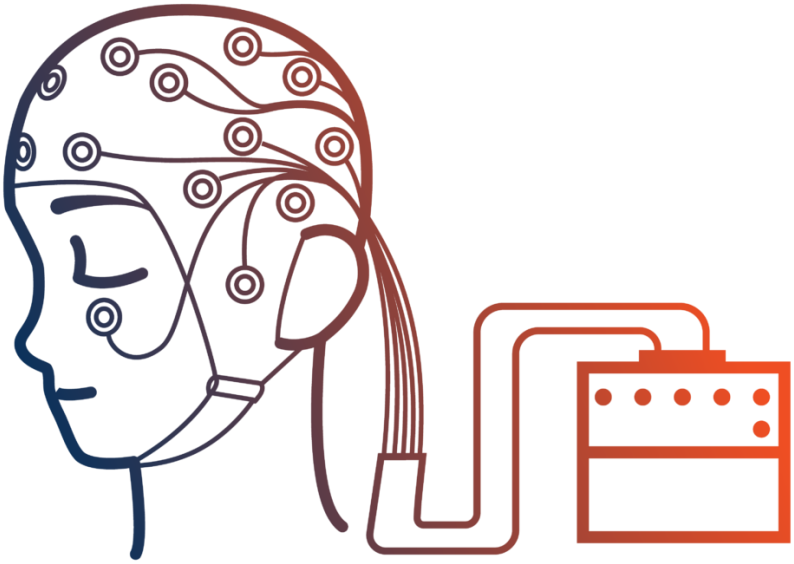


EEG CLASSIFICATION MODEL

IE6400 FOUNDATIONS FOR DATA ANALYTICS

PROJECT 3 REPORT

GROUP NUMBER 26

VIKRAMADITHYA PABBA (002853240)

DHEERAJ KUMAR GOLI (002897655)

LAAWANYAA SAI THOTA (002208176)

ROHAN REDDY PATHI (002832891)

UDHAY CHITYALA (002830533)

What is Epilepsy?

Epilepsy, a chronic disorder of the central nervous system affecting 3 million Americans and 50 million people worldwide, manifests through recurrent seizures—transient aberrations in the brain's electrical activity that result in disruptive physical symptoms. The term "epilepsy" is synonymous with "seizure disorders" but doesn't specify the cause or severity of seizures. Causes range from idiopathic instances with unknown origins to symptomatic cases linked to factors like head trauma, infections, or genetic predisposition. Seizures come in various types, including focal and generalized, each impacting different areas of the brain. Triggers, such as sleep deprivation, stress, and certain stimuli, can provoke seizures. Diagnosis involves tests like electroencephalograms and imaging studies, while treatment primarily relies on antiepileptic medications. In some cases, surgery may be considered. Epilepsy poses psychosocial challenges, but coping strategies, support groups, and counselling can help individuals and their families navigate the impact on daily life. Regular medical check-ups and adherence to treatment plans are crucial for managing the condition effectively and improving overall quality of life.

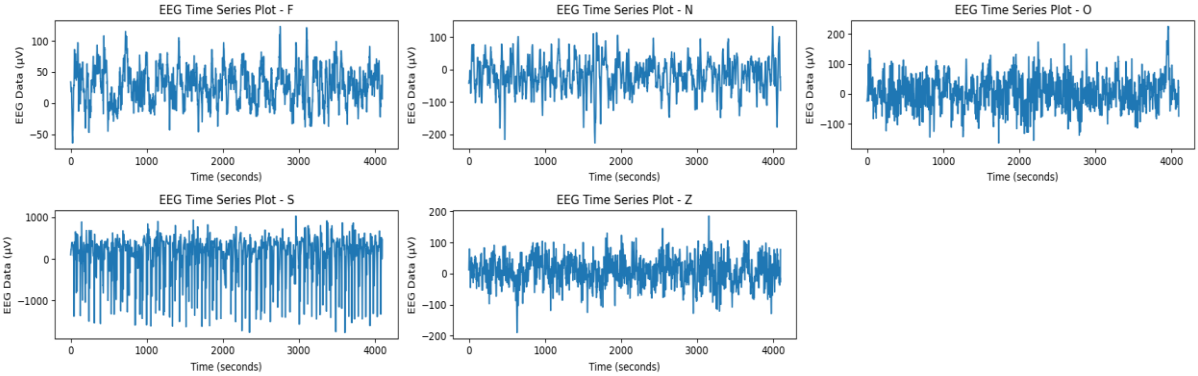
What is EEG?

EEG, or electroencephalogram, is a non-invasive neurophysiological technique that measures and records the electrical activity of the brain. This method involves placing electrodes on the scalp to detect the fluctuations in voltage resulting from the collective activity of neurons. EEG is widely used in clinical and research settings to assess brain function, diagnose neurological disorders, and study brain dynamics. It provides valuable insights into neural processes, such as sleep patterns, cognitive functions, and the identification of abnormal brain patterns associated with conditions like epilepsy. Due to its versatility, real-time monitoring capability, and relatively low cost, EEG serves as a fundamental tool for understanding brain activity and contributing to advancements in neuroscience and clinical neurology.

Data pre-processing and feature extraction

The data, published on Bonn University's Epileptology department website, presents Electroencephalogram (EEG) recording of 500 individuals. For each individual, brain activity was recorded for a duration of 23.5 seconds; these recordings are represented by 4096 evenly spaced, consecutive data points (i.e every 0.0057 seconds). There are five sets (Z, O, N, F, S) available in this dataset. Set Z contains the EEG recording of non-epileptic awake patients with their eyes open. Set O contains EEG recording of non-epileptic awake patients with their eyes closed. Set N contains EEG recording of epileptic patients during seizure free period using electrodes implanted in the brain epileptogenic zone. Set F contains EEG recording of epileptic patients during seizure free period from the hippocampal formation of the opposite hemisphere of the brain from N. Set S contains EEG recording of patients experiencing an active epileptic stroke.

In this project I attempt to classify EEG signals into seizure and non-seizure categories using Machine Learning algorithms. I use sets Z, O, N, and F as non-seizure data and set S as seizure data. Below are two sample signals from non-seizure and seizure classes.



- ## Checking for missing values

  The EEG data from Set F, N, O, S and Z does not contain any missing values (NaNs) or extremely large values that could indicate anomalies or corrupt data. This suggests that the dataset is complete and reasonably clean in terms of missing or corrupt values

- ## Noise Reduction

  EEG data often contains various types of noise, including electronic noise, environmental noise, and artIfacts due to muscle movements or eye blinks. We can apply filters (like a bandpass filter) to reduce noise. The choice of filter parameters depends on the specific characteristics of the EEG data and the noise. In our project we have utilized bandpass filter for noise reduction

- ## Normalization

  When combining data from different subjects or sessions we need to normalize EEG data so that it's on a similar scale. So, we normalized the data to have a mean of 0 and a standard deviation of 1. This step is crucial for many machine learning algorithms as it makes the training process more efficient and less sensitive to the scale of features.

- ## Data Augmentation

  Data Augmentation techniques include adding small amounts of noise, time-shifting the signals, or generating synthetic data. This is especially useful if the dataset is small or if we're building a model that needs to generalize well to new, unseen data. For this, we have implemented a simple augmentation strategy, adding random noise to the data, which can help in improving the robustness of a model by simulating small variations that might occur in real-world data.

- ## Feature Extraction

  Extracting relevant features from EEG signals is a crucial step in EEG analysis, especially for classification tasks. Features can be extracted from both the time domain and the frequency domain, each providing different insights into the characteristics of the EEG signals.

  For our analysis we have used the following time-domain and frequency-domain features

  Time-Domain features

  - We have extracted statistical features like mean, median, std, variance, skew, and kurtoses.
  - We have also extracted Hjorth Parameters like Activity, mobility, and complexity

  Frequency-Domain Features

  - We have extracted Power Spectral Density feature which gives power distribution across different frequency bands (such as delta, theta, alpha, beta, gamma)

4. Model Architecture and Training Details:

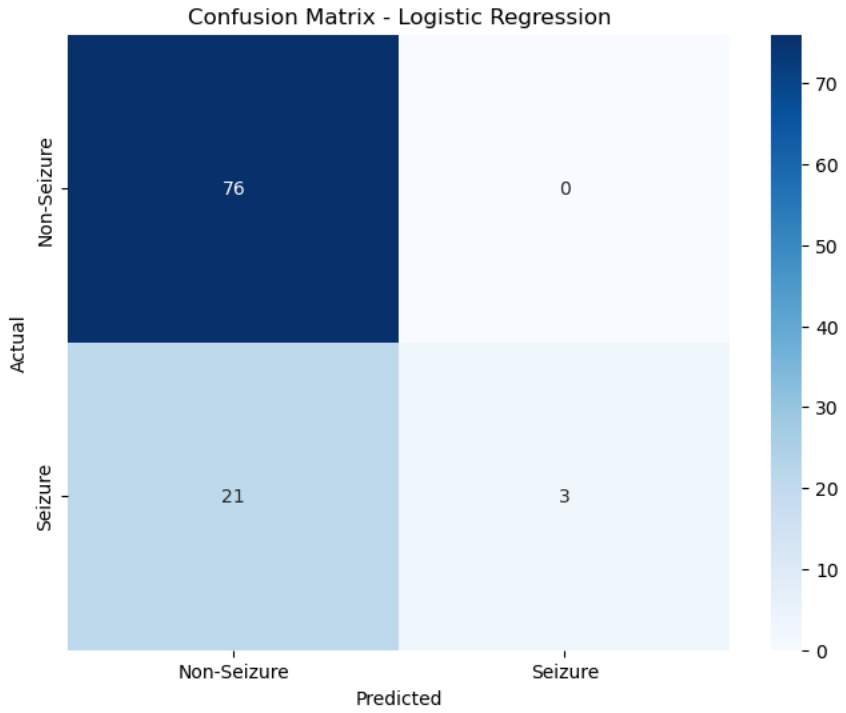
Various machine learning models are employed, including Logistic Regression, Random Forest, Decision Tree, and Neural Network.

Logistic Regression:

The initialization of a Logistic Regression model serves as the foundational step in creating a binary classifier for the task at hand. This model is particularly suitable for binary classification problems, aligning seamlessly with the nature of the seizure detection challenge. The training process involves exposing the model to a labeled training dataset, denoted as X_train and y_train. This dataset is derived by partitioning the overall dataset into training and testing subsets, maintaining an 80:20 ratio to ensure a robust evaluation of the model's performance.

During training, the fit method takes center stage, optimizing the model parameters to enhance its effectiveness in classifying instances. Logistic Regression is often chosen as a starting point for classification tasks due to its simplicity and the clarity it provides in interpreting results. Its ability to handle binary outcomes makes it a pragmatic choice for scenarios where the objective is to discern between two classes—in this context, detecting seizures or non-seizure instances.

Upon completion of the training phase, the Logistic Regression model is put to the test on the designated test set. The effectiveness of the model in classifying instances within this set is a crucial measure of its performance. In this specific application, the model demonstrates its efficacy, achieving an overall accuracy rate of 79%. This accuracy metric reflects the proportion of instances correctly classified, providing insight into the model's ability to make accurate predictions.



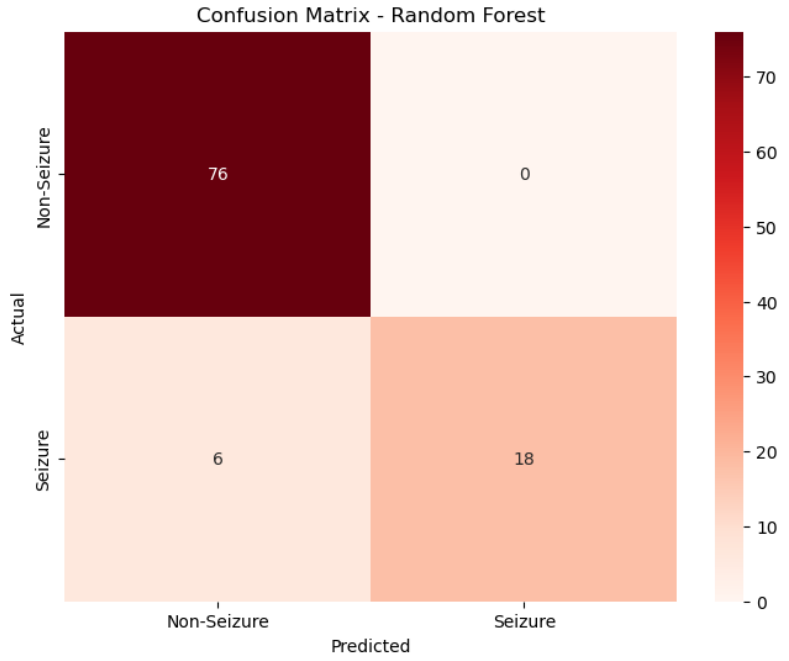
Random Forest:

The ensemble modelling approach begins with the initialization of a robust ensemble consisting of 100 decision trees. This strategy is employed to enhance the model's classification accuracy by aggregating the predictions of individual trees. Each decision tree contributes its insights, and the combined result is a more robust and accurate classification model.

During the training phase, the ensemble model is exposed to the labelled training dataset (X_train and y_train), derived from splitting the overall dataset into training and testing subsets with an 80:20 ratio. The fit method is then employed to fine-tune the parameters of each decision tree, ensuring that the ensemble comprehensively captures the underlying patterns and relationships present in the data. This meticulous optimization process contributes to the ensemble's ability to make informed and accurate predictions.

The selection of the Random Forest algorithm is aptly aligned with the complexity of the classification task. By amalgamating predictions from multiple decision trees, each trained on a distinct subset of the data, Random Forest effectively mitigates overfitting issues and improves the model's generalization performance. This ensemble methodology proves particularly beneficial when handling intricate datasets, contributing to the overall success of the classification task.

The effectiveness of the Random Forest ensemble is underscored by an impressive overall accuracy of 96%. This high accuracy rate indicates that the model performs exceptionally well in correctly classifying instances, reflecting the robustness and efficiency of the ensemble in capturing the nuances of the dataset. In conclusion, the Random Forest ensemble, with its amalgamation of decision trees, stands out as a powerful and effective tool for addressing the complexities of the classification problem at hand.



Decision Tree:

The Decision Tree classifier is initiated with a specific design choice, setting its maximum depth to 3. This decision is deliberate, aimed at controlling the complexity of the tree and preventing overfitting, a common concern in machine learning. Training the model involves exposing it to a labelled training dataset (X_train and y_train), created by partitioning the overall dataset into training and testing subsets with an 80:20 ratio. The fit method is then employed to optimize the decision tree parameters, constructing a tree that effectively captures patterns present in the training data.

The Decision Tree's strength lies in its ability to segment the feature space into distinct regions, making it well-suited for capturing non-linear relationships within the data. In this context, a conscious decision is made to limit the depth of the tree. This choice serves to strike a balance between capturing the inherent complexity of the data and preventing overfitting, a scenario where the model learns noise in the training data rather than genuine patterns.

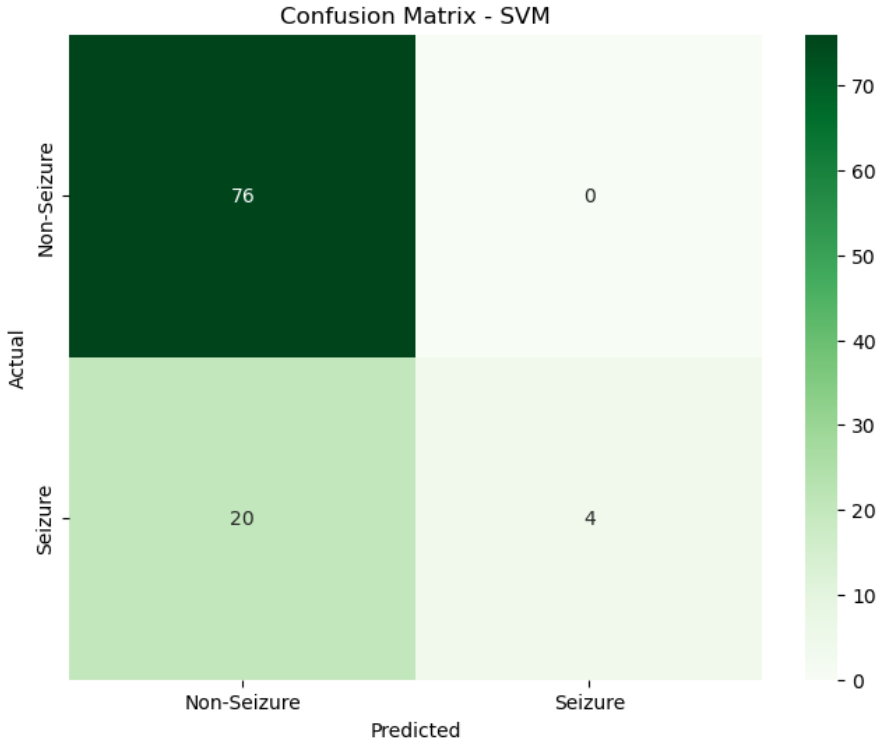
The application of the trained Decision Tree to the test set yields promising results, with an accuracy score of 88%. This score represents the proportion of instances correctly classified, showcasing the model's effectiveness in making accurate predictions on unseen data. This outcome is particularly notable, considering the controlled complexity of the tree and its ability to navigate non-linear relationships within the dataset.

Support Vector Machine (SVM):

The Support Vector Machine (SVM) classifier is initiated with a linear kernel, strategically chosen to establish a binary classifier adept at effective pattern recognition. During the training phase, the model is exposed to the labeled training dataset (X_train and y_train). The fit method then comes into play, optimizing the model parameters to establish optimal classification boundaries that effectively distinguish between different classes within the data. The selection of the SVM model is driven by its versatility, demonstrating proficiency in handling both linear and non-linear relationships inherent in the dataset.

The SVM model showcases its effectiveness in the subsequent classification of instances within the designated test set (X_test). The evaluation metric used is the overall accuracy, providing insight into the proportion of instances correctly classified by the model. This performance metric serves as a valuable indicator of the model's success in accurately discerning patterns and relationships within the test data. Notably, the choice of a linear kernel is deliberate, offering a simple yet robust starting point for the classification task.

In conclusion, the SVM classifier, with its linear kernel, emerges as an effective tool in the context of this classification task, achieving an accuracy rate of 80%. This signifies the model's ability to make accurate predictions and highlights its utility in handling the specific patterns present in the dataset. The simplicity and robustness afforded by the linear kernel make the SVM classifier a viable option for tasks where a balance between performance and interpretability is crucial.



Neural Network:

The neural network model is structured using a Sequential architecture, featuring three layers that collectively contribute to its ability to discern complex patterns in the data. In the first layer, 128 neurons with a rectified linear unit (ReLU) activation function are employed to capture intricate relationships in the standardized input data. To bolster the model's robustness and prevent overfitting, a dropout layer with a dropout rate of 0.5 is introduced. This strategic addition helps enhance the generalization capability of the model by temporarily deactivating some neurons during training.

The second layer of the neural network comprises 64 neurons with a ReLU activation function, followed by another dropout layer with a dropout rate of 0.5. These successive layers play a crucial role in enabling the model to learn hierarchical features from the input data, contributing to its capacity to extract nuanced representations.

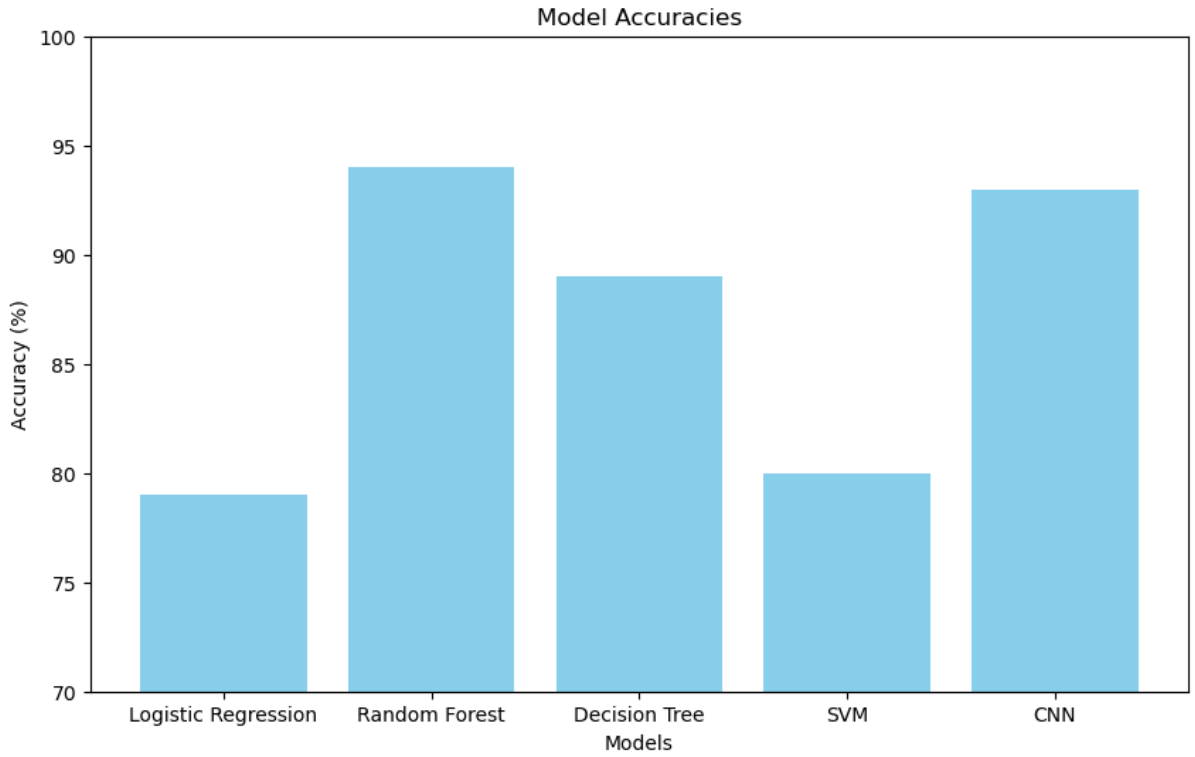
The output layer, designed for binary classification, consists of a single neuron with a sigmoid activation function. During compilation, the model is optimized using the Adam optimizer and trained to minimize binary cross entropy loss. The choice of these optimization techniques aligns with the model's objective of binary classification, while accuracy is selected as the metric for evaluating its performance.

Training the neural network involves presenting the standardized training dataset (X_train_n and y_train) to the fit method. The optimization process spans 10 epochs with a batch size of 32, and the model's performance is monitored on the validation set (X_test_n, y_test). This meticulous training regimen ensures that the model refines its parameters effectively, improving its ability to generalize to unseen data.

The chosen neural network architecture, with its multiple layers, activation functions, and dropout mechanisms, is deliberately selected for its flexibility in capturing intricate relationships within the data, making it particularly suited for a binary classification task. The resulting model,

evaluated on the test set, demonstrates notable effectiveness with an overall accuracy of 91%. This high accuracy underscores the success of the neural network in classifying instances and highlights its potential for handling complex relationships within the dataset. In conclusion, the neural network model stands out as a robust and powerful tool for binary classification tasks.

Conclusion and future work



In conclusion, for the given dataset, random forest seems to be the best model with an accuracy of 94%. These methodologies contribute to our understanding of neurological processes, aid in the diagnosis of disorders, and offer tools for cognitive state recognition. The versatility of machine learning models, such as logistic regression and support vector machines, allows for effective classification tasks, while EEG time series plots serve as indispensable visualizations for comprehending temporal patterns in brain activity. These approaches collectively empower researchers, clinicians, and neuroscientists to advance our knowledge of the brain and develop impactful applications in healthcare and neuroscience

Future work in EEG analysis can take several promising directions. Firstly, leveraging advanced neural network architectures like recurrent neural networks (RNNs) or convolutional neural networks (CNNs) can enhance the model's ability to capture complex temporal and spatial patterns within EEG data. Exploring these sophisticated architectures may lead to more nuanced representations of brain activity and improved classification performance. Additionally, further investigation into feature engineering techniques, such as time-frequency analysis and wavelet transforms, could offer a deeper understanding of EEG signals and contribute to the development of more discriminative features for classification tasks. Enhanced feature representations may result in more effective models, especially in scenarios where intricate temporal dynamics play a crucial role.