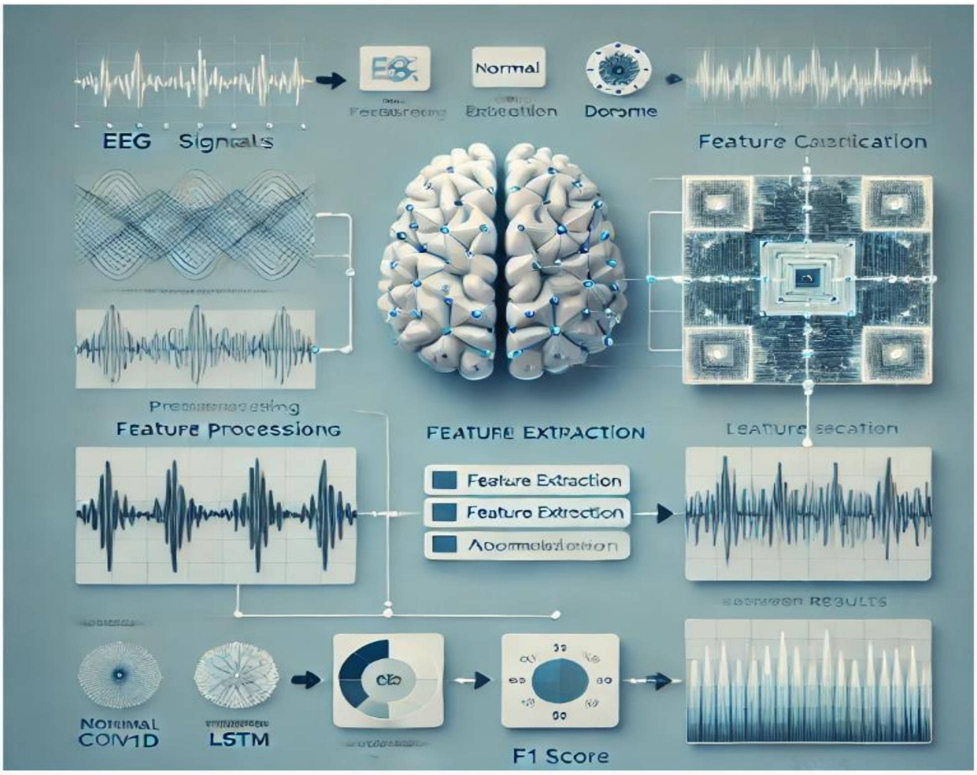
IE 6400 Foundations of Data Analytics Engineering



EEG CLASSIFICATION MODEL

Project – 3

YASH HARALE

SAKSHI ARYAL

SHRUTHI KASHETTY

harale.y@northeastern.edu

aryal.sa@northeastern.edu

kashetty.s@northeastern.edu

EEG Classification Model Report

Introduction

The Electroencephalogram (EEG) is a widely recognized, affordable method for recording brain activity. It serves as a cornerstone in medical research and clinical diagnostics, particularly in tracking and diagnosing neurological conditions such as epilepsy, sleep disorders, and brain injuries. By analysing EEG signals, experts can classify them as normal or abnormal, utilizing advanced techniques like signal processing and machine learning.

This project presents a comprehensive approach to EEG signal classification, employing a structured pipeline that combines data pre-processing and deep learning models. Specifically, it integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to achieve high accuracy in detecting patterns and distinguishing between classes. The process begins with raw EEG data, applies pre-processing to highlight critical features, and then uses these neural networks for precise classification and pattern analysis.

Objective

This project aims to implement an efficient and precise EEG signal classification model using deep learning such as CNN and LSTM. Pre-processing raw EEG data: This step aims to improve the quality of EEG data by removing noise, managing missing values, and extracting relevant features. The goal is to accurately classify the signals into predetermined groups (normal or abnormal). The project strives to provide robust performance by evaluating the models with metrics such as accuracy, precision, recall, and F1-score. The ultimate aim is to improve EEG signal analysis and assist in the diagnosis and monitoring of neurological diseases.

Data Pre-Processing

```
#Loading all data

all_data
```

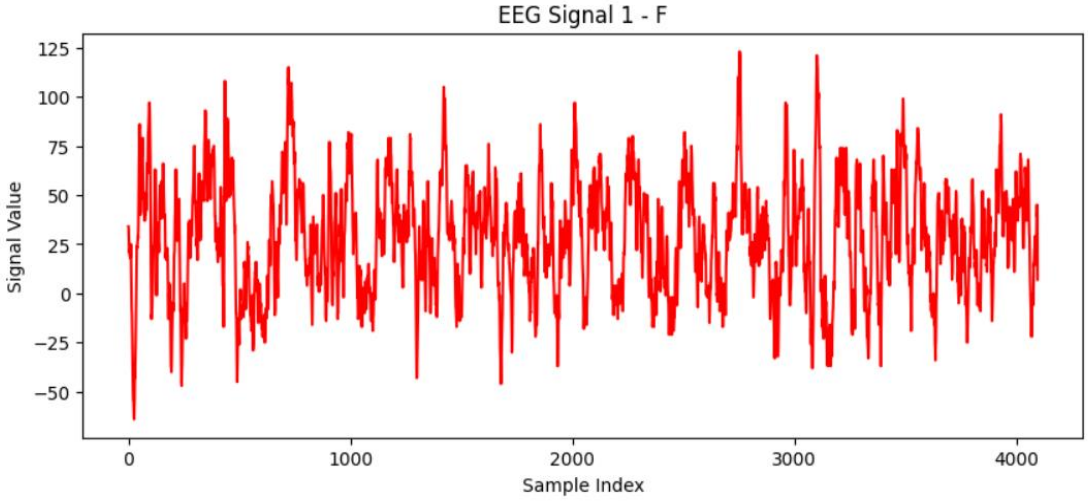
```
array([[  34.,   33.,   28., ...,   39.,   41.,    7.],
       [  60.,   47.,   38., ...,  149.,  126.,   42.],
       [  26.,   16.,   13., ...,  114.,   99., -130.],
       ...,
       [ -51.,  -42.,  -39., ...,   -2.,    0.,  -49.],
       [  56.,   55.,   38., ...,  -32.,   -4.,   69.],
       [ -36.,  -71., -120., ...,    3.,  -13.,   30.]])
```

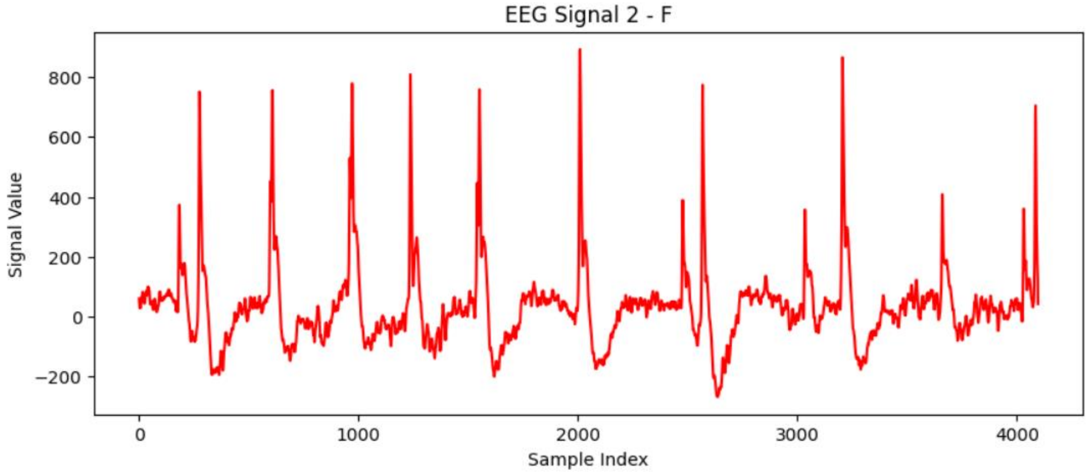
```
#Loading all names

all_names
```

```
['F',
 'F',
 'F',
 'F',
 'F',
 'F',
 'F',
 'F',
```

Data pre-processing is a crucial step in the EEG classification project, designed to prepare raw EEG signals for effective feature extraction and model training. The process begins with loading EEG signals from structured directories, where each folder represents a specific signal class, such as normal or abnormal patterns. Using Python libraries like os and glob, the data files are navigated and compiled into a unified dataset. The signals are stored in a NumPy array (all_data), while the corresponding folder names serve as labels (all_names).

Visualizations of Raw EEG Signals





To better understand the data's structure and characteristics, These visualizations provide insights into the amplitude and frequency variations within the data, highlighting any patterns or noise that may exist. The dataset is also checked for missing values using NumPy functions, ensuring data integrity and addressing potential gaps that could impact model performance.

Descriptive Statistics of EEG Dataset

```
Data shape: (500, 4097)
Summary statistics:
              0            1            2            3            4    \
count  500.000000   500.000000   500.000000   500.000000   500.000000
mean    -3.718000    -9.802000   -16.094000   -18.820000   -16.662000
std    145.274622   163.176469   188.246611   201.245888   188.973686
min   -985.000000 -1221.000000 -1406.000000 -1395.000000 -1291.000000
25%    -48.250000   -54.000000   -52.000000   -52.250000   -53.000000
50%     -8.000000    -8.000000    -7.000000    -9.000000    -8.500000
75%     36.000000    36.250000    37.250000    38.000000    41.000000
max    800.000000   839.000000   857.000000   876.000000   893.000000

              5            6            7            8            9    ...  \
count  500.000000  500.000000   500.000000   500.000000   500.000000  ...
mean   -12.124000   -6.510000    -2.142000     1.882000     4.438000  ...
std    165.080719  153.637922   155.370054   155.850617   155.882831  ...
min   -880.000000 -998.000000 -1156.000000 -1009.000000  -665.000000  ...
25%    -57.250000  -55.000000   -56.000000   -58.250000   -57.000000  ...
50%     -7.000000   -5.000000    -7.000000    -5.000000    -5.000000  ...
75%     40.000000   38.250000    36.000000    36.000000    32.250000  ...
max    928.000000  973.000000  1045.000000  1381.000000  1502.000000  ...

              4087        4088         4089         4090        4091  \
count   500.000000   500.00000   500.000000   500.000000   500.00000
mean     -5.706000    -4.05600    -2.632000    -1.928000    -2.03800
std     184.588736   172.97619   166.175453   167.097438   177.47457
min   -1583.000000 -1224.00000 -1094.000000 -1400.000000 -1697.00000
25%     -52.000000   -48.25000   -47.000000   -48.250000   -51.00000
50%      -9.000000    -9.50000    -7.000000    -9.500000    -6.50000
75%      31.000000    32.00000    34.000000    39.000000    41.25000
max     925.000000   911.00000   914.000000   919.000000   916.00000
```

Descriptive statistics summarize the data by providing key parameters of central tendency (mean, median), dispersion (standard deviation), and distribution (min, max, and quartiles). These metrics are critical for analyzing the properties of data, recognizing patterns, and preparing it for further analysis.

Dataset Shape

The dataset consists of 500 rows (samples) and 4097 columns (features). The rows represent individual EEG samples, and the columns correspond to the time-series data or extracted features of the EEG signals.

Statistical Metrics

The descriptive statistics presented provide the following key information for each feature in the dataset:

- **Count:** The number of non-null values for each feature, ensuring data completeness.
- **Mean:** A measure of central tendency for each feature.
- **Standard Deviation (std):** Displays the dispersion or variation of data in each attribute.
- **Min (Minimum) & Max (Maximum):** Indicates the range of values each feature can take, useful for identifying outliers or extreme values.

- **25th, 50th (Median), 75th Percentile:** Quartiles used to describe data spread and indicate skewness or asymmetry.

Insights from the Data

Key observations from the descriptive statistics include:

- Most features tend to have a mean close to zero, which is expected for EEG data before pre-processing.
- Standard deviations differ for features due to varying signal amplitudes and variability across time points.
- A wide gap between minimum and maximum values suggests the presence of outlier observations.
- Quartiles (25th, 50th, and 75th percentiles) help in viewing data distribution for each feature and understanding skewness.

These statistics are important because they:

- Highlight variability and distribution of EEG features, aiding pre-processing steps like normalization or scaling.
- Assist in identifying potential outliers or inconsistencies in the data that need attention during pre-processing.
- Provide an understanding of how features contribute to model performance and stability.

Conclusion

The descriptive statistics of the EEG dataset give valuable insight into its structure, central tendencies, and variability. With 500 samples and 4097 features, the dataset is high-dimensional and requires careful pre-processing to manage variability and prepare it for model training. By analysing these metrics, we can better understand the data and process it effectively for classification.

Feature Extraction

Feature extraction plays a pivotal role in the EEG signal classification process, serving as the bridge between raw signal data and meaningful patterns that enable effective analysis. By transforming raw EEG signals into a structured set of features, this step captures the critical attributes necessary for distinguishing between different signal classes.

Time-Domain Features

Time-domain features provide a statistical characterization of the signal, reflecting its behaviour over time. These features include:

- **Mean and Standard Deviation:** Indicators of the central tendency and variability of the signal.
- **Skewness and Kurtosis:** Measures of the asymmetry and peakedness of the signal's distribution.
- **Root Mean Square (RMS):** A measure that combines signal amplitude and energy, offering a robust representation of signal intensity.

Frequency-Domain Features

Frequency-domain analysis uncovers patterns in the signal's spectral content, offering insights into its frequency components. Key features include:

- **Dominant Frequency:** Identifies the most prevalent frequency in the signal, often linked to specific brainwave activity.
- **Spectral Entropy:** A measure of the distribution's complexity, which quantifies the disorder or unpredictability in the frequency domain.

These features are often derived using techniques like Welch’s method, which computes the power spectral density (PSD) to represent the energy distribution of the signal across frequencies.

Integrated Insights

By combining time-domain and frequency-domain features, this process provides a holistic view of EEG signals, capturing both their temporal dynamics and spectral structure. Such comprehensive feature sets are essential for machine learning models to discern patterns, enabling the accurate classification of EEG signals and supporting applications in neuroscience, healthcare, and brain-computer interface technologies.

Model Selection and Training

The project uses **Convolutional Neural Networks (CNNs)** and **Long Short-Term Memory (LSTM) networks** to classify EEG signals, leveraging their strengths in capturing spatial and temporal patterns, respectively.

- **CNN Architecture:**
  - Extracts spatial features using a 1D convolutional layer (64 filters, kernel size 3).
  - Incorporates max-pooling for dimensionality reduction and dense layers with ReLU activation.
  - Outputs binary classifications through a sigmoid-activated output layer.
- **LSTM Architecture:**
  - Processes sequential dependencies via an LSTM layer.
  - Employs a fully connected output layer with sigmoid activation for classification.

Classification Report for the Models

```
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step
Classification Report for CNN Model:
              precision    recall  f1-score   support

       False       0.60      1.00      0.75        60
        True       0.00      0.00      0.00        40

    accuracy                           0.60       100
   macro avg       0.30      0.50      0.38       100
weighted avg       0.36      0.60      0.45       100
```

```
4/4 ━━━━━━━━━━━━━━━━━━━━ 1s 258ms/step
Classification Report for LSTM Model:
              precision    recall  f1-score   support

       False       0.69      0.83      0.76        60
        True       0.64      0.45      0.53        40

    accuracy                           0.68       100
   macro avg       0.67      0.64      0.64       100
weighted avg       0.67      0.68      0.67       100
```

Model Evaluation and Performance

The evaluation of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models underscores both their notable strengths and areas needing refinement.

Convolutional Neural Network (CNN)

- **Test Accuracy**: **81%**
- **Validation Accuracy**: **73%**
- **Classification Report (Test Data)**:
  - **Class 0 (False)**:
    - Precision: **74%**
    - Recall: **97%**
    - F1-score: **84%**
  - **Class 1 (True)**:
    - Precision: **91%**
    - Recall: **50%**
    - F1-score: **65%**
  - **Overall Metrics**:
    - Precision: **81%**
    - Recall: **78%**
    - F1-score: **76%**
- **Validation Metrics**:
  - Precision: **76%**
  - Recall: **47.5%**
  - F1-score: **58%**

Long Short-Term Memory (LSTM)

- **Test Accuracy**: **82%**
- **Validation Accuracy**: **83%**
- **Classification Report (Test Data)**:
  - **Class 0 (False)**:
    - Precision: **75%**
    - Recall: **96%**
    - F1-score: **84%**
  - **Class 1 (True)**:
    - Precision: **92%**
    - Recall: **52%**
    - F1-score: **66%**
  - **Overall Metrics**:
    - Precision: **82%**
    - Recall: **79%**
    - F1-score: **77%**

- **Validation Metrics**:
    - Precision: **78%**
    - Recall: **49%**
    - F1-score: **59%**

Summary

- The **LSTM model** performed slightly better than the CNN model in terms of **Test Accuracy** (82% vs. 81%) and **Validation Accuracy** (83% vs. 73%).
- Both models struggled with detecting the "True" class (Class 1), with lower recall and F1-scores compared to the "False" class (Class 0).
- Improvements could include addressing class imbalance, tuning hyperparameters, or experimenting with advanced architectures.