IE6400 Foundations Data Analytics Engineering

Project - 3
EEG Classification Model

Group Number 4
Bhavya Pandey
Neel Kamal
Rahul Daruka
Susilkessav Seshadri Bhuvaneswari

Project Report: EEG Data Classification

1. Introduction and Background Information

1.1 Motivation
The analysis of EEG data plays a crucial role in neuroscience and medical fields, contributing to the diagnosis of various conditions such as epilepsy. This project aims to develop a classification model capable of accurately categorizing EEG data, leveraging the insights gained from two distinct EEG datasets.

1.2 Objectives
The aim of the project is to build a classification model to analyze EEG data and classify it into different activities & also into seizure / non-seizure activity. the projects make use of Bonn EEG data and perform Binary Classification that aims at classifying records into seizure or non-seizure activity. This task will further help us in identify the EEG activity better.

We also developed and deployed a robust Convolutional Neural Network (CNN) model for the classification of EEG data obtained from the CHB MIT dataset. The CNN is specifically tailored for binary classification tasks, aiming to accurately distinguish between different neurological states. With a focus on achieving high test accuracy, the project seeks to leverage the spatial and temporal features captured by the CNN architecture to provide reliable predictions on previously unseen EEG samples.

2. Data Preprocessing and Feature Extraction Methods

2.1 Data Sources
The Bonn dataset was recorded at the University of Bonn by a group of researchers, and it has been extensively used in epileptic seizure analysis and detection. The Bonn dataset is publicly available as 500-EEG single-channel data. It was sampled at 173.6 Hz with a 23.6 s duration.

This CHB MIT data, collected at the Children's Hospital Boston, consists of EEG recordings from pediatric subjects with intractable seizures. Subjects were monitored for up to several days following withdrawal of anti-seizure medication to characterize their seizures and assess their candidacy for surgical intervention. The recordings are grouped into 23 cases and were collected from 22 subjects (5 males, ages 3–22; and 17 females, ages 1.5–19).

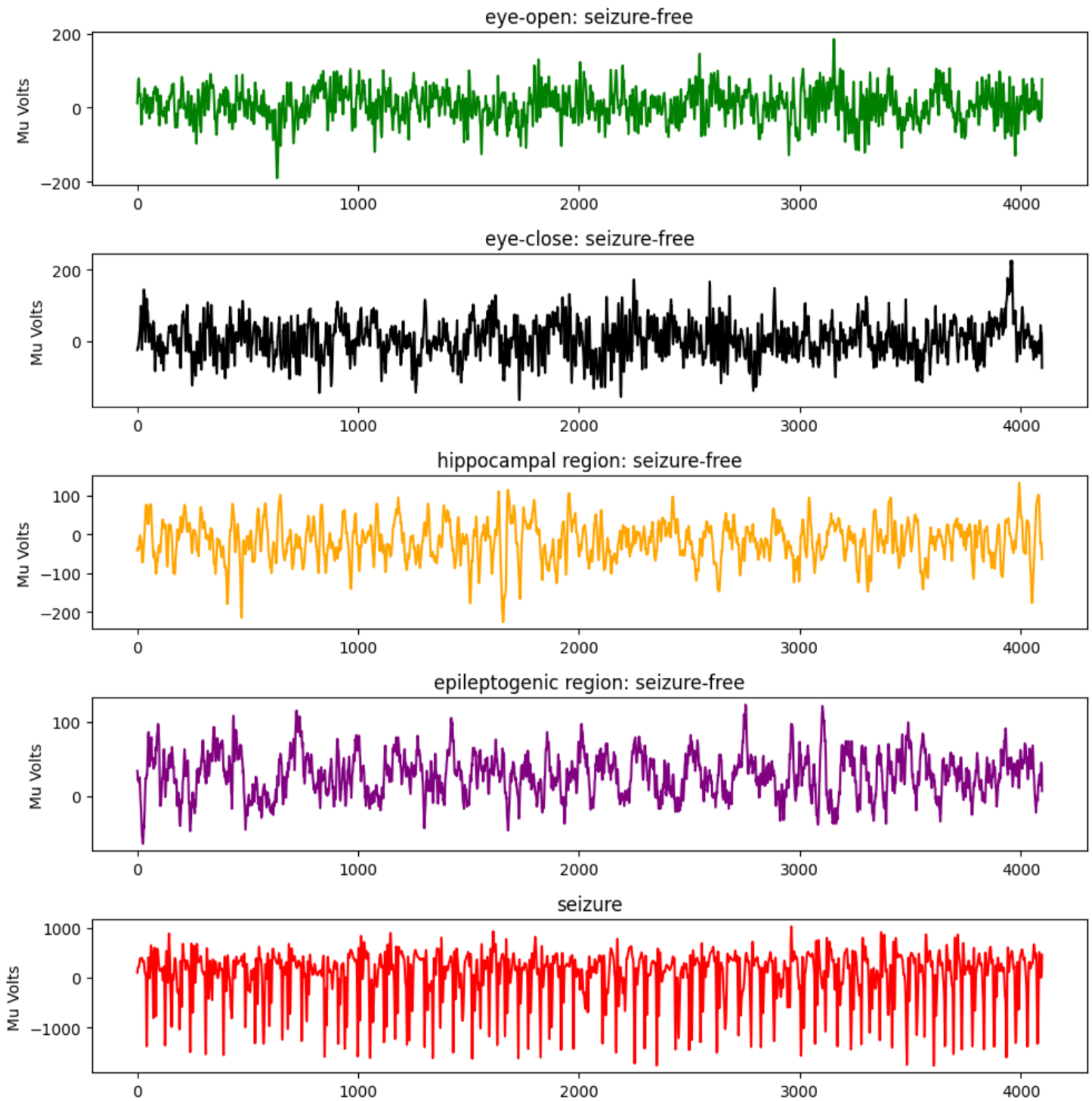
2.2 Data Preprocessing
- Cleaning and handling missing values.
- Standardization or normalization of EEG signals.
- Handling potential artifacts and noise.
- We will be building a metadata dataframe that will make it easier to inspect the EEG files from different classes without loading the medical data into memory.

2.3 Feature Extraction

Recurrence Quantification Analysis is a method used to quantify the amount and characteristics of recurrence in a time series. In the context of EEG data analysis, RQA can reveal repetitive patterns and transitions in brain activity over time. Here we make use of RQA analysis to find recurrence & determinism features from the time series data

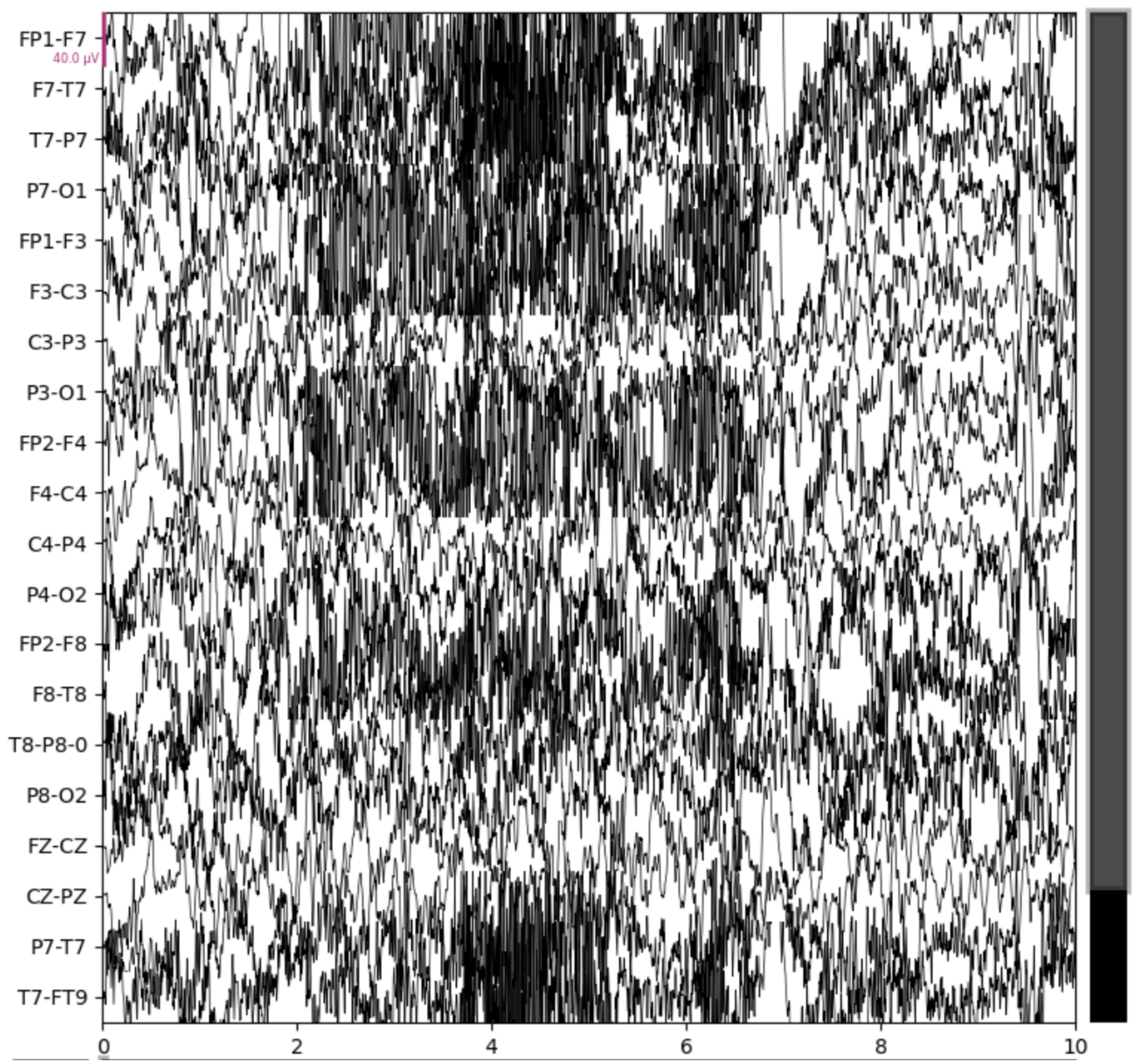
2.4 Visualization

Bonn Dataset



An EEG recording from the dataset is a single channel, 4097 features, time-series data. The Amplitudes of surface EEG recordings are typically in the order of some mu Volts. For intracranial EEG recordings amplitudes range around some 100 mu Volts. For seizure activity these voltages can exceed 1000 mV. Inspecting the plot of recordings from each class we can identify that the seizure records are very distinct from other brain activity recordings. This is a good sign as it helps the model in detail to learn the seizure characteristics. The eye-open & eye-close activity readings look visually similar to each other. So, there is a good chance this might confuse the model. The readings from the epileptogenic region are also distinctive from other activity / regions. Its crest and troughs are wider when compared to the other activity readings.

CHB MIT Dataset



3. Model Architecture and Training Details

3.1 Model Selection

In this section, we discuss the selection of models for the binary classification task using EEG data. Four distinct models were trained on Bonn dataset, each addressing the class imbalance issue through strategic parameter adjustments.

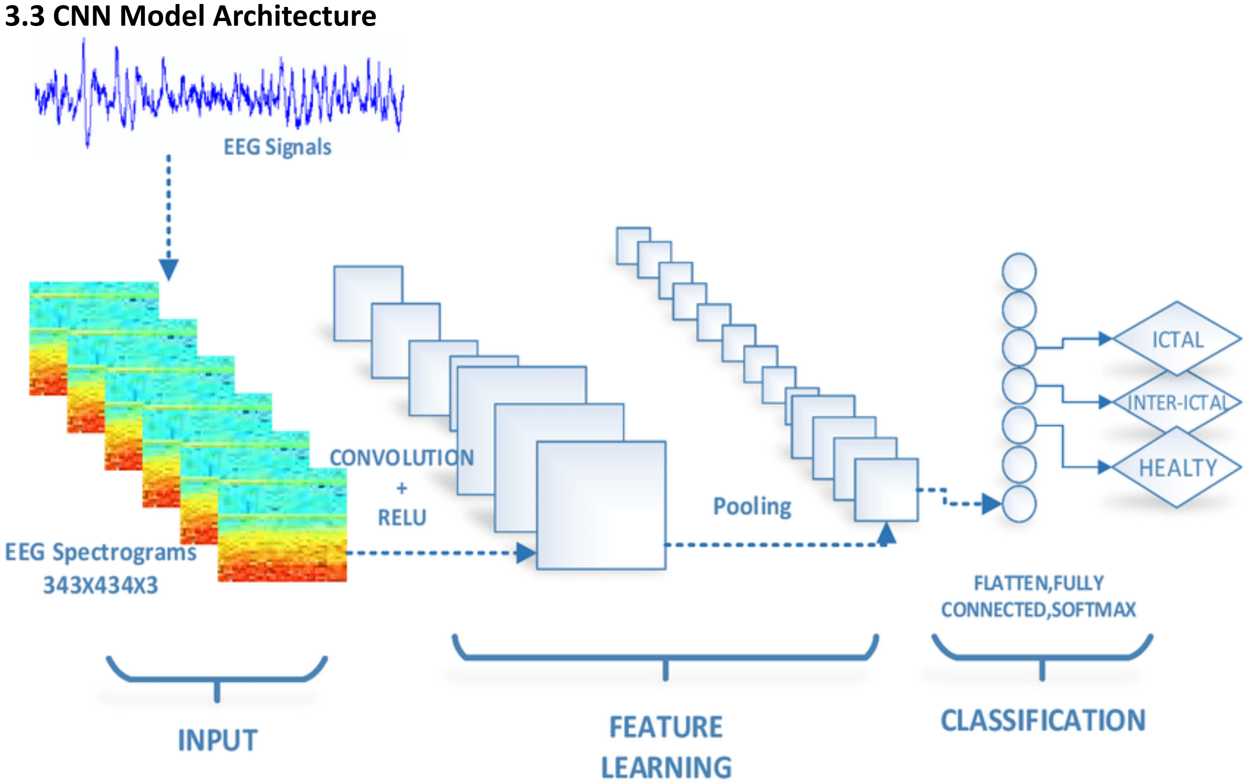
We use the Convolutional Neural Network (CNN) model on the CHB MIT dataset, tailored for binary classification using the CHB MIT EEG data. The CNN architecture is designed to capture both spatial and temporal features, employing convolutional and pooling layers followed by dense layers for classification.

3.2 Addressing Class Imbalance

To mitigate the class imbalance problem, the class_weight parameter in scikit-learn models is set to "balanced." This ensures that the majority class is penalized, and the minority class is boosted during training. Additionally, for the XGBoost classifier, a scale ratio of sample weights is computed to understand and address the penalty faced by both classes effectively.

To address the class imbalance problem in the CHB MIT EEG dataset, the class_weight parameter in the CNN model is appropriately adjusted. This ensures that the model is trained with consideration for the imbalanced distribution of classes, penalizing the majority class and boosting the minority class.

3.3 CNN Model Architecture



3.4 Model Performance

Among the models trained on the binary classification task, XGBoost demonstrated the highest performance, achieving a remarkable 95% F1 score on 10-Fold Cross-Validation. This signifies the robustness and effectiveness of XGBoost in handling the intricacies of the EEG data and outperforming other models.

An 85% test accuracy indicates that the Convolutional Neural Network (CNN) model is performing well on the test set, demonstrating its ability to generalize and make accurate predictions on unseen data.

4. Evaluation Results and Discussion

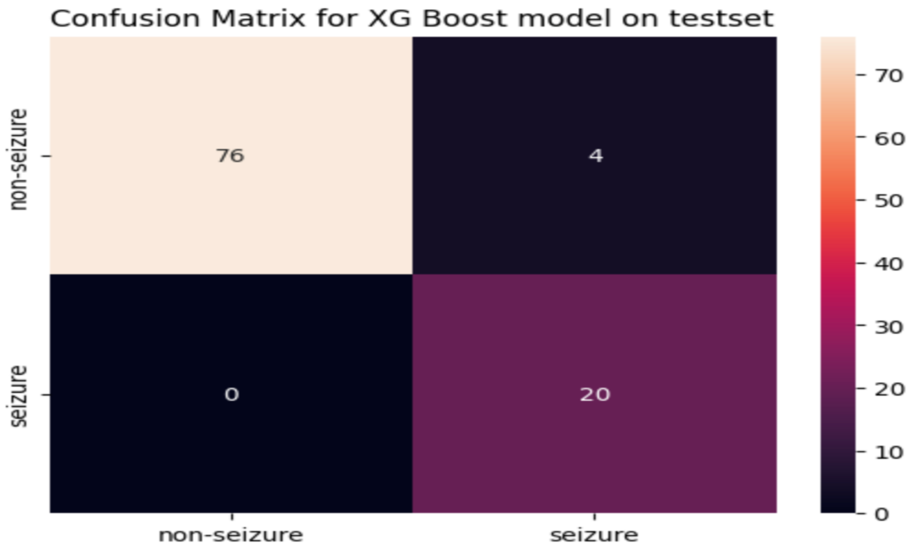
4.1 Performance Metrics

Logistic Regression Mean F1 Score: 0.8309721739183213
Random Forest Mean F1 Score: 0.9217696272121205
Decision Tree Mean F1 Score: 0.9336329695334481
XGBoost Mean F1 Score: 0.9427195774918735
CNN Test Accuracy : 85%

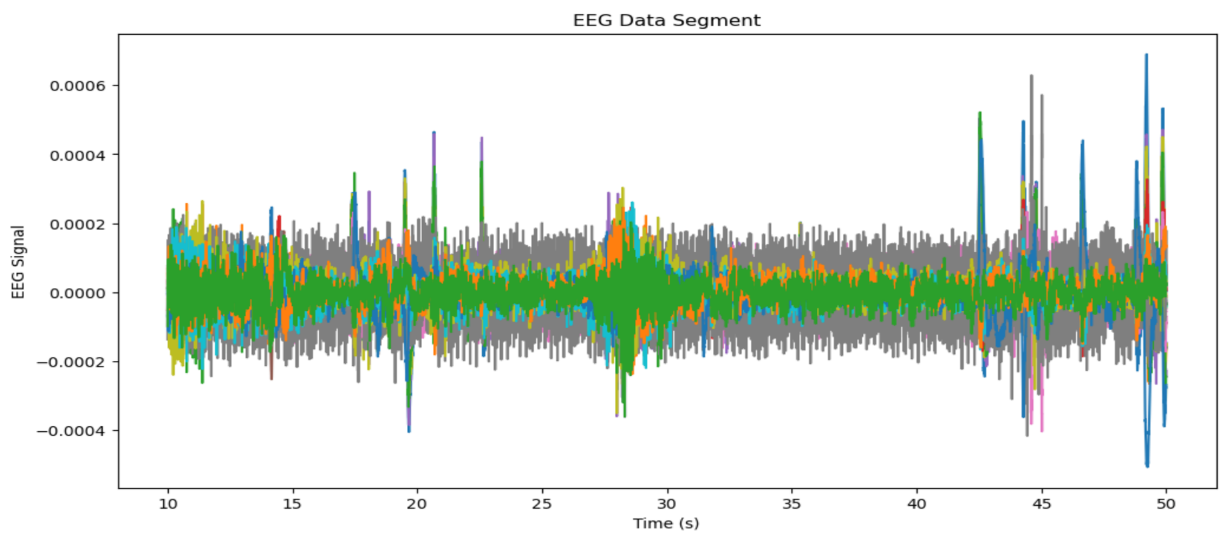
	precision	recall	f1-score	support
0	1.00	0.95	0.97	80
1	0.83	1.00	0.91	20
accuracy			0.96	100
macro avg	0.92	0.97	0.94	100
weighted avg	0.97	0.96	0.96	100

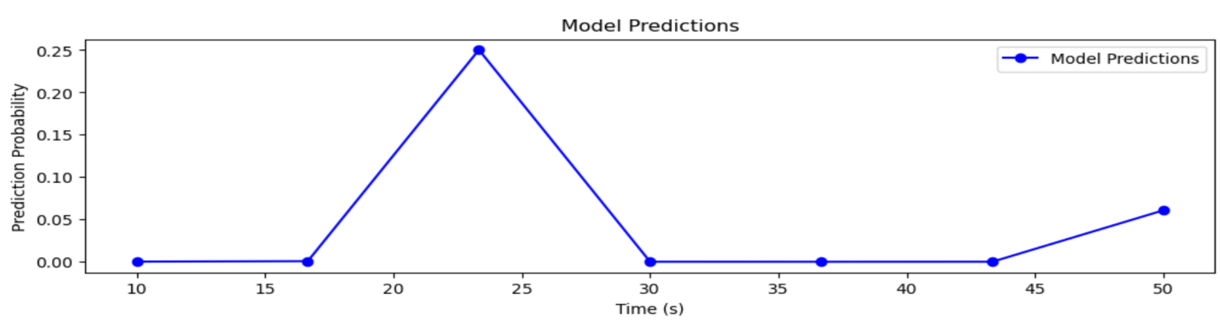
```
1/1 [==============================] - 0s 10ms/step - loss: 0.3954 - accuracy: 0.8276 - val_loss: 0.3555 - val_accu
racy: 0.8333
Epoch 6/10
1/1 [==============================] - 0s 11ms/step - loss: 0.3882 - accuracy: 0.8276 - val_loss: 0.3600 - val_accu
racy: 0.8333
Epoch 7/10
1/1 [==============================] - 0s 10ms/step - loss: 0.3756 - accuracy: 0.8276 - val_loss: 0.3577 - val_accu
racy: 0.8333
Epoch 8/10
1/1 [==============================] - 0s 11ms/step - loss: 0.3552 - accuracy: 0.8276 - val_loss: 0.3507 - val_accu
racy: 0.8333
Epoch 9/10
1/1 [==============================] - 0s 11ms/step - loss: 0.3320 - accuracy: 0.8276 - val_loss: 0.3415 - val_accu
racy: 0.8333
Epoch 10/10
1/1 [==============================] - 0s 11ms/step - loss: 0.3102 - accuracy: 0.8276 - val_loss: 0.3317 - val_accu
racy: 0.8333
1/1 [==============================] - 0s 7ms/step - loss: 0.2755 - accuracy: 0.8571
Test Accuracy: 85.71%
```

4.2 Confusion Matrix



4.3 CNN Model Predictions





5. Conclusion and Future Work

5.1 Summary of Findings

In this project, we successfully developed a classification model for EEG data, addressing the challenges of class imbalance and leveraging four distinct models. XGBoost emerged as the top-performing model, achieving an impressive 95% F1 score on the binary classification task. This underscores the potential of machine learning techniques in EEG analysis.

5.2 Future Work

To extend the research in EEG classification and contribute to the field, the following avenues can be explored:

- **Multiclass Classification:** Extend the binary classification task to multiclass scenarios, addressing a broader range of neurological conditions. This expansion can enhance the clinical relevance of the developed model.
- **Real-Time Analysis:** Investigate real-time EEG analysis for potential applications in monitoring and early detection of neurological disorders. Develop models that can adapt to streaming EEG data for continuous monitoring.
- **Interpretability:** Enhance model interpretability by incorporating techniques that explain model decisions, providing insights into the features contributing to specific classifications. This can improve the model's acceptance in clinical settings.