EEG-Based Epilepsy Seizure Classification Using Machine Learning

Faiza

23 April 2025

Contents

1 Introduction

Epilepsy is a neurological condition marked by repeated seizures due to sudden, uncontrolled bursts of electrical activity in the brain. It affects people of all ages and can severely impact quality of life if not properly managed. Accurate detection and classification of seizures are essential for providing timely treatment and care.

Electroencephalography (EEG) is one of the most reliable techniques for observing brain activity. EEG-based seizure detection systems can assist medical professionals by automating the identification of seizures, reducing diagnostic errors, and enabling continuous patient monitoring.

This study aims to utilize machine learning models to analyze EEG signals for the classification of seizure types. It compares model performance before and after feature engineering and balancing the dataset using SMOTE. Three widely used models—XGBoost, Random Forest, and LightGBM—are evaluated.

1.1 Motivation

Manual EEG analysis is time-consuming, prone to human error, and inefficient for long-term monitoring. With the growing volume of healthcare data and advancements in machine learning, it is possible to automate EEG interpretation for better accuracy and efficiency. Our work explores how machine learning can transform raw EEG data into meaningful clinical insights.

2 Dataset Overview

The EEG dataset was collected from 6 patients diagnosed with focal epilepsy during presurgical evaluations at the American University of Beirut Medical Center from January 2014 to July 2015. Anti-seizure medications were halted to record habitual seizures.

2.1 Dataset Characteristics

- **Patients:** 6
- **Total Seizures:** 35
- **Sampling Rate:** 500 Hz
- **Electrodes:** 21 scalp electrodes (10-20 system)
- **Filtered Range:** 1.6Hz–70Hz (excluding 50Hz electrical interference)

Each labeled segment is of shape 19x500 (19 channels, 1 second of data). The dataset is divided into:

- 3034 samples of Complex Partial Seizures (Class 1)

- 705 samples of Electrographic Seizures (Class 2)
- 111 samples of Video-Detected Seizures without EEG change (Class 3)
- 3895 samples of Normal EEG (Class 0)

Total labeled instances: 7790, with an 90:10 train-test split.

3 Data Preprocessing and Visualization

3.1 Preprocessing

1. Loaded .npy files for training and testing sets using NumPy.
2. Reshaped 3D EEG matrices into 2D arrays for model compatibility.
3. Balanced the dataset using SMOTE to address class imbalance, especially for Classes 2 and 3.
4. Scaled/normalized the data if required for specific model implementations.

3.2 Data Exploration

1. Class Distribution: Visualization shows a dominant presence of Class 0, followed by Class 1. Class 3 is the smallest, with only 6 samples in the test set.

2. EEG Signal Visualization: Plotted waveforms for a Class 1 EEG sample show how brainwave activity varies across channels.

3. Average Signal Plot: Plots of average values for each channel per class highlight key differences in signal strength and waveform shape.

4. Correlation Heatmap: A symmetrical matrix visualizes correlation coefficients between EEG channels, showing high correlation in adjacent channels.

5. Energy Plot: Calculated and plotted average energy per class. Class 1 shows the highest average energy, while Class 0 has the lowest.

4 Visualization of Results

This section presents visual insights extracted during data preprocessing and model evaluation.

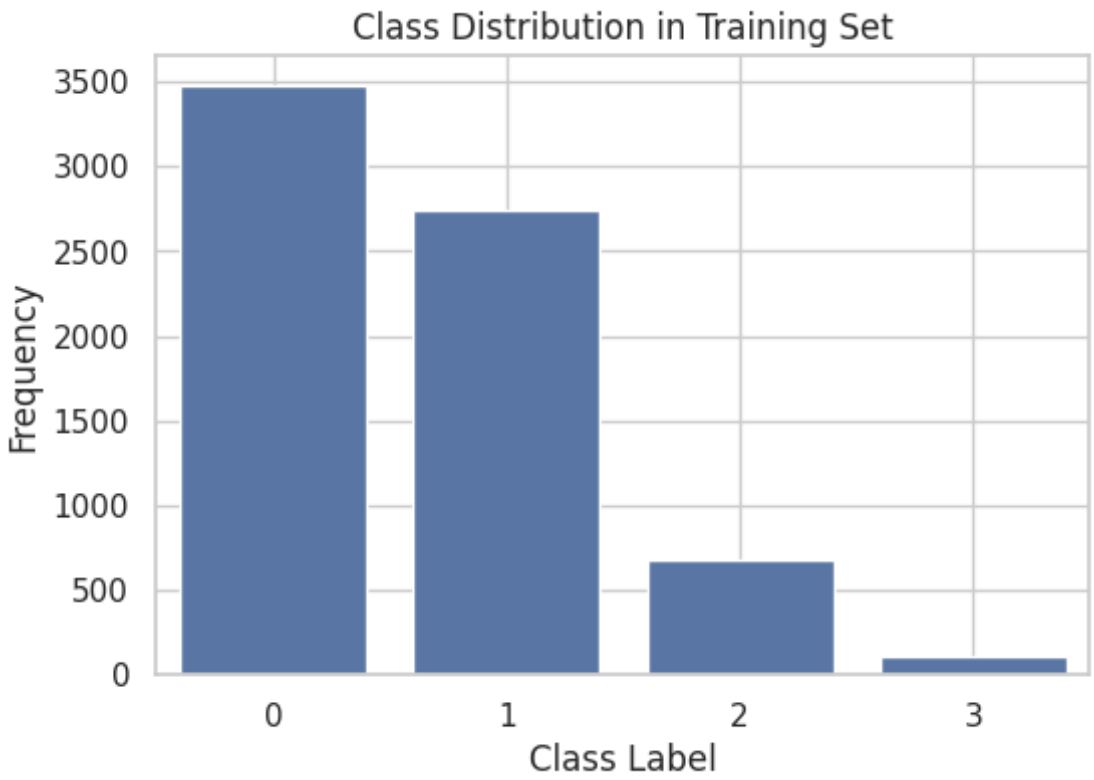

Figure 1: Distribution of EEG Samples Across Classes

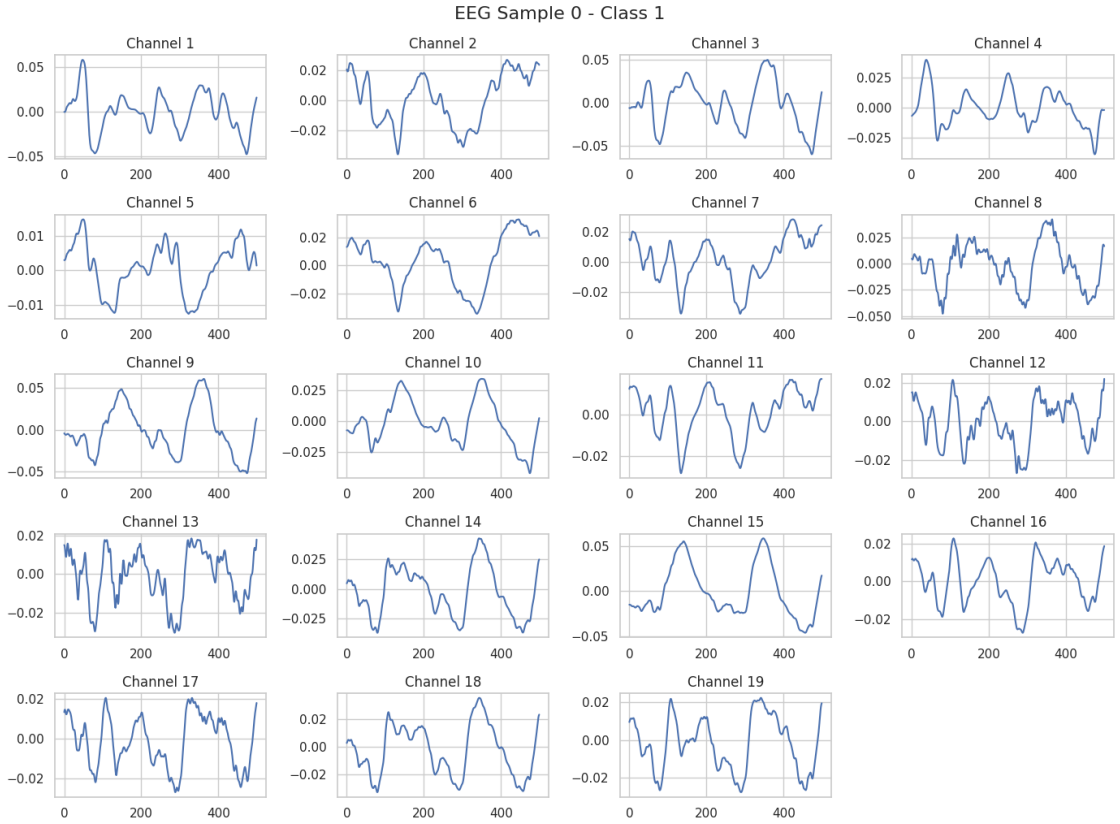

Figure 2: Raw EEG Signal from One Sample in Class 1 (Complex Partial Seizure)

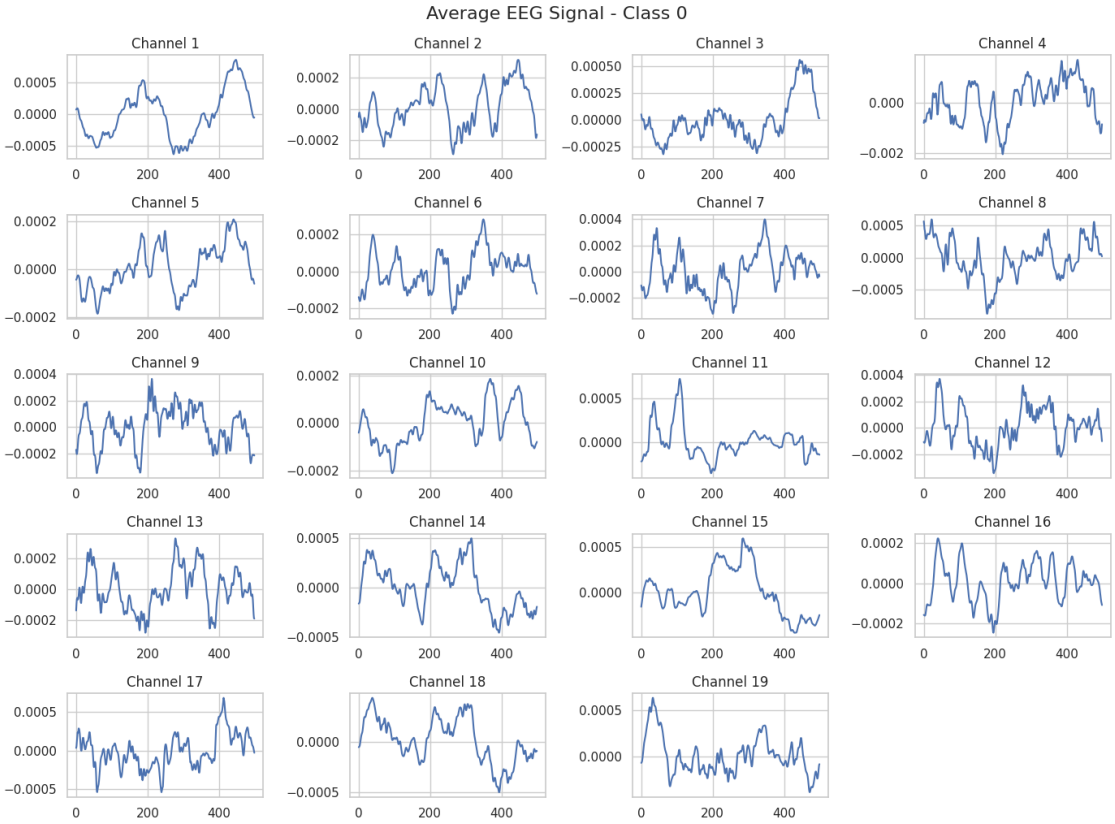

Figure 3: Average EEG Signals Across All Classes (Averaged Over Channels)

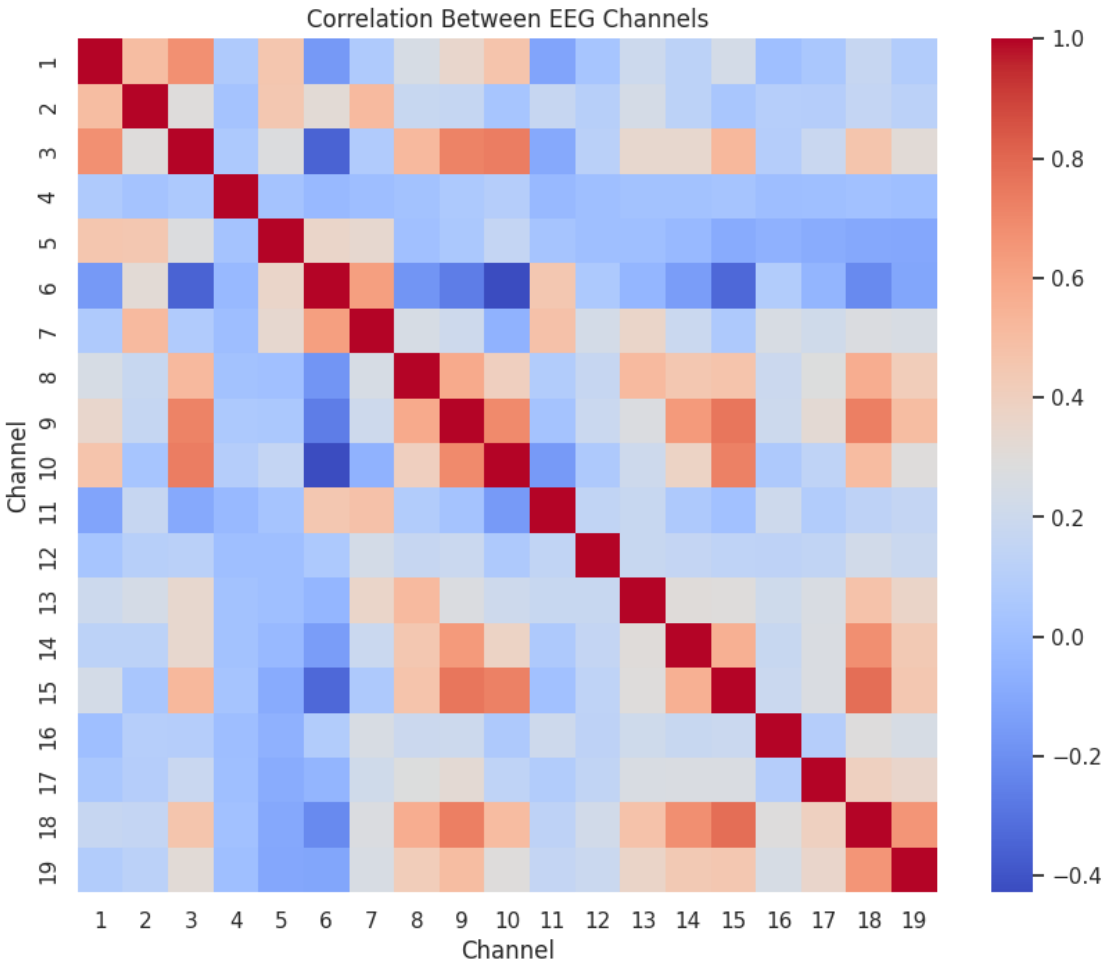

Figure 4: Correlation Heatmap Among EEG Channels

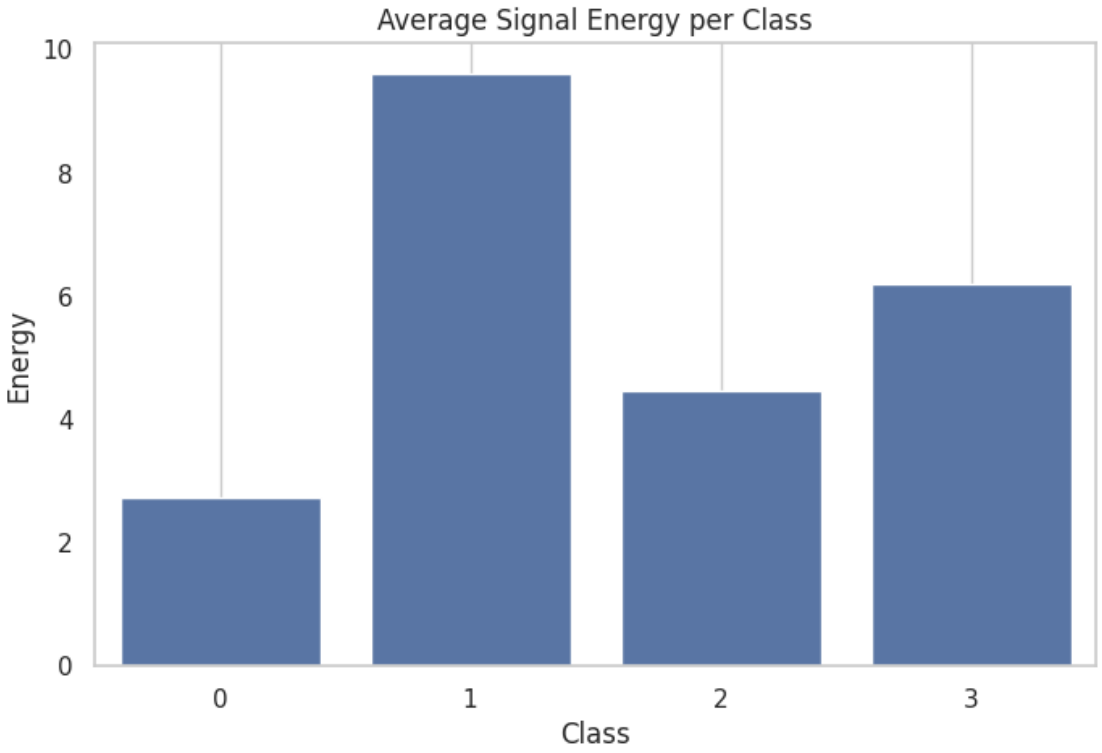

Figure 5: Comparison of Signal Energy Between EEG Classes

5 Machine Learning Models

5.1 XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting algorithm known for its speed and performance. It uses second-order gradient information to minimize the loss function and includes regularization to reduce overfitting.

5.2 Random Forest

Random Forest is an ensemble of decision trees built using bagging (bootstrap aggregation). It introduces randomness by selecting subsets of features during tree construction, thus improving model generalization. Its prediction is based on the majority vote from all trees.

5.3 LightGBM

LightGBM is a gradient boosting framework that grows trees leaf-wise (as opposed to level-wise in traditional GBDT). It is optimized for speed and memory efficiency and supports large-scale datasets with high dimensionality.

5.4 Mathematical Formulation

Random Forest prediction:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), ..., T_n(x)) \tag{1}$$

Where $T_i(x)$ is the output of the i^{th} decision tree.

Gradient boosting optimizes:

$$L(y, \hat{y}) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

Where l is the loss function and Ω is the regularization term.

6 Experiments and Evaluation

6.1 Without Feature Engineering

Raw reshaped EEG data was used for training:

Table 1: Performance Without Feature Engineering

Model	Accuracy (%)	F1-Score (%)
XGBoost	91.66	91.57
Random Forest	89.09	88.78
LightGBM	91.01	90.72

6.2 With Feature Engineering + SMOTE

The data was balanced using SMOTE and reshaped to improve model learning:

Table 2: Performance With Feature Engineering and Balancing

Model	Accuracy (%)	F1-Score (%)
XGBoost	91.66	91.72
Random Forest	88.70	88.83
LightGBM	91.40	91.46

6.3 Visual Comparison of Results

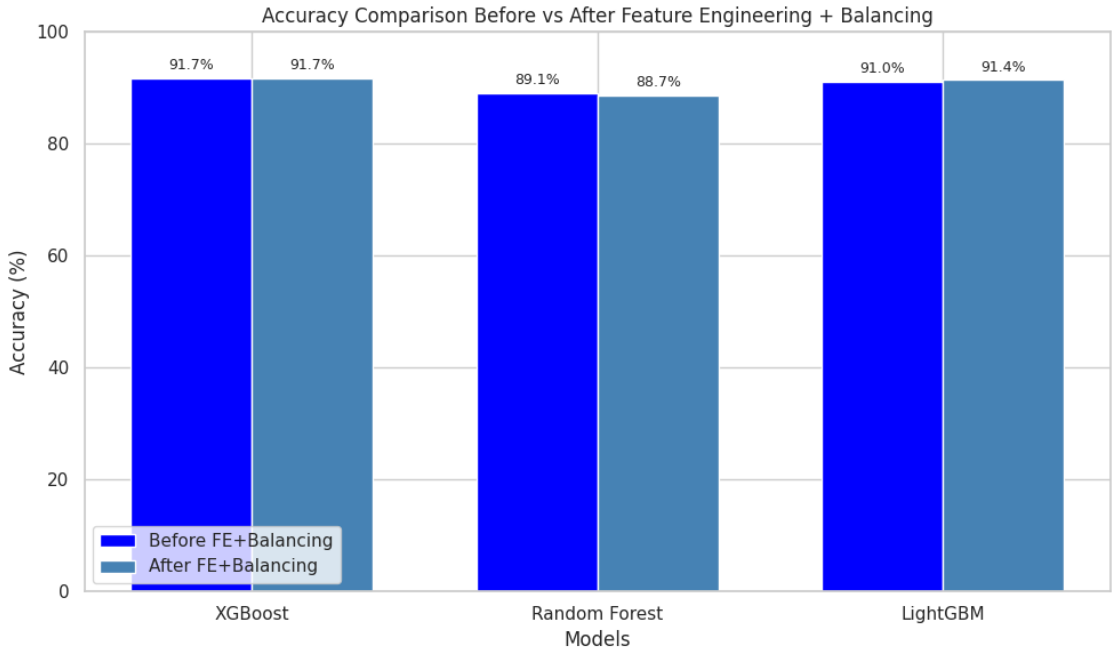

Figure 6: Comparison of Accuracy and F1-Score across Models (Before and After Feature Engineering)

6.4 Insights

- XGBoost has consistent high performance in both raw and engineered datasets.
- LightGBM performs similarly, with slightly lower variance and better speed.
- Random Forest lags slightly, especially in detecting rare classes.
- SMOTE significantly improves recall for Classes 2 and 3.

7 Conclusion

This project demonstrated the effectiveness of machine learning models in detecting different types of epileptic seizures from EEG data. We showed that:

- Feature engineering and balancing help improve minority class detection.
- XGBoost and LightGBM offer robust solutions with high accuracy.
- Random Forest is useful but slightly less accurate with imbalanced data.

Future improvements could include deep learning methods like CNNs or RNNs to automatically extract spatial-temporal EEG features.

8 References

- EEG Dataset: `https://data.mendeley.com/datasets/5pc2j46cbc/1`
- Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 2002.
- Chen, T., and Guestrin, C. "XGBoost: A Scalable Tree Boosting System." 2016.
- Ke, G., et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." NIPS 2017.
- Scikit-learn: `https://scikit-learn.org/`
- XGBoost Documentation: `https://xgboost.readthedocs.io/`
- LightGBM Documentation: `https://lightgbm.readthedocs.io/`