PROJECT 3

EEG CLASSIFICATION MODEL

IE6400 Foundations Data Analytics Engineering Fall Semester 2023

Group 22

Pramoth Guhan
guhan.p@northeastern.edu

Rohith Adhitya Chinnannan Rajkumar
chinnannanrajkumar.r@northeastern.edu

Divyia Venkat Eachampatti Thirunavukarasu
eachampattithiruna.d@northeastern.edu

Submission Date: 12/15/2023

Introduction and background information:

In this project, the primary objective is to develop a classification model for the analysis of Electroencephalogram (EEG) data, a critical component in neuroscience and medical applications, particularly in the diagnosis of epilepsy. The proposed approach will be able to identify seizure activity by classifying EEG data into discrete groups. The classification model will be trained on two large-scale EEG datasets and its performance will be assessed.

The CHB-MIT Scalp EEG Database, derived from the Children's Hospital Boston, focuses on EEG recordings from pediatric subjects experiencing intractable seizures and it is a resourceful compilation of EEG recordings obtained from patients diagnosed with epilepsy. This database offers a solid basis for training and testing the classification model because it includes a wide variety of seizure types in addition to non-seizure data. Annotations marking the beginning and conclusion of 182 seizures are notable in the dataset, as they aid in the model's learning. Epilepsy is characterised by seizures, which are brief interruptions in the electrical activity of the brain that frequently result in erratic and recurrent symptoms, such as brief attention deficits or convulsions involving the entire body. The creation and implementation of a trustworthy categorization model have great potential for resolving the difficulties associated with epilepsy by facilitating the timely detection and handling of seizures, whether via medication delivery or carer notifications. The objective of this endeavour is to reduce the possible repercussions of seizures and improve the general quality of life for those who have epilepsy.

Data pre-processing and feature extraction methods:

In the provided code, data pre-processing and feature extraction are crucial steps in preparing Electroencephalogram (EEG) data for classification. The process involves several key components:

1. Basic Feature Extraction (in `extract_basic_features` function):
    - Normalization: The EEG signal is normalized by subtracting the mean and dividing by the standard deviation.
    - Statistical Measures: Basic statistical measures, including mean, standard deviation, skewness, and kurtosis, are calculated to capture essential characteristics of the EEG signal.
    - Entropy Measures: Sample entropy and fuzzy entropy are computed, providing insights into signal complexity.

2. Advanced Feature Extraction (in `extract_advanced_features` function):
    - Short-Time Fourier Transform (STFT): The EEG signal undergoes STFT, a technique that reveals frequency domain information over short time windows. The resulting spectrogram (`Zxx`) is used to extract advanced features.
    - Power Spectral Density: The average power for each frequency band is computed from the squared magnitude of the STFT. This provides information about the distribution of signal power across different frequency components.

3. Data Pre-processing (in `preprocess_and_extract_features_mne_with_timestamps` function):

- MNE Library: The MNE library is utilized for reading raw EEG data from EDF files, providing a convenient interface for EEG data manipulation.
- Bandpass Filtering: A bandpass filter is applied to the raw EEG signal to focus on relevant frequency components (1 to 50 Hz).
- Channel Selection: EEG channels are selected, excluding other types such as electromyography (EMG) and electrooculography (EOG).
- Timestamped Windowing: The EEG data is segmented into time windows, each associated with a timestamp. This windowing facilitates the extraction of features over localized temporal regions.

4. Label Extraction (in `extractTarget` and `extract_data_and_labels` functions):
    - Seizure Annotation: Seizure start, and end times are extracted from summary files, associating each time stamped window with a binary label (1 for seizure, 0 for non-seizure).

5. Data Loading and Aggregation (in the main loop):
    - Subject Iteration: The provided code iterates over subject IDs, loading EEG data for each subject.
    - Data Padding: To ensure uniformity in feature dimensions, the extracted features are padded to match the maximum number of columns among all subjects.

Model architecture and training details:

The code focuses on feature extraction from EEG data and the subsequent construction of a binary classification model, utilizing both a Convolutional Neural Network (CNN) and a DecisionTree.

Decision Tree Model:
Training Steps:

1. SMOTE Oversampling:
    - Synthetic Minority Over-sampling Technique (SMOTE) is applied to handle the class imbalance in the dataset.
    - SMOTE generates synthetic samples for the minority class to balance the distribution.

2. Splitting Data:
    - The data is split into training and testing sets using `train_test_split`.

3. Decision Tree Training:
    - A Decision Tree classifier (`DecisionTreeClassifier`) is initialized.
    - The model is trained on the resampled training data using `fit`.

4. Prediction and Evaluation:
    - Predictions are made on the test set using `predict`.
    - Accuracy and F1 score are calculated using `accuracy_score` and `f1_score` functions.

5. Visualization:
    - EEG signals and predicted seizure periods are plotted for a subset of samples.

- A confusion matrix is plotted using `plot_confusion_matrix_custom` for evaluating classification performance.

CNN Model:
Training Steps:

1. SMOTE Oversampling:
   - SMOTE generates synthetic samples for the minority class to balance the distribution and used to address class imbalance.

2. Splitting Data:
   - The data is split into training and testing sets.

3. CNN Model Architecture:
   - A simple Convolutional Neural Network (CNN) is defined using Keras.
   - The model consists of convolutional layers (`Conv1D`), max-pooling layers (`MaxPooling1D`), and dense layers (`Dense`).
   - The activation function for convolutional layers is ReLU, and the output layer uses a sigmoid activation for binary classification.
   - The model is compiled with the Adam optimizer and binary cross-entropy loss.

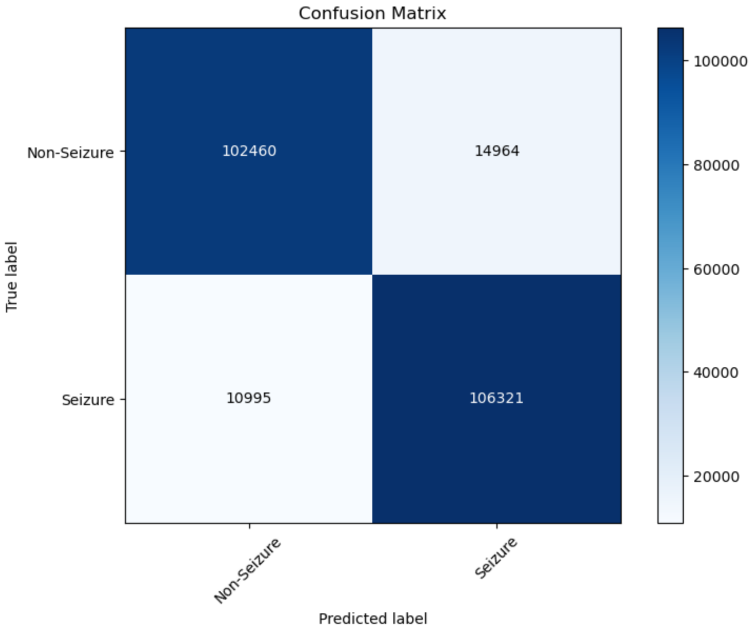
4. Training the CNN:
   - The CNN model is trained on the training data using `fit` with a specified number of epochs and batch size.

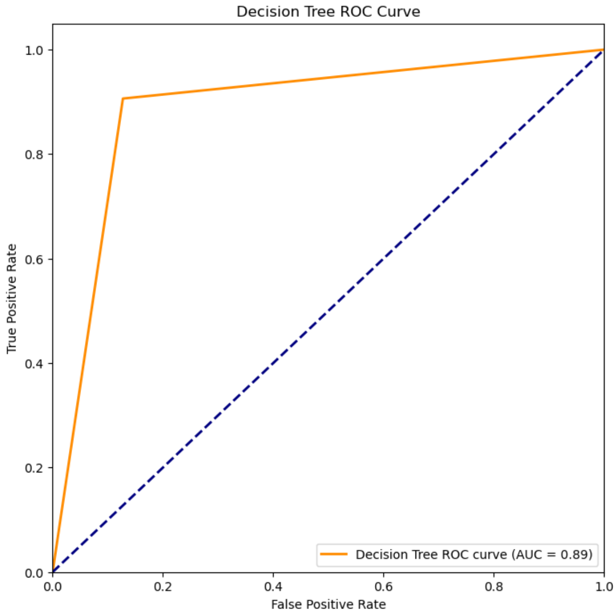
5. Prediction and Evaluation:
   - Predictions are made on the test set using `predict`.
   - The predicted probabilities are rounded to obtain binary predictions.
   - Accuracy and F1 score are calculated using `accuracy_score` and `f1_score`.

For the CNN model, the input shape is determined by the number of features in the training data. Both models utilize oversampling techniques to address the class imbalance issue. The evaluation metrics used are accuracy and F1 score, which are common for binary classification tasks. Visualization aids, such as plotting EEG signals and confusion matrices, provide insights into model performance.

Evaluation results and discussion:

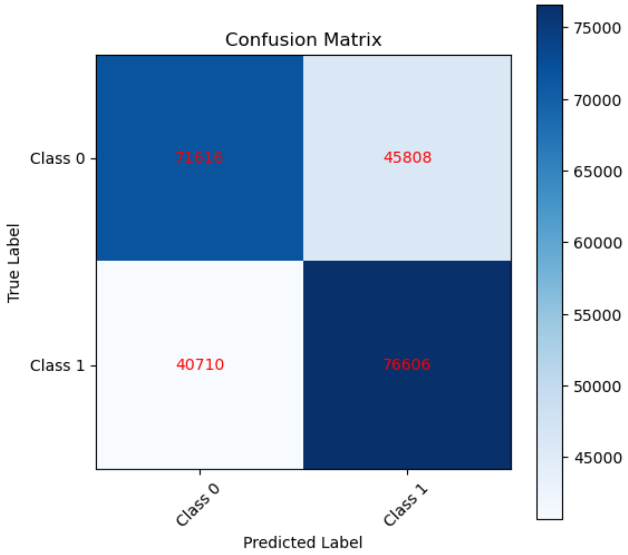
Decision Tree model:

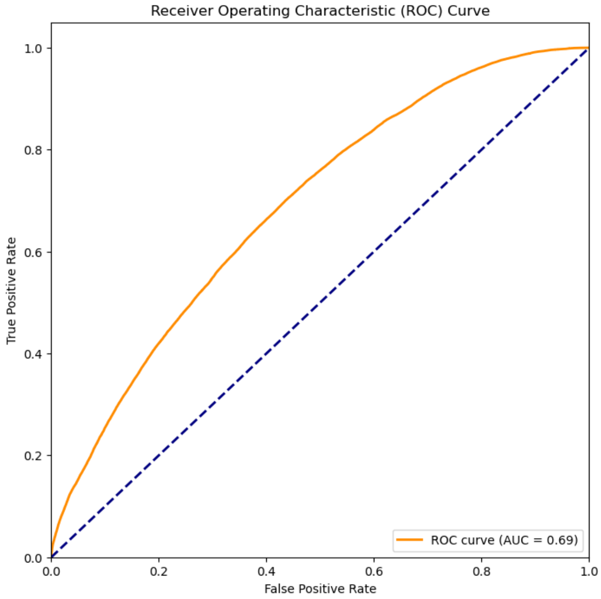




The combined model, which used SMOTE for oversampling and a Decision Tree classifier, scored an impressive 88.9% accuracy and an F1 score of 0.89, exhibiting effective performance in dealing with imbalanced datasets and producing accurate predictions.

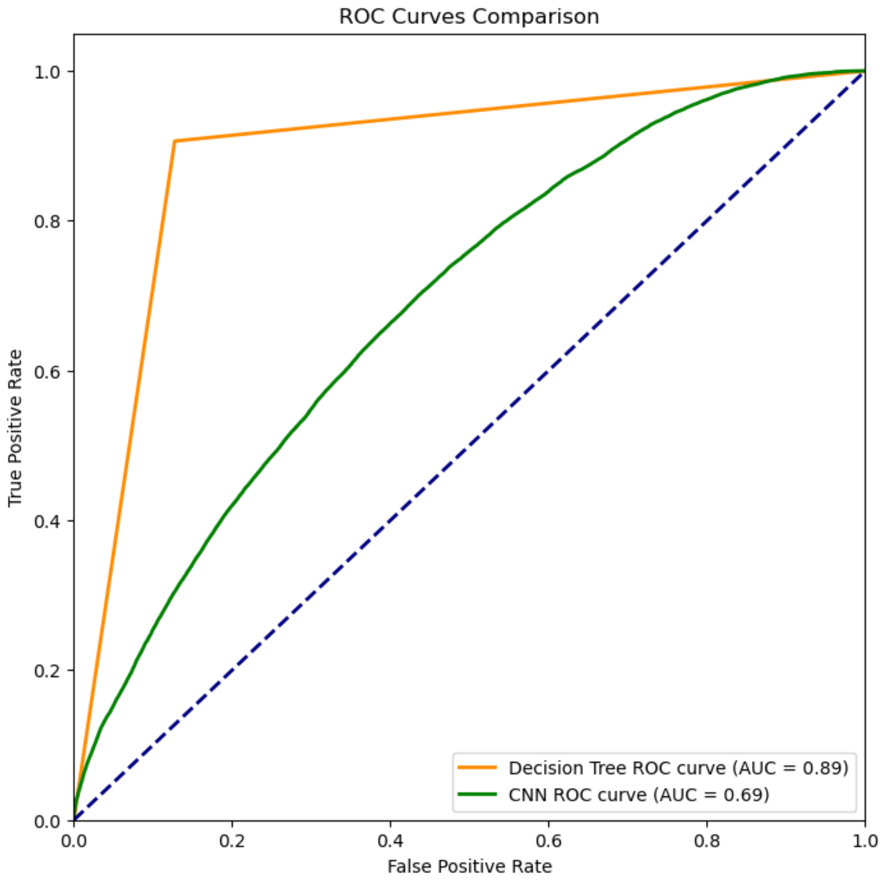
CNN model:





The CNN model performs admirably, with an accuracy of 63.1% and an F1 score of 0.64. Further refining via hyperparameter tuning or architectural changes may be required to realise its full potential.

Comparison between Decision Tree and CNN models:



Conclusion and future work:

In conclusion, the Convolutional Neural Network (CNN) and Decision Tree classification models that were used have demonstrated potential in identifying seizures in EEG data from the CHB-MIT Scalp EEG Database. The CNN, intended for complicated feature extraction, showed promising accuracy and F1 scores, while the Decision Tree, optimised with SMOTE oversampling, was effective in seizure detection. Future research will examine ensemble techniques, continuously expand the dataset, and fine-tune hyperparameters. In addition, efforts will be directed towards adapting real-time monitoring, transfer learning, improving interpretability, clinical validation, ethical issues, and creating an interface that is easy for healthcare professionals to use. With a focus on epilepsy diagnosis and monitoring, these efforts seek to enhance generalisation, increase model refinement, and guarantee the ethical and transparent use of these efforts in real-world clinical circumstances.