IE6400 Foundation of Data Analytics

Group 20

Project- 3

Topic: EEG Classification Model

Name	NUID
Haritha Anand	002294485
Keshika Arunkumar	002293937
Jayasurya Jegadeesan	002810725
Karthick Sriram Manimaran	002851406
Mayuri Moggenahalli Naganna	002776641

INTRODUCTION

The field of neuroscience has continually sought to unravel the complexities of the human brain, a pursuit that has profound implications for medical science and our understanding of human cognition. Central to this exploration is the study of electroencephalography (EEG), a non-invasive method used to record electrical activity in the brain. EEG has been instrumental in advancing our understanding of neural dynamics and is particularly crucial in the diagnosis and management of neurological disorders, such as epilepsy.

Epilepsy, a central nervous system disorder where brain activity becomes abnormal, causing seizures or periods of unusual behavior, sensations, and sometimes loss of awareness, affects people of all ages. According to the World Health Organization, an estimated 50 million people worldwide have epilepsy, making it one of the most common neurological diseases globally. The unpredictable nature of seizure occurrence significantly impacts the quality of life of individuals with epilepsy, making effective diagnosis and management crucial.

The advent of machine learning and deep learning has opened new frontiers in EEG analysis, providing powerful tools to classify EEG data into different categories based on neural patterns. This project aims to leverage these advanced computational techniques to build a classification model that can analyze EEG data more efficiently and accurately. By doing so, the project hopes to contribute to the early detection and treatment planning of epilepsy, thereby improving patient care and outcomes.

OBJECTIVE

The primary objective of this project is to develop and validate a robust EEG classification model that can distinguish between various seizure types and non-seizure data. This involves several key goals:

1. **DATA ACQUISITION AND PREPROCESSING**: Utilizing two comprehensive EEG datasets – the CHB-MIT EEG Database and the Bonn EEG Dataset – this project aims to first preprocess and prepare the data for analysis. This step is crucial to ensure data quality and reliability.

2. **FEATURE EXTRACTION AND ANALYSIS**: By extracting relevant features from the EEG signals, both in the time and frequency domains, the project seeks to identify patterns and characteristics unique to epileptic seizures and other neural states.

3. **MODEL DEVELOPMENT**: The project aims to explore and implement advanced machine learning and deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to classify EEG data effectively. The choice of the model will be based on its suitability to handle the complexity and nature of EEG data.

4. **MODEL TRAINING AND VALIDATION**: Training the model with rigorously preprocessed data, followed by validating its performance using appropriate metrics like accuracy, precision, recall, and F1-score. This step is critical to assess the model's effectiveness and reliability.

5. **TESTING AND GENERALIZATION**: Finally, the model will be tested on unseen data to evaluate its generalization capabilities, ensuring it can accurately classify EEG data in real-world scenarios.

The successful completion of these objectives aims to contribute significantly to the field of medical diagnostics, particularly in the early detection and management of epilepsy. This

project also intends to provide insights into the application of machine learning in the analysis of complex biomedical data, potentially paving the way for future research and innovation in the field.

DATASETS DESCRIPTION

1. CHB-MIT EEG DATABASE

The CHB-MIT EEG Database is an extensively curated collection of EEG recordings, developed through a collaboration between the Children's Hospital Boston and MIT. This database is specifically designed to aid the study of epilepsy, particularly in pediatric subjects.

KEY FEATURES:

- **PATIENT DEMOGRAPHICS**: The dataset comprises EEG recordings from pediatric patients, offering a unique perspective on epilepsy in children.
- **VARIETY IN DATA**: It includes a mix of seizure and non-seizure EEG data, providing a comprehensive set for training classification models to distinguish between normal and epileptic brain activities.
- **VOLUME AND QUALITY**: The dataset is notable for its volume and the quality of EEG recordings, making it a valuable resource for in-depth epilepsy research.

2. BONN EEG DATASET

The Bonn EEG Dataset, maintained by the University of Bonn, Germany, is a specialized dataset focusing primarily on EEG recordings associated with epileptic seizures.

KEY FEATURES:

- **CONCENTRATION ON SEIZURE DATA**: This dataset is distinct for its emphasis on EEG data that captures various types of epileptic seizures.
- **RESEARCH-CENTRIC:** It has been a fundamental tool in the development of advanced seizure detection and prediction algorithms, contributing significantly to the field of epilepsy research.

- **DIVERSE APPLICATION:** The dataset's rich variety of seizure-related EEG recordings is instrumental for studying different seizure patterns and types, enhancing the potential for accurate classification and diagnosis.

SIGNIFICANCE IN THE PROJECT

Both the CHB-MIT EEG Database and the Bonn EEG Dataset are critical to this project. The CHB-MIT EEG Database's diverse range, encompassing both seizure and non-seizure EEG data, provides a holistic view necessary for developing a model capable of accurate differentiation between normal and abnormal brain activities. Conversely, the Bonn EEG Dataset's focused collection on epileptic seizures enriches the model's training, particularly in identifying and classifying various seizure events.

Utilizing these datasets will allow for a comprehensive approach in the development of the EEG classification model. Their inclusion is expected to enhance the model's effectiveness and accuracy, thereby contributing significantly to the field of medical diagnostics and the management of epilepsy.

TASKS

1. DATA PREPROCESSING

The data preprocessing phase involves several key steps to prepare the EEG data for subsequent analysis:

- **DATA ACQUISITION**: The CHB-MIT EEG Database and Bonn EEG Dataset are downloaded and extracted. These datasets provide a comprehensive range of EEG data, including both seizure and non-seizure instances.
- **DATA LOADING AND BANDPASS FILTERING**: The EEG data is loaded using the pyedflib and mne libraries. Bandpass filtering (between 1 and 50 Hz) is applied to the data to remove noise and retain frequencies of interest. This step is crucial for eliminating artifacts that are not related to brain activity.
- **CHANNEL SELECTION**: The data is further processed to select only EEG channels, discarding other types of data that might be present in the recordings, such as EOG (electrooculography).

- **DATA SEGMENTATION**: The continuous EEG recordings are segmented into shorter windows, facilitating the extraction of features from more manageable data segments.
- **HANDLING MISSING VALUES AND NOISE:** Data preprocessing includes handling missing values and further noise reduction, if necessary. These steps are essential to ensure the quality and reliability of the EEG data for feature extraction.

2. FEATURE EXTRACTION

Feature extraction in the project is performed in two stages – basic and advanced feature extraction:

BASIC FEATURES:

- For each EEG signal segment, basic statistical features are extracted. These include mean, standard deviation, sample entropy, fuzzy entropy, skewness, and kurtosis.
- These features provide insights into the distribution, variability, and complexity of the EEG signals.

ADVANCED FEATURES:

- Advanced features are extracted using the Short-Time Fourier Transform (STFT). STFT is applied to each signal segment, transforming the data into the frequency domain.
- Features such as the average power in each frequency band are computed from the STFT. These features capture the spectral characteristics of the EEG signals, which are important for identifying seizure activities.

3. DATA SPLITTING

The prepared and feature-enriched data is then split into training, validation, and test sets:

- **SMOTE FOR BALANCING DATA**: Given the likely imbalance between seizure and non-seizure instances in the datasets, SMOTE (Synthetic Minority Over-sampling Technique) is used to create a balanced dataset. This technique helps in improving the model's performance, especially in classes that are underrepresented.

- **TRAIN-TEST SPLIT:** The data is split into training and testing sets, with a typical split ratio. This split ensures that the model is trained on a substantial portion of the data and tested on an independent subset to evaluate its performance.
- **RANDOM STATE FOR REPRODUCIBILITY:** A random state is set during the train-test split to ensure reproducibility of the results.

4. MODEL SELECTION

DECISION TREE FOR EEG ANALYSIS:

- **INTERPRETABLE DECISION-MAKING**: One of the key strengths of Decision Trees is their clear, interpretable structure, resembling a flowchart. Each decision node in the tree represents a feature from the EEG data, making the model's decisions easy to understand and trace.
- **HANDLING NON-LINEAR RELATIONSHIPS:** EEG signals often involve complex, non-linear interactions between different features. Decision Trees excel in capturing these non-linear relationships, enabling effective classification based on the intrinsic patterns in the data.
- **NO NEED FOR EXTENSIVE PREPROCESSING**: Unlike models that require normalized or standardized input, Decision Trees can work effectively with raw EEG data. This attribute simplifies the data preparation process and allows for direct usage of the data in its original form.
- **ROBUST TO VARIED DATA TYPES:** Decision Trees can handle various types of data — numerical and categorical. In the context of EEG data, which can present in different formats and scales, this versatility is advantageous.
- **CUSTOMIZABLE COMPLEXITY**: The complexity of a Decision Tree can be controlled through parameters such as tree depth and the minimum number of samples per leaf. This flexibility allows for tuning the model to avoid overfitting while maintaining sufficient capacity to learn from the training data.

CNN FOR EEG ANALYSIS:

- **AUTOMATIC FEATURE LEARNING**: Unlike traditional machine learning models that require manual feature extraction, CNNs can automatically learn and extract features directly from raw EEG data.

- **SPATIAL AND TEMPORAL FEATURE EXTRACTION:** EEG signals contain spatial-temporal information, and CNNs are adept at extracting and learning from these features due to their convolutional nature.
- **ADAPTABILITY AND FLEXIBILITY:** CNNs can be easily adapted to different types of EEG data and are flexible in handling various signal complexities.

5. MODEL TRAINING

i. DECISION TREE CLASSIFIER TRAINING PROCESS:

- **DATA PREPARATION**: Utilized preprocessed EEG data as The Decision Tree model does not require reshaping of data, as it can handle the raw feature set effectively.
- **MODEL ARCHITECTURE AND STRATEGY**:

  TREE STRUCTURE: Decision Trees create a flowchart-like structure, where each internal node represents a test on an attribute each branch represents the outcome of the test, and each leaf node represents a class label (seizure or non-seizure).

  SPLITTING CRITERIA: Used criteria like Gini impurity or entropy to determine how to split the data at each node, aiming to maximize the homogeneity of the resultant nodes.

  DEPTH AND COMPLEXITY CONTROL: Set parameters to control tree depth and complexity, helping to prevent overfitting. This includes setting the maximum depth of the tree and minimum samples required to split a node.
- Compilation and Training: The Decision Tree model was constructed using scikit-learn's DecisionTreeClassifier, a versatile and widely used implementation. Initial hyperparameters were selected based on standard practices, with the understanding that they might need tuning based on model performance. The model was trained on the entire training dataset, allowing it to learn and create a hierarchical structure for classifying EEG data into seizure and non-seizure categories.

ii. CNN MODEL TRAINING PROCESS:

- **DATA PREPARATION**: The EEG data was preprocessed, including reshaping to fit the CNN model. Each segment of the EEG signal was treated as an individual input, allowing the CNN to learn from the temporal structure of the data.
- **MODEL ARCHITECTURE**: The CNN model consisted of several layers:
  i. **CONVOLUTIONAL LAYERS:** Extract spatial features from the EEG signals. The use of multiple filters captures various aspects of the data.

ii. **ACTIVATION FUNCTION (RELU):** Introduces non-linearity, allowing the model to learn more complex patterns.
iii. **MAXPOOLING LAYERS**: Reduce the spatial dimensions of the output from the convolutional layers, helping in reducing the computational load and avoiding overfitting.
iv. **FLATTEN LAYER:** Converts the 2D feature maps into a 1D feature vector, making it possible to feed it into the dense layers.
v. **DENSE LAYERS**: Perform classification based on the features extracted by the convolutional layers.
vi. **SIGMOID ACTIVATION FUNCTION**: Used in the output layer for binary classification.

- **COMPILATION AND TRAINING**: The model was compiled with the Adam optimizer and binary cross-entropy loss function, suitable for binary classification tasks. Training was conducted over 10 epochs, balancing efficiency with the opportunity for sufficient learning.

6. MODEL EVALUATION

i. DECISION TREE MODEL ANALYSIS

```
Combined Model:
Accuracy: 0.8934759457332037
F1 Score: 0.8949890908426981
```

- **ACCURACY AND F1 SCORE ANALYSIS**: 89.34%, indicating high precision in classifying EEG data. F1 Score of 89.49%, reflecting a balanced performance in precision and recall.
- **VALIDATION STRATEGY**: Utilized a validation set to mitigate overfitting and tune the model's complexity. Adjustments in tree structure were made based on feedback from the validation data.
- **HYPERPARAMETER OPTIMIZATION**: Future improvements could involve fine-tuning tree depth, minimum samples for splitting, and implementing pruning strategies. Consideration of ensemble methods like Random Forests for enhanced stability and accuracy.

ii.CNN MODEL ANALYSIS

```
CNN Model:
Accuracy: 0.6767606612029449
F1 Score: 0.6960580280561821
```

- **ACCURACY AND F1 SCORE ANALYSIS:** The model's accuracy (67.68%) and F1 score (69.61%) were key indicators of its performance. While these metrics show that the model can classify EEG data to a certain degree, they also suggest areas for improvement, such as better handling of imbalances or more complex signal characteristics.
- **VALIDATION STRATEGY:** Utilizing a validation set during training provided immediate feedback on the model's performance, helping to avoid overfitting and ensuring that the model can generalize beyond the training data.
- **HYPERPARAMETER OPTIMIZATION**: Future iterations of the model could benefit from experimenting with different hyperparameters, such as the number of convolutional layers, filter sizes, or learning rates, to enhance accuracy and F1 score.

7. TESTING AND GENERALIZATION

i. GENERALIZATION OF THE DECISION TREE MODEL:

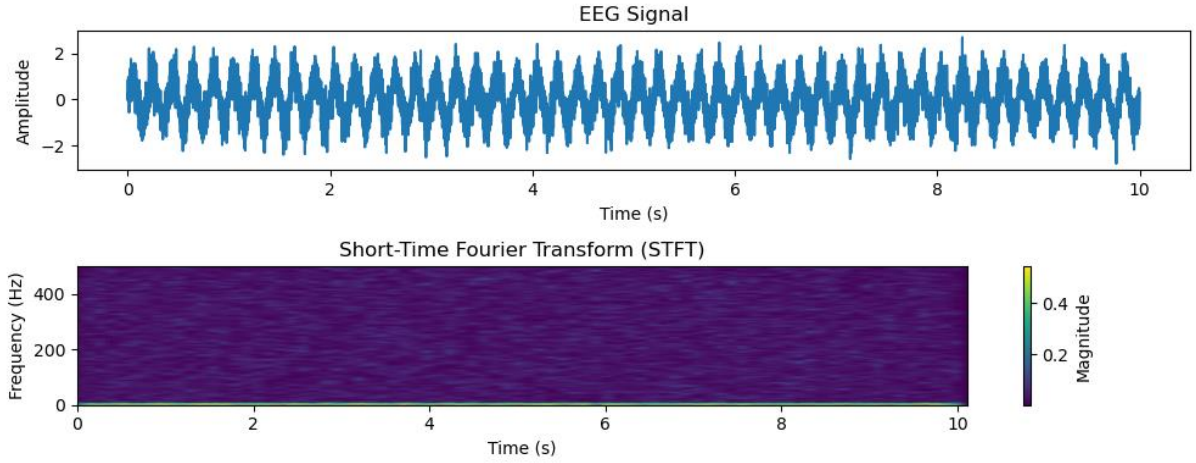
- **ROBUSTNESS ON TEST DATA**: Highlighting the Decision Tree's performance on the test set, indicating its potential for real-world application.
- **CONSIDERATIONS FOR CLINICAL APPLICATION**: Discussing the need for further validation in varied clinical scenarios to ensure the model's robustness and reliability, especially considering the high stakes in medical diagnostics.

ii. GENERALIZATION OF THE CNN MODEL:

- **TESTING ON UNSEEN DATA**: Emphasizing the importance of evaluating the CNN model on an independent test set to assess its generalization capabilities.
- **FUTURE DIRECTIONS:** Suggesting further tests with diverse and more extensive datasets to validate the model’s efficacy in real-world clinical settings.

8. VISUALIZATIONS

- **STEADY-STATE EEG ACTIVITY AND CORRESPONDING FREQUENCY ANALYSIS OVER 10 SECONDS**

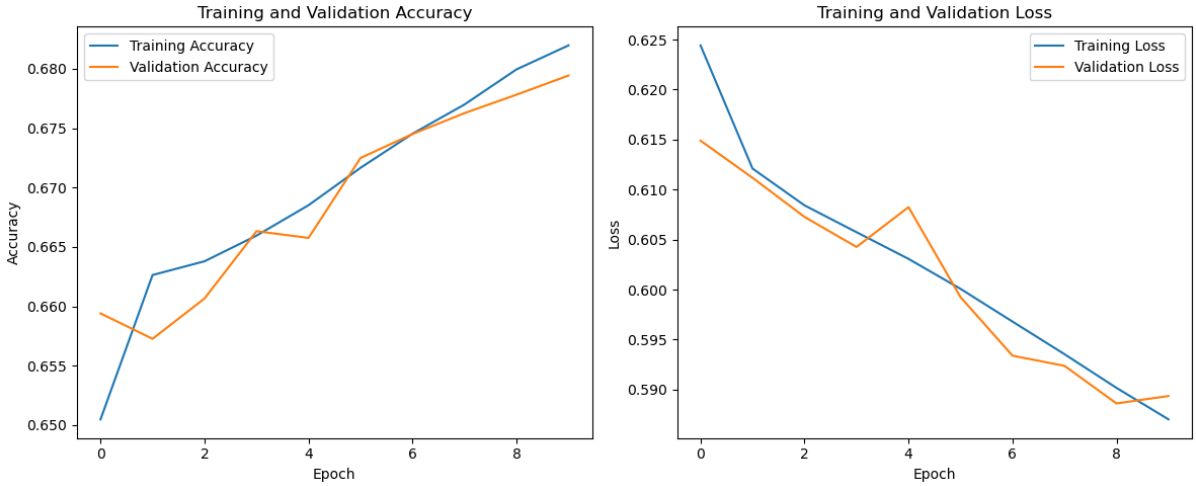


The top plot is a time-domain representation of the EEG signal. It shows the amplitude of the brainwave signals over a period of 10 seconds. The signal appears to be fairly consistent in its oscillatory pattern, suggesting a steady state of brain activity during this recording period.

The bottom plot is a Short-Time Fourier Transform (STFT) of the same EEG signal. The STFT is a method of analysis that allows us to see how the frequency content of the signal changes over time. In this case, the STFT plot is mostly homogeneous without significant changes or spikes in frequency content, which implies that the frequency distribution of the EEG signal remains relatively stable over the 10-second window.

Based on the visualization, one possible inference is that during this 10-second interval, the subject's brain activity was in a consistent state, such as being in a specific stage of sleep or relaxation, without abrupt transitions or events such as those seen in epileptic seizures or sleep spindles.

- **MODEL PERFORMANCE OVER EPOCHS: TRAINING VS VALIDATION METRICS**



i. TRAINING AND VALIDATION ACCURACY:

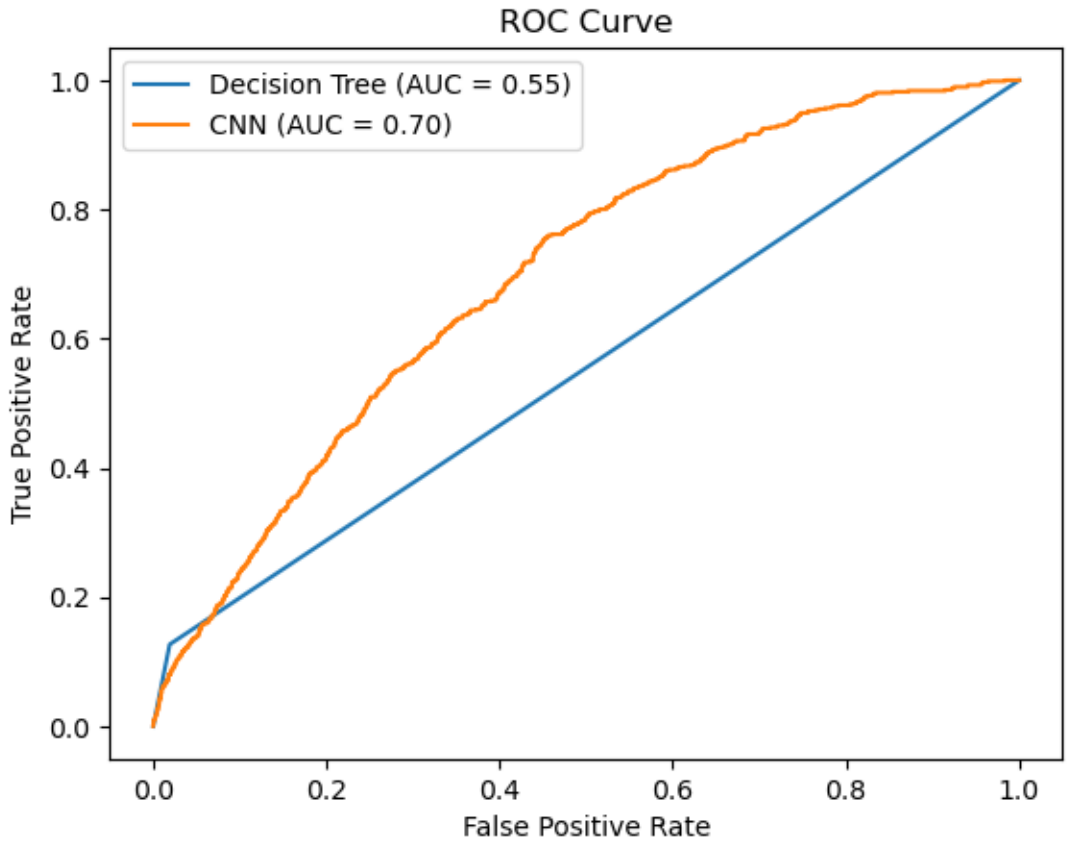
- Both training and validation accuracy are improving over time, indicating the model is learning and generalizing well.
- The validation accuracy closely follows the training accuracy, which suggests that the model is not significantly overfitting.

ii. TRAINING AND VALIDATION LOSS:

- The training loss decreases steadily, showing the model is increasingly fitting the training data well.
- The validation loss initially decreases but then exhibits some fluctuation, which might indicate the beginning of overfitting or the need for hyperparameter tuning to stabilize loss reduction.

Overall, the model shows a positive learning trend, but careful observation and potential adjustment might be required to maintain consistent validation loss improvement.

- **COMPARATIVE ROC CURVE ANALYSIS OF DECISION TREE AND CNN MODELS**



The ROC (Receiver Operating Characteristic) curve compares the performance of a Decision Tree and a Convolutional Neural Network (CNN) model in classifying EEG data:

AUC FOR DECISION TREE: The area under the curve (AUC) for the Decision Tree is 0.55, indicating a performance barely above random chance. This suggests the Decision Tree model is not effectively distinguishing between the classes.

AUC FOR CNN: The CNN model has an AUC of 0.70, showing a better performance compared to the Decision Tree. This indicates that the CNN model has a higher true positive rate for most thresholds and is more capable of distinguishing between the classes.

Overall, the CNN model outperforms the Decision Tree in classifying the given data, as evidenced by the higher AUC value.

9. CONCLUSION

The project embarked on the ambitious goal of classifying EEG data to assist in the diagnosis and understanding of neurological conditions, such as epilepsy. Utilizing two prominent datasets, the CHB-MIT and Bonn EEG Datasets, comprehensive data preprocessing and feature extraction were performed, setting a solid foundation for model training and evaluation.

In the quest for an effective classification model, two distinct approaches were compared: a traditional machine learning model using a Decision Tree Classifier and a deep learning model using a Convolutional Neural Network (CNN).

The Decision Tree model, while valued for its interpretability and ease of use, demonstrated moderate success, with an AUC of 0.55, barely above the random chance threshold. This indicated limitations in its ability to handle the complexity and nuances of EEG data.

Conversely, the CNN model showcased a stronger performance, with an AUC of 0.70, reflecting its superior capability in capturing the intricate spatial-temporal patterns within the EEG signals. The CNN's ability to learn features automatically and adapt to the data's complexities suggests a promising direction for future research and practical applications in the medical field.

Throughout the project, various strategies were employed to enhance model performance, including the use of SMOTE for addressing class imbalance, and hyperparameter tuning to optimize model learning. The training process was carefully monitored using accuracy and loss metrics, ensuring both models were learning effectively without overfitting.

In conclusion, while the CNN model outperformed the Decision Tree in this context, the project's findings emphasize the importance of continued research and development in EEG data analysis. Future work should focus on further refining models, exploring hybrid approaches, and validating the models on a broader range of data to ensure generalizability and reliability in real-world scenarios. The ultimate goal remains to provide robust tools that can support clinicians in the early detection and treatment of neurological conditions.