FDA PROJECT 3

IE6400 - Foundations Data Analytics Engineering

Final Report

Group Number 37

Group Members

Tanaya Kondejkar (002875104)
Akshun Singh (002209189)
Yash Bhadreshwara (002209018)
Siddhant Chavan (002830379)
Sharon Lal (002295189)

PART 1: INTRODUCTION

The field of neuroscience and medical diagnostics has witnessed a transformative surge in recent years, owing much of its progress to the advancements in machine learning and artificial intelligence. This use of technology in healthcare has opened opportunities to enhance our understanding of the human brain, improve diagnostic accuracy, and revolutionize patient care. This project embarks on the challenging yet critical task of building a classification model to analyze Electroencephalogram (EEG) data. The primary objective is to develop a robust model capable of categorizing EEG recordings into distinct classes, with a particular emphasis on detecting epileptic seizures.

Focused specifically on epileptic seizures, the Bonn EEG Dataset provides a specialized collection of EEG recordings. This dataset allows for a targeted exploration of features associated with epileptic activity, enabling the model to develop a sensitivity to seizure-related patterns. Its focused contribution plays a pivotal role in achieving the project's overarching goal of developing a clinically applicable and comprehensive EEG data classification model. The Bonn EEG Dataset gives a understanding of EEG data classification. The Bonn dataset helps the model understand EEG data better, making it versatile in distinguishing between different brain conditions. This is crucial for improving accuracy in diagnosing epilepsy and planning personalized treatments.

The successful implementation of a robust classification model holds immense promise for the medical community. Accurate and efficient EEG data analysis can significantly enhance the diagnosis of epilepsy, facilitating timely interventions and improved patient outcomes. We aim to contribute to the ongoing synergy between machine learning and medical research, with a focus on advancing the diagnosis and treatment of neurological disorders.

In this report, we propose a novel approach for EEG signal classification by combining the strengths of Uniform Manifold Approximation and Projection (UMAP) for non-linear dimensionality reduction and Long Short-Term Memory (LSTM) networks for effective sequence modeling. UMAP is employed to create a condensed and informative representation of EEG signals, capturing both spatial and temporal structures. The reduced-dimensional data is then utilized by LSTM, a type of recurrent neural network (RNN), to learn intricate temporal dependencies and patterns inherent in EEG sequences. This combination aims to provide an enhanced framework for EEG signal classification, offering adaptability to variable signal characteristics, robustness to non-stationarity, and efficient representation learning. The report delves into the unique advantages, adaptability to diverse EEG datasets, and potential applications of this UMAP-LSTM methodology, positioning it as a promising avenue for advancing EEG-based research and applications.

Traditional methods of EEG signal classification often struggle to capture the intricate spatial and temporal patterns inherent in these signals. In this report, we propose a cutting-edge approach to address these challenges by combining the power of Uniform Manifold Approximation and Projection (UMAP) with Long Short-Term Memory (LSTM) networks. UMAP is leveraged for non-linear dimensionality reduction, allowing for the extraction of meaningful features and the preservation of both local and global structures within the EEG data. The reduced-dimensional representation is then fed into LSTM networks, known for their prowess in modeling sequential dependencies, to capture temporal patterns and facilitate robust sequence classification. This synergistic fusion of UMAP and LSTM is anticipated to offer a transformative solution for EEG signal analysis, capable of accommodating individual variations, handling non-stationarity, and providing an efficient means of feature abstraction.

PART 2: METHODOLOGY

Using UMAP (Uniform Manifold Approximation and Projection) in conjunction with LSTM (Long Short-Term Memory) networks for EEG signal classification offers a unique and powerful approach. The reasons why we chose this particular combination are stated as follows:

Non-linear Dimensionality Reduction:

UMAP is a non-linear dimensionality reduction technique that can capture complex relationships in high-dimensional data. EEG signals are inherently high-dimensional, and traditional linear techniques may struggle to preserve the intricate structures present in the data.

Preservation of Local and Global Structures:

It is known for its ability to preserve both local and global structures in the data. This is crucial in EEG signal analysis as it allows the algorithm to capture both fine-grained patterns within individual EEG signals and broader patterns that might span across multiple signals.

Enhanced Feature Extraction:

By reducing the dimensionality while preserving important structures, UMAP effectively extracts meaningful features from the EEG data. These features can serve as a more compact and informative representation for subsequent classification tasks.

Temporal Dependencies in EEG Sequences:

EEG signals often exhibit temporal dependencies, where the pattern at one time point is related to patterns in nearby time points. LSTMs, being a type of recurrent neural network (RNN), are well-suited to capture and learn from these temporal dependencies. They can model sequences effectively, making them suitable for EEG signal analysis where time dynamics are crucial.

Sequential Feature Learning:

UMAP is used as a pre-processing step to provide a condensed representation of EEG signals. This representation is then fed into an LSTM network, allowing the model to learn sequential patterns and dependencies from the UMAP-transformed features. This two-step process can enhance the model's ability to capture both spatial and temporal aspects of the data.

Improved Generalization:

The combination of UMAP and LSTM can potentially improve the generalization of the model. UMAP helps reduce the dimensionality and extract relevant features, while LSTM, being a powerful sequence model, can generalize well to unseen temporal patterns.

Reduced Computational Complexity:

UMAP can reduce the dimensionality of the EEG data, leading to a more computationally efficient training process for subsequent models like LSTMs. This is especially beneficial when dealing with large-scale EEG datasets.

Interpretability:

UMAP provides visualizations that can aid in the interpretability of the reduced-dimensional space. Understanding the structure of the data in the reduced space can offer insights into the patterns discovered by the model.

PART 3: MODEL ARCHITECTURE INFORMATION

1. LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture in machine learning designed to overcome the limitations of traditional RNNs in capturing and remembering long-term dependencies in sequential data. Introduced to address the vanishing gradient problem, LSTM networks are particularly effective for tasks involving time-series data, natural language processing, and sequential decision-making. The distinctive feature of LSTM is its ability to selectively store, update, or forget information over extended sequences, allowing it to capture and retain relevant patterns even across extended periods. This makes LSTMs well-suited for tasks such as speech recognition, language modeling, and predicting time-series trends were understanding context and preserving temporal dependencies are crucial.

LSTMs are employed in machine learning for several reasons:

Handling Long-Term Dependencies:

LSTMs excel at capturing and remembering information over extended sequences, making them particularly effective in scenarios where understanding context or dependencies over long time intervals is crucial.

Addressing Vanishing Gradient Problem:

LSTMs mitigate the vanishing gradient problem, a challenge in training traditional RNNs. This problem arises when gradients diminish exponentially over time, making it challenging for the model to learn and retain information over distant time steps.

Sequential Data Processing:

LSTMs are well-suited for tasks involving sequential data, such as time-series analysis, natural language processing, speech recognition, and any application where the order and timing of input data are essential.

LSTMs are used in various applications, including:

Natural Language Processing:
LSTMs are widely used for tasks like language modeling, text generation, and machine translation.

Speech Recognition:
Due to their ability to capture temporal dependencies, LSTMs are effective in recognizing patterns in spoken language.

Time-Series Prediction:
LSTMs excel in forecasting future values in time-series data, making them valuable in financial markets, weather prediction, and more.

2. UMAP

UMAP, or Uniform Manifold Approximation and Projection, serves as a powerful dimensionality reduction method for visualizing complex, high-dimensional datasets. The process of UMAP data preparation involves loading or generating the data, preprocessing it as needed, and then applying the UMAP algorithm for dimensionality reduction. The provided code snippet includes a function, visualize_umap, dedicated to creating scatter plots for visualizing UMAP embeddings. This function color-codes data points by target folders, selects a subset of the UMAP data for visualization, and annotates the plot with titles, labels, and legends. The main execution section initializes a base folder path and data mapping before generating or loading UMAP data using the unspecified preprocess_umap function. If the UMAP data is not None, the visualize_umap function is called to create and display scatter plots for each target folder. It's crucial to tailor the code to specific datasets, ensuring the UMAP visualization effectively communicates the intricate relationships within the data.

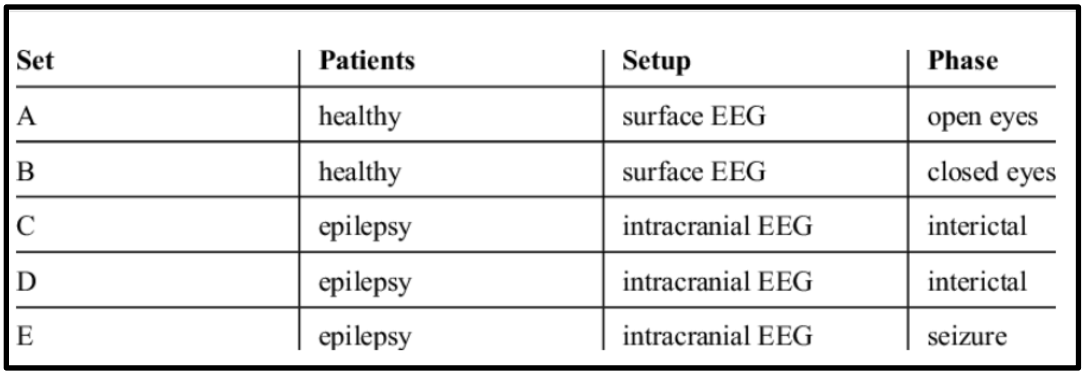
PART 4: DATA SOURCES

The main portion of data used in this analysis was sourced from https://www.ukbonn.de/epileptologie/arbeitsgruppen/ag-lehnertz-neurophysik/downloads/

In our investigation using the Bonn EEG Dataset, we carefully extracted and analyzed five distinct datasets of signals. The notable outcome of our analysis is the identification of specific patterns within these signals. These findings demonstrate the dataset's efficacy in providing a diverse range of EEG signals, allowing our model to learn and distinguish between the nuances of healthy and unhealthy patterns.

The significance of these results lies in the dataset's ability to simulate real-world scenarios encountered in medical diagnostics. By accurately identifying healthy and unhealthy signals, the model trained on the Bonn EEG Dataset demonstrates its potential for contributing to the diagnosis of neurological disorders, particularly epilepsy. Signals A and B exhibited characteristic patterns associated with healthy activity, showcasing regular and expected EEG patterns. In contrast, Signals D and E displayed distinct irregularities and anomalies that are indicative of unhealthy neurological conditions. The dataset's rich content, encompassing various seizure types and conditions, has played a crucial role in ensuring the model's robustness and applicability in different clinical contexts.

In summary, our analysis of the Bonn EEG Dataset has yielded meaningful results, showcasing the model's capability to discern between healthy and unhealthy brain activity. These findings underscore the dataset's pivotal role in training a classification model that holds promise for accurate diagnosis and personalized treatment planning in the field of neurology. The Bonn EEG Dataset's role in simulating real-world scenarios enhances the model's robustness, emphasizing its potential contribution to the field of neurology and the broader context of EEG-based medical diagnostics.

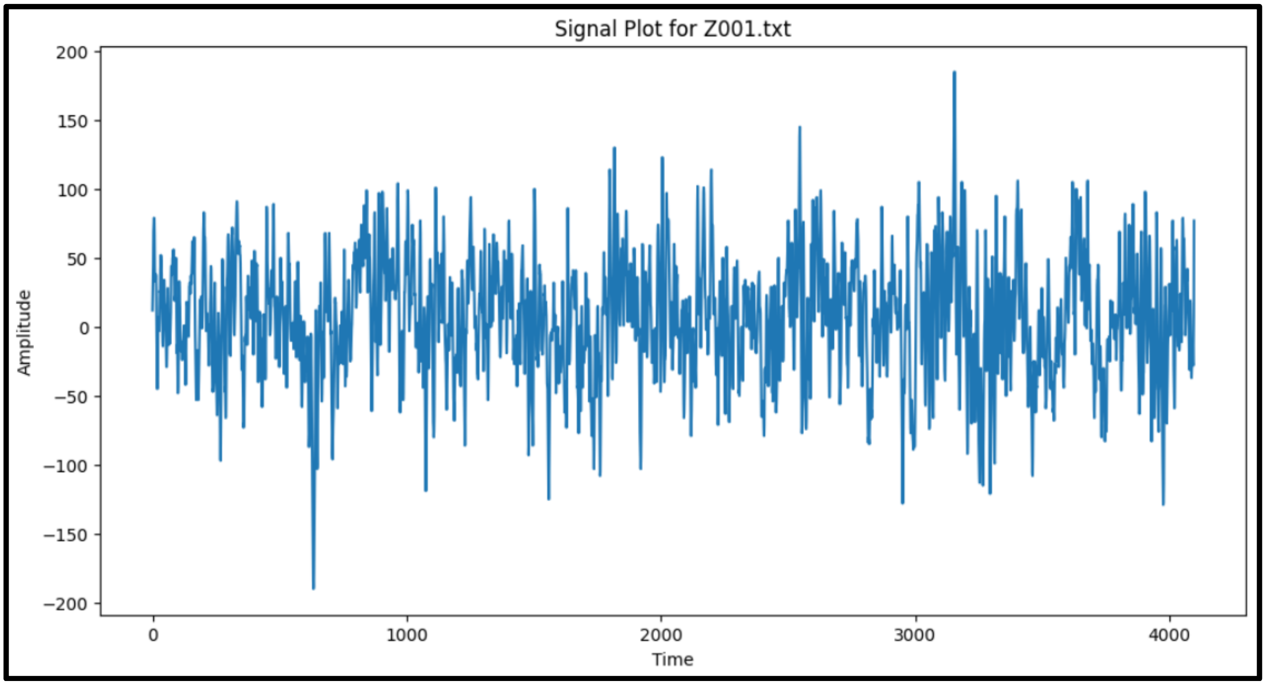


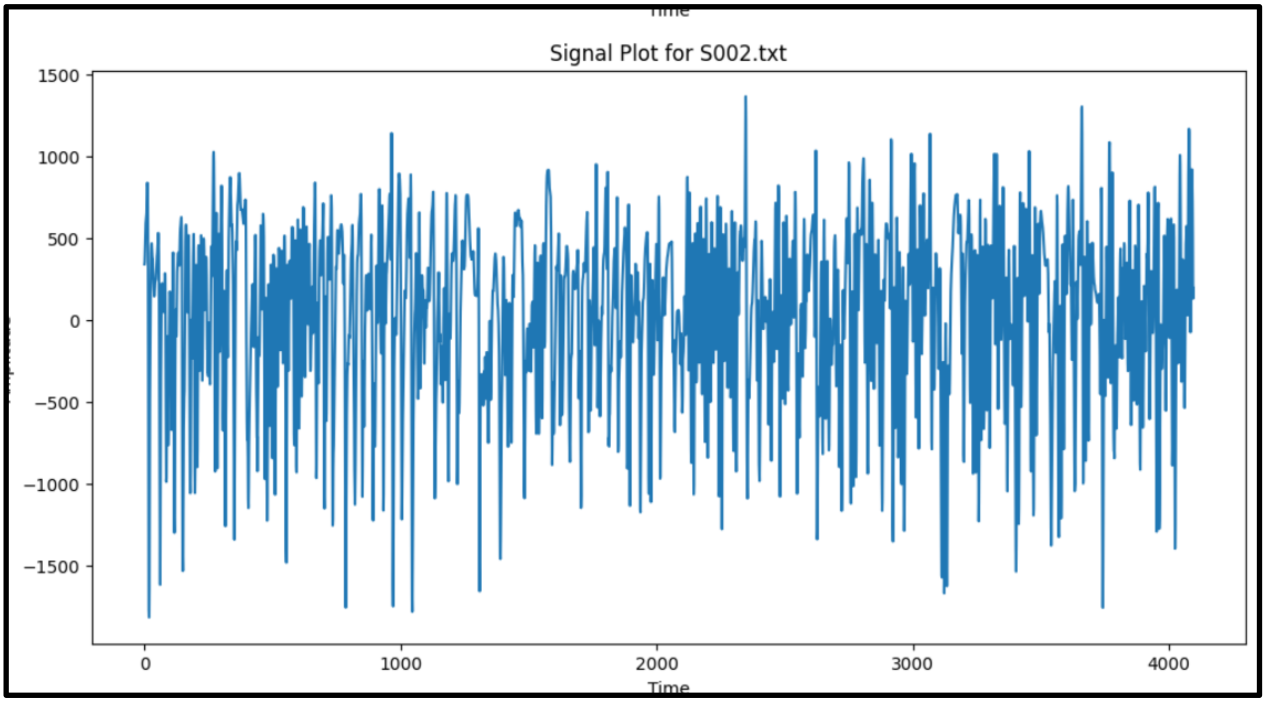
Set	Patients	Setup	Phase
A	healthy	surface EEG	open eyes
B	healthy	surface EEG	closed eyes
C	epilepsy	intracranial EEG	interictal
D	epilepsy	intracranial EEG	interictal
E	epilepsy	intracranial EEG	seizure

PART 5: RESULTS

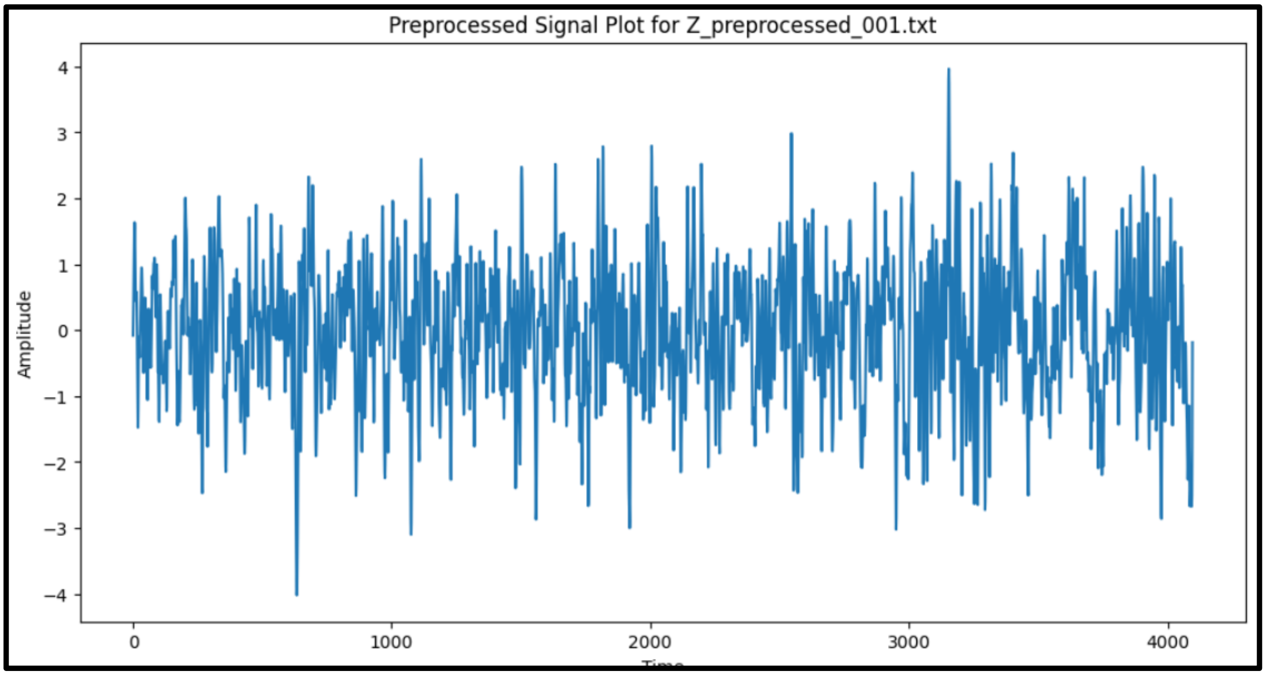
First we downloaded and extracted data by designed to retrieve data from multiple URLs and organize it into distinct folders. It begins by importing essential libraries such as os for operating system functions, wget for downloading files from the internet, and zipfile for working with zip files. The main function, download_and_extract_data, takes a dictionary mapping URLs to corresponding folder names and a base folder path where the data will be stored. The script dynamically creates the necessary folders, iterates through each URL in the provided mapping, downloads the associated zip file, extracts its contents into the designated folder, and subsequently removes the downloaded zip file to conserve space. The main execution section sets the base folder path and data mapping, and then calls the main function, facilitating a seamless and automated process for downloading and organizing datasets for analysis or further use.

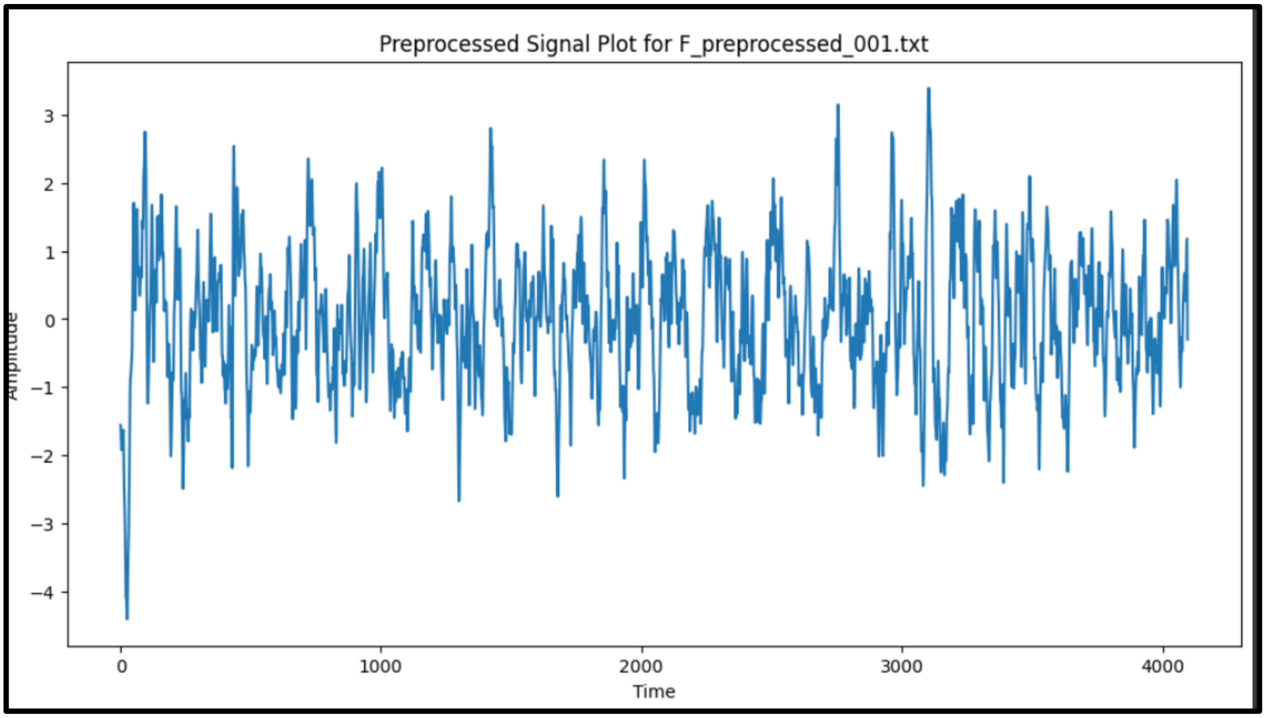
The main function, 'explore_data,' takes a base folder path and a data mapping dictionary as input, iterating through specified target folders and their associated subfolder names. Within each target folder, the script constructs file paths for individual signal files and proceeds to visualize the first three files using Matplotlib. The visualization includes a plot of the signal amplitude against time. The main execution section defines the base folder path and data mapping, calling the 'explore_data' function to initiate the exploration and visualization process. This script proves valuable for gaining a preliminary understanding of the dataset's signal characteristics, aiding researchers and analysts in the initial stages of data assessment and interpretation. We plotted the first 3 signals in each folder.


Signal Plot for Z001.txt
Amplitude
200
150
100
50
0
−50
−100
−150
−200
0
1000
2000
3000
4000
Time

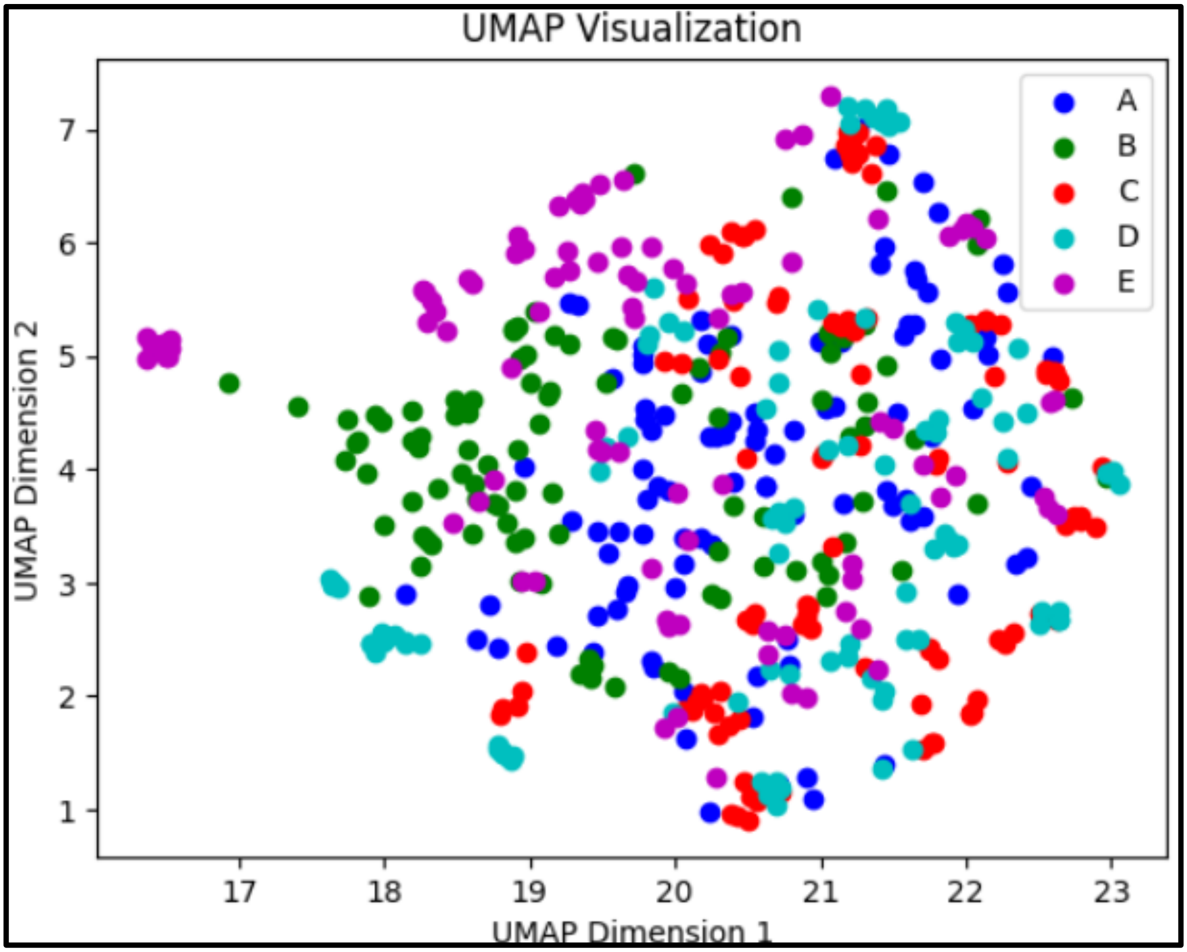

Time
Signal Plot for S002.txt
1500
1000
500
0
−500
−1000
−1500
0
1000
2000
3000
4000
Time

This function, 'preprocess_eeg_data,' takes a base folder path and a data mapping dictionary, iterating through specified target folders and their associated subfolder names. It systematically addresses file naming conventions and constructs file paths before implementing key data preprocessing steps. These steps include imputing missing values using a mean imputer, applying bandpass filtering to reduce noise, and standardizing the data to ensure consistency across the dataset. The script further allows for potential data augmentation if needed. Notably, it meticulously handles file processing, providing informative print statements indicating successful data processing and saving. In the main execution section, the script sets the base folder path and data mapping before calling the main function, thereby initiating the comprehensive EEG data preprocessing process. This script stands as a valuable tool for researchers and analysts seeking to prepare EEG data for subsequent investigations. After preprocessing, the signals look like this.



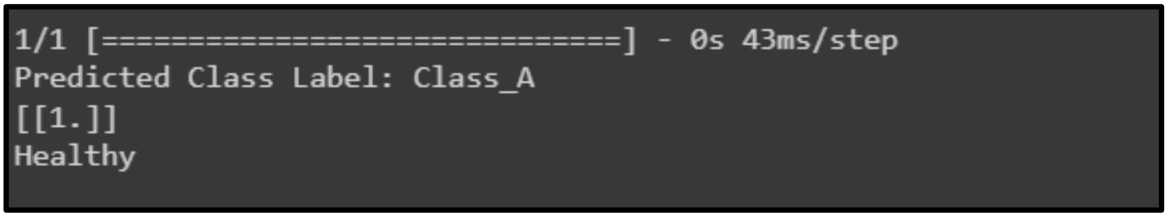

Preprocessed Signal Plot for F_preprocessed_001.txt
Amplitude
3
2
1
0
−1
−2
−3
−4
0
1000
2000
3000
4000
Time

The next function, 'extract_features_umap,' systematically processes preprocessed data files within specified target folders, loading and aggregating them for subsequent feature extraction. If data is present, it standardizes the aggregated data and applies UMAP to reduce dimensionality, yielding a two-dimensional representation. The script further provides a visualization function, 'visualize_umap,' to depict the UMAP embeddings for each target folder with distinct colors. In the main execution section, the script sets the base folder path and data mapping, calling the 'extract_features_umap' function to initiate feature extraction and UMAP visualization. This script serves as a valuable tool for researchers seeking to gain insights from EEG data by transforming it into a lower-dimensional space for improved interpretability and visualization. 'split_data,' takes UMAP features, target folders, and an optional parameter specifying the number of samples per folder. It employs 'LabelEncoder' to convert target folder labels into numerical values, ensuring consistency in the number of expected samples. The script then splits the data into training, validation, and test sets using 'train_test_split' with a specified test size and random seed. In the main execution section, the script sets the base folder path, defines data mapping, extracts UMAP features using the previously defined function, and proceeds to split the data. The resulting output includes encoded labels, along with the shapes of the training, validation, and test sets.



After this, we build a LSTM model for classification of the signals. For testing, we also provided a random image and asked the model to identify if it is healthy or not. The signals in Class A and B were labeled healthy later on and others were epilepsy signals.

The model correctly identified the signals as shown below.



```
1/1 [==============================] - 0s 43ms/step
Predicted Class Label: Class_A
[[1.]]
Healthy
```

PART 6: CONCLUSION

In this project, we tried to develop a model for the preprocessing, feature extraction, and dimensionality reduction of EEG (Electroencephalogram) data from the dataset. We incorporated functionalities for downloading, extracting, and preprocessing the data. Utilizing UMAP (Uniform Manifold Approximation and Projection), the script successfully extracts informative features and produces a reduced-dimensional representation of the EEG data. The preprocessing steps, including imputation, filtering, and standardization, contribute to enhancing the quality and consistency of the dataset.

Furthermore, the code systematically splits the preprocessed UMAP features into training, validation, and test sets, facilitating the creation of a structured dataset for machine learning applications. The LabelEncoder ensures numerical representation of target folder labels, enabling seamless integration with various classification algorithms.

The visualization of UMAP embeddings provides valuable insights into the underlying structure of the EEG data, aiding researchers and analysts in understanding patterns and relationships within the dataset.

‘

PART 7: FUTURE SCOPE

The developed script lays a solid foundation for further exploration and improvement. Future enhancements and research avenues include:

- **Model Training:** Integrate machine learning models, such as classification algorithms, to leverage the preprocessed and reduced-dimensional features for EEG data classification tasks.

- **Hyperparameter Tuning**: Conduct systematic experiments to optimize UMAP and other preprocessing parameters to enhance the quality of the reduced-dimensional representation.

- **Data Augmentation:** Explore and implement data augmentation techniques to further diversify the dataset, potentially improving model generalization.

- **Incorporate Domain Knowledge:** Integrate domain-specific knowledge and additional metadata to enhance the interpretability of the results and potentially guide preprocessing decisions.

- **Real-Time Processing:** Adapt the script for real-time EEG data processing, enabling applications in live monitoring or feedback systems.

- **Scalability:** Assess and enhance the scalability of the script to handle larger datasets efficiently.

- **Integration with Neuroinformatics Tools:** Explore integration with neuroinformatics tools and databases for a more comprehensive analysis of EEG data.

In conclusion, the model presented here serves as a robust framework for EEG data preprocessing and analysis, with numerous opportunities for future research and development to advance our understanding of brain activity patterns and improve the applicability of EEG data in various domains.