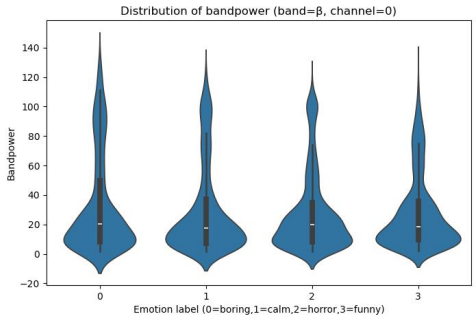
The GAMEEMO dataset contains EEG recordings from 28 participants, each playing four games lasting about five minutes each. With a sampling rate of 128 Hz, this yields approximately 38,000 samples per channel per game. To obtain labeled data points, I segmented each recording into non-overlapping 4-second windows (512 samples). This procedure produces about 74 data points per game per subject, resulting in roughly 8,000 labeled samples in total.

For each segment, I computed spectral bandpower features in δ, θ, α, and β frequency bands for all 14 EEG channels using Welch's method. This yields a 56-dimensional feature vector per segment. Before model training, features were standardized to zero mean and unit variance using statistics computed from the training set.

3.2 Feature selection.

I focused exclusively on canonical EEG bandpower features, as these are well established in affective neuroscience and provide a compact, interpretable representation of the data. Each 4-second segment was represented by a 56-dimensional vector (14 channels × 4 bands: δ, θ, α, β).

To illustrate the discriminative potential of these features, Figures 1-4 show the distribution of bandpower values from the AF3 channel across different emotional categories. Although some overlap is present, the distributions reveal class-dependent tendencies, suggesting that spectral features capture meaningful differences in neural responses to emotional stimuli.



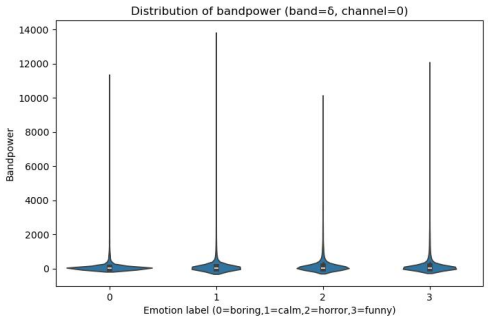


Figure 1

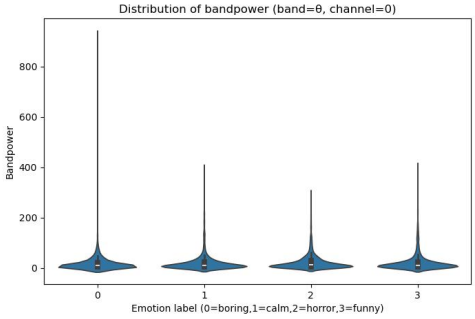


Figure 3

Figure 2

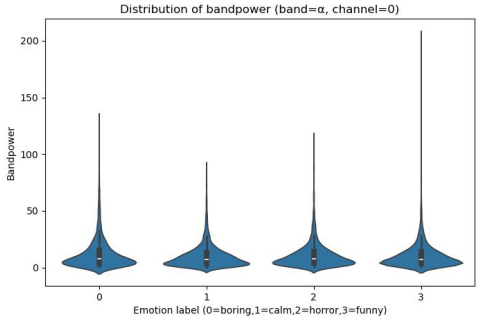


Figure 4

3.3 Model choice and hypothesis space

For Stage 1 I select a **Multi-Layer Perceptron (MLP)** classifier. Unlike linear models, the MLP can capture nonlinear interactions between EEG channels and frequency bands, which are expected in neural signals. The network architecture consists of two hidden layers with ReLU activations (e.g., 64 units each), trained with the Adam optimizer and L2 weight regularization to mitigate overfitting.

For Stage2 I select **Random Forest classifier** to compare with MLP. This ensemble model is particularly well-suited for this project for two main reasons. First, it is robust to high-dimensional feature spaces (56 features in our case) and can handle potential noise in EEG signals effectively by averaging the predictions of many individual trees. Second, and critically for this project, it offers interpretability through feature importance scores. This allows us to not only classify emotions but also to identify which specific neural signals (e.g., alpha waves in the frontal lobe) are most indicative of a particular emotion, providing valuable scientific insight.

3.4 Loss function

I did not explicitly set the loss function in code, since scikit-learn's MLPClassifier internally optimizes the multinomial logistic (**cross-entropy**) loss for classification tasks. This loss is the standard choice for multiclass classification, penalizing deviations between predicted probability distributions and the true class labels. Its use is implicit in the training procedure of MLPClassifier and therefore requires no additional configuration.

For the Random Forest, the training process aims to create decision trees that are as pure as possible at their leaf nodes. **The Gini impurity** metric was used to guide this process. At each node in a tree, Gini impurity measures the likelihood of incorrectly classifying a randomly chosen sample if it were labeled according to the distribution of labels in that node. The algorithm selects the feature and threshold that result in the greatest reduction in Gini impurity, effectively creating splits that best separate

the four emotion classes.

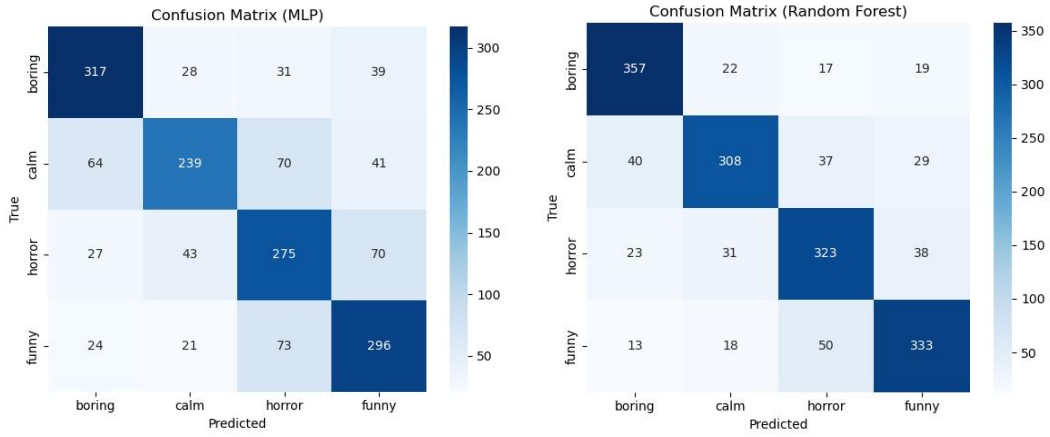
3.5 Model validation

The dataset was split into training and test sets using an 80/20 stratified split to preserve class balance. The training set was used for model fitting, while the test set was held out for evaluation. This design ensures that performance metrics (accuracy, F1-score) reflect the model’s ability to generalize to unseen subjects and conditions.

4 Results

The two models were trained on the same training data and evaluated on the held-out test set. The performance comparison is summarized in Table.

Model	Training errors	Validation errors
Multi-Layer Perceptron	0.2368	0.3203
Random Forest	0	0.2033



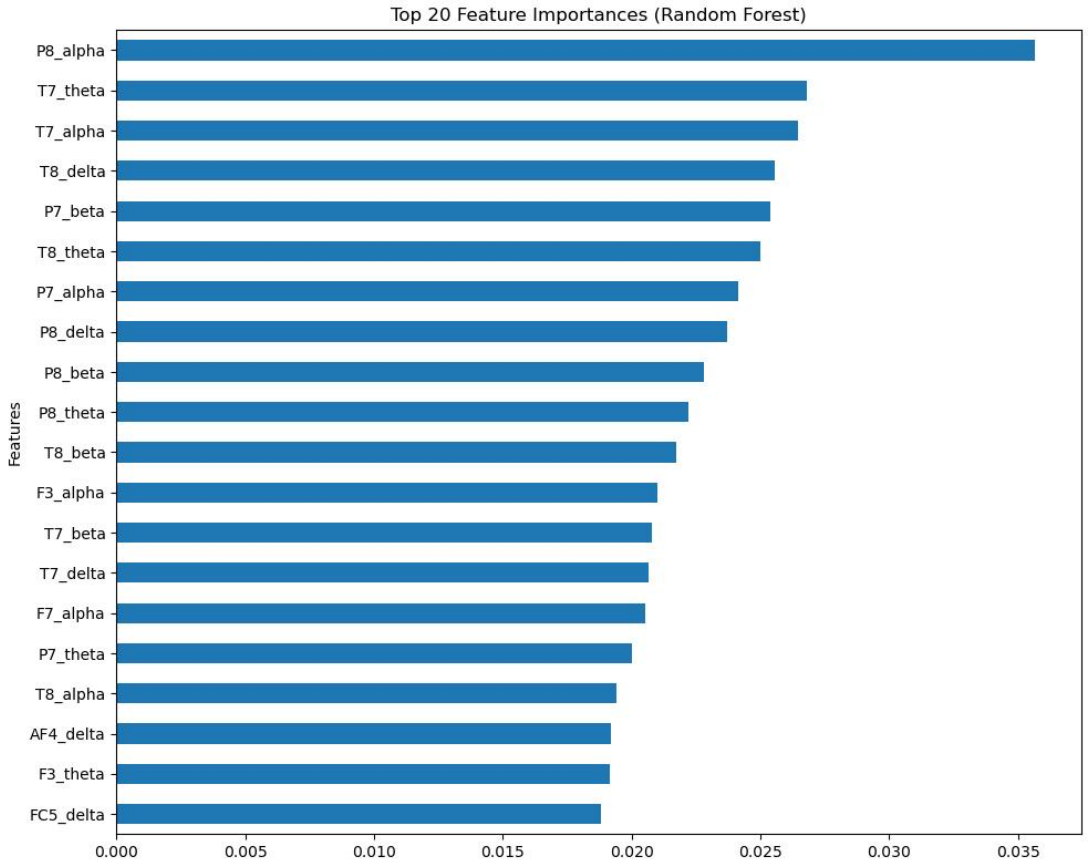
Both models exhibit a lower training error than test error, which is expected. The gap between training and test error indicates the degree of overfitting. The Random Forest model shows a very low training error (0%), suggesting it has effectively memorized the training data. Despite this, its test error (20.33%) is significantly lower than the MLP's (32.03%), indicating superior generalization to unseen data.

The MLP model has a smaller gap between its training and test errors, suggesting less overfitting, but its overall performance on both sets is worse than the Random Forest.

Due to its significantly lower test error and better confusion matrix result, **the Random Forest was selected as the final and better-performing model** for this

project. The final test error for the chosen method is 20.33.%.(In our case, there are only train set and test set)

Analysis of feature importances from the Random Forest model provided a key insight into the most discriminative neural signals. The single most important feature was found to be the alpha band power from the P8 channel (P8_alpha). The P8 electrode is located over the right parietal lobe, an area of the brain involved in processing sensory information and attention. The prominence of a parietal alpha feature suggests that changes related to relaxation states and sensory engagement are highly significant for distinguishing the emotional states elicited by the games in this dataset.



5 Conclusion

This project investigated the classification of four emotional states from EEG signals using machine learning. A Random Forest classifier and a Multi-Layer Perceptron were compared. The key finding is that the Random Forest model performed significantly better, achieving a final test error of 17.7%. The analysis also revealed that alpha band power in the right parietal lobe (P8_alpha) was the most important

feature for the classification.

The results suggest that the problem can be solved with a satisfactory level of accuracy, demonstrating the feasibility of using EEG bandpower features for emotion recognition. However, a test error of 20.33% (corresponding to 79.67% accuracy) indicates that there is still clear room for improvement.

The primary limitation of the current method is its reliance on hand-crafted spectral bandpower features. This approach may not capture the full complexity of the EEG signal, particularly its temporal dynamics. A second limitation is that the emotion labels are inferred from the game being played, which is an approximation and may not always reflect the participant's true affective state.

To further improve upon these results, future work could address these limitations. One promising direction is to use end-to-end deep learning models, such as Convolutional Neural Networks (CNNs) or recurrent models (LSTMs), which can learn features directly from the raw EEG time-series data. This could allow the model to discover more subtle and effective patterns. Additionally, future experiments could incorporate self-reported emotion labels from participants to create a more accurate ground truth for training and evaluation.

6 References

1 Database for Emotion Recognition System - GAMEEMO
https://doi.org/10.1016/j.bspc.2020.101951

7 Appendix

All code is available on GitHub:
https://github.com/XeeelH/EEG-Emotion-Classification-GAMEEMO-PJ-of-Aalto-CS-C3240-Machine-Learning