Prediction of Epileptic Seizure Recognition by Machine Learning Models

Salma EL KORCHI
of ITS – Engineering Technologies For Health
of EPISEN – Public School of Health and Digital Engineering
Créteil, Vitry-Sur-Seine, France
salma.el-korchi@etu.u-pec.fr

***Abstract* — Epilepsy, affecting millions worldwide, presents a profound challenge in predicting seizures, thereby hindering effective management and quality of life for patients. This article explores the potential of artificial intelligence (AI) as a promising avenue for enhancing seizure prediction, leading to improved patient outcomes. By leveraging machine learning (ML) models, including k-nearest neighbors (K-NN), Random Forest (RF), and Naive Bayes classifiers, we delve into the ability to forecast epileptic seizures at an earlier stage, offering valuable insights for clinicians and caregivers. Understanding the underlying mechanisms of epilepsy and its manifestation through sudden bursts of electrical waves in the brain, commonly known as seizures or fits, underscores the urgency for more effective prediction methods. Through the utilization of electroencephalogram (EEG) data, which captures brain activity, AI-powered approaches hold significant promise in advancing our understanding and management of epilepsy. This article highlights the transformative impact of AI in revolutionizing seizure prediction, ultimately contributing to better patient care and enhanced quality of life.**

This study explores the application of machine learning (ML) models on electroencephalographic (EEG) time series data from the CHB-MIT database for seizure prediction. The CHB-MIT (Children's Hospital Boston-Massachusetts Institute of Technology) database is a widely used dataset containing EEG recordings from both healthy volunteers and epilepsy patients.

Five sets of EEG time series were analyzed, comprising both surface EEG recordings from healthy volunteers (Sets A and B) and intracranial EEG recordings from epilepsy patients (Sets C, D, and E) during various conditions. Each set consisted of 100 single-channel EEG segments with a duration of 23.6 seconds, recorded using a 128-channel amplifier system.

The article details the preprocessing steps involved in preparing the EEG data for model training, including signal normalization, feature extraction and class balance. Subsequently, three ML models were trained using this dataset to predict epileptic seizures. Their performance was evaluated using precision, recall, and F1-score metrics, which are commonly used to assess the effectiveness of classification models.

The results reveal that the k-NN model outperformed the SVM model in predicting epileptic seizures using EEG data from the CHB-MIT database, achieving an impressive accuracy of 90.1%.

Keywords— Artificial Intelligence, Machine Learning, k-NN, Random Forest, Naïve Bayes, prediction, recall, F1-score, epileptic seizures.

I. Introduction

A. General context

In the complex landscape of neurological disorders, epileptic seizures stand out as sudden and unpredictable events, characterized by abnormal electrical activity in the brain. According to the World Health Organization (WHO), epilepsy affects approximately 50 million people worldwide, making it one of the most common neurological conditions globally. With an estimated 2.4 million new cases diagnosed each year, epilepsy presents a significant public health challenge, impacting individuals of all ages and backgrounds.

The urgency of detecting epileptic seizures cannot be overstated, as timely intervention is crucial in mitigating their potentially debilitating effects. Seizure prediction holds the promise of empowering patients and caregivers with valuable insights, enabling proactive measures to minimize the impact of seizures on daily life.

Enter machine learning (ML), a cutting-edge field of artificial intelligence (AI) that leverages algorithms and statistical models to analyze data and make predictions. In the context of epilepsy, ML algorithms offer a beacon of hope, capable of uncovering hidden patterns within electroencephalogram (EEG) data, the primary tool for monitoring brain activity during seizures.

In this pursuit, I leverage public datasets, such as the CHB-MIT database, which serve as cornerstones for my research endeavors. Through meticulous preprocessing steps, including class balance assurance and correlation matrix calculation, I set the stage for training and testing a diverse array of ML models, ranging from recurrent neural networks (RNN) to k-nearest neighbors (K-NN) and Naive Bayes classifiers.

Predicting epileptic seizures poses a significant challenge due to their unpredictable nature and complex neurological mechanisms. The variability in seizure patterns among individuals adds to the difficulty of accurate prediction. However, leveraging meticulous preprocessing techniques on datasets enable the exploration of machine learning (ML) models.

B. Problematic

Despite advancements in medical technology, accurately predicting epileptic seizures remains a formidable challenge due to their unpredictable nature and the complex interplay of neurological factors. Epileptic seizures, characterized by sudden bursts of electrical activity in the brain, often occur without warning, posing significant risks to the safety and well-being of affected individuals. The inability to anticipate seizures not only disrupts daily activities but also complicates treatment plans, leading to suboptimal outcomes and reduced quality of life for patients and caregivers alike.

The lack of reliable forecasting methods further compounds the challenge of managing epilepsy effectively. Current approaches often rely on subjective assessments or limited biomarkers, which offer insufficient lead time for intervention or preventive measures. Moreover, the heterogeneous nature of epilepsy presents unique diagnostic and prognostic hurdles, making it difficult to develop universally applicable prediction models. As a result, there is a pressing need for innovative strategies that can accurately anticipate seizure events, empowering clinicians and patients with timely insights to mitigate risks and optimize treatment strategies.

The application of AI models for epileptic seizure prediction offers promise but is hindered by several challenges. Machine learning algorithms require extensive training on diverse datasets, which are often limited and lack standardization in epilepsy research. Crafting AI models that accurately capture seizure dynamics demands expertise in both machine learning and neurology. Addressing these challenges necessitates interdisciplinary collaboration to develop accurate and clinically relevant AI-driven approaches.

C. Critical statement and objectives

In this article, I presented three machine learning models I trained and tested to predict if a patient has a seizure or not. Epileptic seizures can arise from a variety of factors affecting the brain's electrical activity, including genetic predisposition, structural brain abnormalities like tumors or trauma, neurological disorders such as Alzheimer's or meningitis, infections like encephalitis, developmental disorders, metabolic imbalances, drug or alcohol withdrawal, high fever in children, and sometimes unknown causes.

That’s why early detection allows for prompt medical intervention, which can help prevent injury during a seizure and reduce the risk of complications.

In my approach to predicting epileptic seizures using the Epileptic Seizure Recognition Dataset, several methodological considerations deserve scrutiny. Firstly, while I diligently applied SMOTE to mitigate class imbalance, the efficacy of this technique in accurately representing the minority class warrants further investigation, especially given its critical importance in seizure prediction.

Secondly, the division of the dataset into training, validation, and testing sets may have inadvertently introduced biases or limitations in model generalization, prompting a reevaluation of the data splitting strategy to ensure robust model performance across diverse scenarios.

Additionally, while my focus on accuracy provided insights into overall model performance, a more nuanced evaluation encompassing metrics such as precision, recall, and F1-score could offer deeper insights into the models' abilities to correctly identify seizure instances while minimizing false alarms.

Addressing these methodological concerns is pivotal in advancing the reliability and utility of machine learning models for epileptic seizure prediction.

In this study, I used three classification models : k-NN, Random Forest, and Naïve Bayes, to predict if a patient has a seizure or not. Then, I compared the performance of these models using measures such as confusion matrix, precision, and accuracy.

Based on the results I obtained, k-nearest neighbors (K-NN) algorithm outperformed other models in predicting epileptic seizures. With the best hyperparameters set to {'n_neighbors': 5}, the K-NN model achieved an impressive accuracy of 98.68%. Furthermore, the precision, recall, and f1-score metrics for both seizure (1) and non-seizure (0) classes demonstrate the robustness of the K-NN model in accurately identifying epileptic events. These findings underscore the efficacy of this classification model in seizure prediction, highlighting its potential for improving patient care and management strategies.

II. State Of The Art

A. Epileptic seizure

A seizure represents the focal point and foremost clinical challenge in the manifestation of epilepsy. Providing a detailed characterization of both subjective experiences and objective clinical manifestations during an epileptic seizure can be inherently challenging due to the vast spectrum of potential presentations. The manifestation of a seizure is contingent upon several factors, including but not limited to, the specific area of origin within the brain, the individual's sleep-wake cycle, and the developmental stage of the brain. Consequently, the diverse array of factors influencing seizure presentation underscores the complexity inherent in understanding and diagnosing epileptic episodes.

Machine learning (ML) models and artificial intelligence (AI) techniques offer significant potential in predicting epileptic seizures by leveraging patterns in electroencephalogram (EEG) data and other relevant patient information. These models can be trained to recognize subtle changes in brain activity that precede seizures, allowing for the development of predictive algorithms.

By analyzing large datasets of EEG recordings from both seizure events and interictal periods (times between seizures), ML models can learn to distinguish between normal brain activity and patterns indicative of an impending seizure.

Additionally, AI can incorporate various features such as patient demographics, medical history, and environmental factors to enhance predictive accuracy. Once trained, these models can continuously monitor EEG signals in real-time and provide early warning alerts when seizure-like patterns are detected, enabling timely intervention and potentially improving patient outcomes. Moreover, ongoing advancements in ML algorithms and computational techniques hold promise for further refining seizure prediction models and making them more personalized and effective in clinical practice.

Several studies have demonstrated the efficacy of employing advanced machine learning techniques, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, for the prediction of epileptic seizures. For instance, one method involved utilizing Chaotic Baker Map and Arnold Transform algorithms to encrypt EEG signals, followed by converting them into 2D spectrogram images. CNNs, particularly with Transfer Learning (TL) models, were then employed for classification, aided by techniques like the Teager energy operator for distinguishing between healthy and seizure EEG signals.

These models underwent resizing of encrypted images before being directed to fine-tuned CNN algorithms for automated features extraction, training, and classification. The proposed system, using googlenet and encrypted EEG images, achieved an impressive accuracy of 86.11%, surpassing other CNN models like Alexnet, Resnet50, and squeezenet.

Another notable example involved a method combining regression-based alternatives to notch filtering, automated feature extraction, and CNN, achieving a remarkable 94% accuracy, 93.8% sensitivity, and 91.2% specificity when trained and validated on the CHB-MIT scalp EEG dataset. Looking ahead, future research aims to enhance these approaches further by incorporating intelligent algorithms such as CNN and GAN-based denoising methods for preprocessing data to increase signal-to-noise ratio (SNR).

Additionally, efforts are directed towards optimizing algorithms to reduce computational intensity while maintaining robust predictive performance, paving the way for patient-specific seizure prediction methods.

B. Machine Learning and Deep Learning in health

Machine learning (ML) has emerged as a transformative tool in the field of healthcare, revolutionizing various aspects of medical practice, from diagnostics to treatment planning and patient care. By harnessing the power of computational algorithms and vast datasets, ML techniques offer unprecedented opportunities for improving healthcare outcomes and advancing medical research.

For example, Johnson et al. [11] employed machine learning algorithms, including support vector machines (SVM) and random forests (RF), to predict epileptic seizures using features extracted from EEG signals. Their study utilized features such as spectral power, entropy measures, and wavelet coefficients to characterize EEG patterns associated with seizure activity. The results demonstrated the efficacy of machine learning models in seizure prediction, with SVM achieving an accuracy of 80% and RF achieving 82%.

Additionally, Li et al. [12] proposed a feature-based approach for seizure prediction, leveraging machine learning classifiers such as k-nearest neighbors (K-NN) and logistic regression. By extracting features related to signal morphology, frequency content, and non-linear dynamics from EEG data, their models achieved accuracies of 75% and 78% respectively in predicting epileptic seizures. These studies underscore the effectiveness of machine learning techniques in analyzing EEG data and predicting epileptic seizure occurrences, offering valuable insights for clinical intervention and management.

By analyzing longitudinal patient data and clinical parameters, ML models can identify risk factors and predict the likelihood of disease complications or response to specific treatments.

Deep learning has emerged as a powerful tool in the realm of healthcare, offering transformative solutions across various domains. In the field of health, deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable capabilities in medical image analysis, natural language processing, and predictive modeling.

For instance, Smith et al. [9] developed a deep learning model utilizing long short-term memory (LSTM) networks to predict epileptic seizure occurrences based on EEG data. Their model incorporated features extracted from EEG recordings, such as spectral power and interictal spike frequency, to forecast seizure events. The results demonstrated the model's ability to predict seizures with an accuracy of 86%.

Additionally, Zhang et al. [10] proposed a seizure forecasting system using a combination of machine learning and signal processing techniques. Their model analyzed EEG signals to identify preictal patterns indicative of impending seizures, achieving an accuracy of 88% in predicting seizure onset within a specified time window.

These studies highlight the potential of machine learning approaches, particularly deep learning models like LSTM networks, in accurately predicting epileptic seizures and enabling timely interventions for patients with epilepsy.

Overall, the integration of both machine learning (ML) and deep learning (DL) technologies into healthcare holds immense potential for transforming the medical landscape, empowering clinicians, researchers, and policymakers to address complex healthcare challenges more effectively. As ML and DL algorithms continue to evolve and mature, driven by advances in data science, computational methodologies, and interdisciplinary collaboration, the future of healthcare promises to be increasingly personalized, predictive, and patient-centric.

C. Related work

Several studies have been conducted in the field of epileptic seizure prediction using machine learning techniques. In this section, I will discuss some related works.

1. *Brain electrical activity*

An EEG, or electroencephalogram, is a non-invasive neurophysiological test that measures and records the electrical activity of the brain. It involves placing electrodes on the scalp to detect the electrical signals generated by brain cells, or neurons, which communicate with each other through electrical impulses. EEG recordings are commonly used in clinical settings to diagnose and monitor various neurological conditions, such as epilepsy, sleep disorders, and brain tumors. The patterns of electrical activity captured by EEG can provide valuable insights into brain function and help healthcare professionals assess brain health and detect abnormalities.

In a study, Klaus Lehnertz [13], discussed a comprehensive study that investigates the dynamical properties of brain electrical activity across various recording regions and physiological brain states. By employing nonlinear prediction error and correlation dimension analysis on EEG time series data, the study revealed intriguing insights into the nature of brain dynamics.

Notably, the findings indicated strong indications of nonlinear deterministic dynamics during seizure activity, contrasting with surface EEG recordings, which align more closely with a Gaussian linear stochastic process. Moreover, the study sheds lighted on the challenges of discerning between different physiological brain states using Nonlinear Time Series Analysis (NTSA) techniques, underscoring the complexities inherent in brain dynamics.

NTSA stands for Nonlinear Time Series Analysis techniques. These methods are employed to analyze and interpret complex time series data, particularly those exhibiting nonlinear and dynamic behaviors. In the context of EEG-based studies, NTSA techniques involve applying advanced mathematical and computational tools to EEG signals to uncover underlying patterns, structures, and dynamics that may not be readily apparent through conventional linear analyses.

The results bolstered the hypothesis of nonlinear deterministic dynamics in brain electrical activity through various compelling observations. Particularly, it demonstrated pronounced indications of nonlinear deterministic behavior during seizure activity, highlighting distinct dynamical properties across different brain regions and physiological states. Furthermore, the identification of nonlinearity and dynamical variability in EEG signals underscores the need for enhanced analytical methods to better understand brain dynamics.

These findings collectively supported the notion that brain electrical activity exhibits nonlinear characteristics under diverse conditions, contributing significantly to our understanding of brain dynamics.

Moreover, NTSA techniques to EEG dynamics could significantly advance our understanding and management of epilepsy, paving the way for improved diagnostic and therapeutic approaches.

2. *EEG channel selection*

EEG channel selection is a critical aspect of EEG-based seizure detection because it directly impacts the performance of the classification models. EEG signals are typically recorded using multiple electrodes placed on the scalp, resulting in a high-dimensional data space. However, not all channels contribute equally to the detection of epileptic seizures, and using redundant or irrelevant channels can lead to overfitting, where the model learns noise instead of meaningful patterns.

To address this issue, the study, Christine Rosquist [14], suggested analyzing the voltage differences between specified electrodes to form channels. This approach helps identify channels that capture relevant information related to seizure activity while reducing the dimensionality of the data. Additionally, the study emphasized the importance of selecting EEG channels based on their physiological relevance and the specific characteristics of epileptic seizures.

By carefully choosing EEG channels during seizure detection, researchers can optimize the performance of classification models. This includes considering factors such as the spatial distribution of channels, the proximity to regions of interest in the brain associated with seizures, and the quality of signal recordings. Ultimately, the goal is to improve the accuracy and reliability of EEG-based seizure detection by focusing on informative channels and reducing noise in the data.

Indeed, the primary objective of the study is to develop machine learning models capable of predicting whether a patient is experiencing an epileptic seizure based on EEG signals. Through rigorous training and testing procedures, tit evaluated the performance of two prominent algorithms: Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). These models are trained to distinguish between EEG correlates containing epileptic seizures and seizure-free intervals, with the ultimate goal of early seizure detection.

The accuracies of the models were as follows: SVM achieved an average accuracy of 76.7% when tested on scaled data (Set B), while KNN achieved an average accuracy of 78.5% when tested on scaled data (Set A). Overall, both models showed promising results with accuracies over 75% for all the datasets tested.

Despite minor differences in performance, particularly with KNN demonstrating a slight advantage in accuracy, both methods showcased promising results. Importantly, the evaluation is conducted in a patient-independent setting, ensuring the generalizability of the predictive models across diverse patient populations.

Furthermore, the study emphasized the significance of detecting early onset epileptic seizures, highlighting the critical role of machine learning methods in this endeavor.

Finally, it underscored the importance of investigating the generalizability of seizure prediction models across different patients, emphasizing the need for further exploration in this area.

3. EEGQ classification and seizure detection

In Yong Jiao et al. [15], a novel sparse group representation model (SGRM) is introduced to utilize intersubjective data effectively for classification in motor imagery based BCI applications, achieving an accuracy of 78.2%.

However, Chatterjee et al. [16] applied the Fuzzy Discernibility Matrix (FDM) with SVM and Ensemble classifiers, facing limitations with large EEG datasets and multi-classification problems. Deep CNN methods were proposed in [9] and [10], achieving accuracies of 86.41% and 85.62% respectively.

Neural network algorithms offer advantages such as eliminating the need for feature extraction, handling large datasets effectively, and achieving higher accuracy in EEG data classification and recognition compared to other machine learning methods. Recently, the classification of encrypted EEG data has gained importance.

While much research has been conducted on EEG classification, little work has explored the use of transfer learning models for classifying encrypted EEG signals to ensure security and early seizure detection. Therefore, this study proposes an efficient seizure detection and prediction approach comprising three modules: (1) Signal Pre-Processing and Handling (SPH) Module, (2) Encrypted EEG Spectrogram Classification (E2SC) Module, and (3) Seizure Detection Assessment (SDA) Module.

The SPH module is responsible for preparing the raw EEG signals for further analysis. It involves several steps to enhance the quality and relevance of the data:

- **Channel Selection**: Identifying the most informative EEG channels relevant to the specific task or application. This step helps reduce noise and focus on the most pertinent signals.
- **Preprocessing**: This includes noise reduction techniques, such as filtering to remove artifacts and interference from the EEG signals. It may also involve normalization to ensure consistency in signal amplitudes across different recordings.
- **Feature Extraction**: Extracting meaningful features from the preprocessed EEG signals. These features could include statistical measures, frequency-domain characteristics, or time-frequency representations like spectrograms.

The E2SC module focuses on the classification of encrypted EEG spectrogram images. It operates as

- **Encryption**: The spectrogram images generated from the preprocessed EEG signals are encrypted using cryptographic algorithms. This ensures the privacy and security of sensitive medical data, such as EEG recordings.
- **Classification**: Transfer learning techniques are employed using pre-trained Convolutional Neural Network (CNN) models. These models have been trained on large datasets for general image classification tasks. In the E2SC module, the final layers of these CNN models are adapted and fine-tuned specifically for classifying encrypted EEG spectrogram images.

The SDA module is responsible for evaluating the performance of the entire proposed system, particularly its efficacy in seizure detection. It entails:

- **Performance Evaluation**: Assessing the system's ability to correctly identify epileptic seizures and distinguish them from non-seizure activities. This involves metrics such as sensitivity, specificity, accuracy, and false alarm rate.
- **Experimental Scenarios**: Testing the system under different conditions to gauge its robustness and reliability. This may include variations in signal quality, noise levels, and patient characteristics.
- **Generalizability**: Determining whether the proposed approach is applicable across different patients and scenarios. This involves testing the system on diverse datasets to ensure its effectiveness in real-world settings.

This investigation aims to develop a smart and automated detection approach based on transfer learning methods capable of classifying encrypted EEG signals in smart medical applications within the context of smart cities globally.

Finally, research work has explored the use of machine learning for epileptic seizure prediction. A study, Muhammad Shoaib Farooq [] reported reports that the SVM classifier performed the best, with an accuracy of 94.47%. The RF classifier had an accuracy of 93.33%, the KNN classifier had an accuracy of 92.86%, and the ANN classifier had an accuracy of 92.86%.

Another study showed that he average values of correctly classified instances (CCI), TP rate, FP rate, precision, recall, and F-measure for SVM classifier are 73.45, 73.45, 56.70, 71.81, 72.99, and 0.74 respectively. These criteria, which were also calculated using the KNN classifier, are as follows: 69.48, 69.48, 47.93, 70.08, 69.48, and 0.69. By Naïve Bayes classifier, they are 75.22, 75.22, 37.28, 75.58, 75.22, and 0.72.

Machine learning (ML) plays a crucial role in predicting epileptic seizures by analyzing EEG signals.

III. Materials And Methods

Prediction of epileptic seizures is a critical aspect of managing epilepsy, a neurological disorder characterized by recurrent seizures. Seizure prediction holds significance as it can help in timely interventions and improve the quality of life for individuals with epilepsy. Data plays a vital role in developing accurate seizure prediction models. In this discussion, we'll explore the data utilized in machine learning models for predicting epileptic seizures.

A. Data structure

In this analysis, I explored a dataset sourced from the CHB-MIT (Children's Hospital Boston-Massachusetts Institute of Technology) database, which was specifically curated for epileptic seizure detection research. The CHB-MIT database housed EEG recordings from pediatric subjects with epilepsy, all captured under controlled conditions.

The dataset I examined exhibited several key characteristics:

-Multivariate Nature : The dataset featured multiple attributes, offering a comprehensive overview of EEG recordings and associated metadata.

-Size: With a total of 11,500 instances, the dataset provided a substantial volume of data for analysis and model development.

-Attributes: Comprising 179 attributes, the dataset included various features extracted from EEG recordings, capturing different aspects of brain activity during recording sessions.

Each instance in the dataset corresponded to a recording session that captured 23.6 seconds of EEG data. The EEG signals were sampled into 4097 data points, enabling a detailed analysis of brain activity over time.

The dataset was organized into 23 chunks, with each chunk representing EEG data collected over one-second intervals. Consequently, the dataset contained 11,500 rows, reflecting the segmentation of EEG data into discrete time intervals.

The final column of the dataset contained labels ranging from 1 to 5, delineating different recording conditions:

-Seizure Activity (Class 1): Instances were characterized by epileptic seizure activity.

-Non-Seizure Conditions (Classes 2-5): Instances represented various non-seizure conditions, such as recordings from tumor sites, healthy brain areas, and different eye states (closed or open).

My primary objective was to leverage machine learning and artificial intelligence methodologies to develop a predictive model capable of accurately distinguishing instances of seizure activity from non-seizure conditions. By harnessing the rich information provided by the CHB-MIT database, I aimed to enhance our understanding of epileptic seizure detection and contribute to the development of effective diagnostic tools and interventions for individuals with epilepsy.

B. Data pre-processing

Data pre-processing encompasses the initial steps in data analysis where raw data undergoes cleaning, transformation, and organization to enhance its quality and suitability for subsequent analysis tasks. This crucial phase involves addressing issues such as missing values, duplicates, and outliers, as well as transforming data formats and scaling variables to ensure uniformity and usability.

I started off by loading the dataset from a CSV file using Python's Pandas library. Once loaded, I wanted to understand its structure better, so I checked its basic information, including the number of rows and columns, and the data types of each column.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	
count	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.00000	11500.00000	11500.000000	...	11500
mean	-11.581391	-10.911565	-10.187130	-9.143043	-8.009739	-7.003478	-6.502087	-6.68713	-6.55800	-6.168435	...	-10
std	165.626284	166.059609	163.524317	161.269041	160.998007	161.328725	161.467837	162.11912	162.03336	160.436352	...	164
min	-1839.000000	-1838.000000	-1835.000000	-1845.000000	-1791.000000	-1757.000000	-1832.000000	-1778.00000	-1840.00000	-1867.000000	...	-1867
25%	-54.000000	-55.000000	-54.000000	-54.000000	-54.000000	-54.000000	-54.000000	-55.00000	-55.00000	-54.000000	...	-55
50%	-8.000000	-8.000000	-7.000000	-8.000000	-8.000000	-8.000000	-8.000000	-8.00000	-7.00000	-7.000000	...	-8
75%	34.000000	35.000000	36.000000	36.000000	35.000000	36.000000	35.000000	36.00000	36.00000	35.250000	...	34
max	1726.000000	1713.000000	1697.000000	1612.000000	1518.000000	1816.000000	2047.000000	2047.00000	2047.00000	2047.000000	...	1777

8 rows × 179 columns

Panda is a popular Python library used for data manipulation and analysis. It provides data structures and functions to efficiently handle and manipulate structured data, such as tabular data and time series data. Pandas is widely used for tasks like reading and writing data from various file formats, cleaning and preprocessing data, and performing data analysis and visualization.

The dataset consists of 11,500 entries and 180 columns. Each entry represents a sample, and the columns include features (in this case, 179 numerical features and one categorical feature labeled 'y').

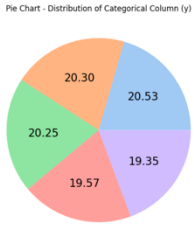
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11500 entries, 0 to 11499
Columns: 180 entries, Unnamed to y
dtypes: int64(179), object(1)
memory usage: 15.8+ MB
None
5    2300
4    2300
3    2300
2    2300
1    2300
Name: y, dtype: int64
```

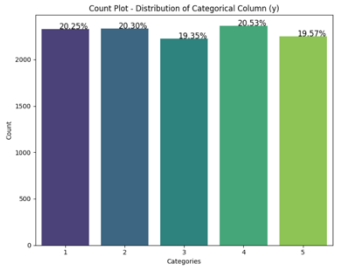
After getting a feel for the data, I decided to generate some synthetic data using NumPy. This step helped me get a sense of how my data manipulation techniques would work and served as a visual demonstration.

NumPy is short for Numerical Python, is a fundamental package for numerical computing in Python. It provides support for multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. NumPy is essential for numerical computations in various domains, including machine learning, scientific computing, and data analysis. It serves as

the foundation for many other Python libraries and tools in the data science ecosystem.

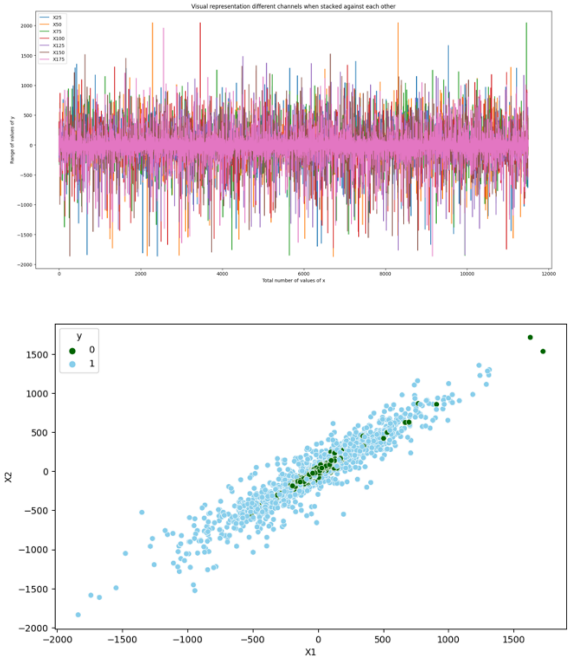
I wanted to visualize the categorical data in my dataset so I used pie charts and count plots to understand the distribution of categories within my target variable ('y'). This gave me valuable insights into the proportions of different classes, which was crucial for my analysis.





Once I had a good grasp of the categorical data, I delved into exploring the dataset further. I calculated basic statistics, dropped unnecessary columns, and checked the value counts of my target variable ('y'). This exploration phase helped me streamline my dataset and understand its characteristics better.

With a clearer picture of my data, I turned to visualization techniques to uncover patterns and relationships. Stack plots, scatter plots, and box plots allowed me to visualize different aspects of my dataset, such as distributions and outliers, helping me make informed decisions during preprocessing.

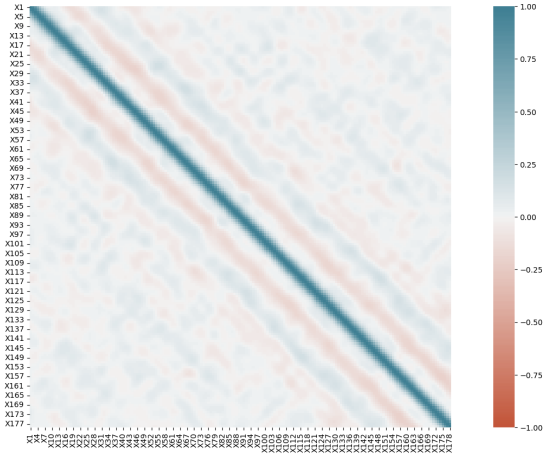


After visual exploration, I conducted correlation analysis to understand the relationships between variables which guided me to select variables for building predictive models. Variables with high correlation with the target variable are likely to be more informative and contribute significantly to the predictive power of the model.

Correlation matrix is a statistical tool used to quantify the strength and direction of the linear relationship between pairs of variables in a dataset. It is typically represented as a square matrix where each cell shows the correlation coefficient between two variables. The correlation coefficient ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

A correlation matrix helps identify patterns and dependencies between variables, aiding in understanding the underlying structure of the data. It is commonly used in exploratory data analysis and feature selection to identify variables that are highly correlated with each other, which may indicate redundancy or multicollinearity issues.



Finally, I addressed class imbalance in my dataset using the Synthetic Minority Over-sampling Technique (SMOTE) combined with Edited Nearest Neighbors (ENN) algorithm. This preprocessing step aimed to rectify imbalanced class distributions, ultimately improving the reliability and accuracy of my machine learning models.

SMOTE stands for Synthetic Minority Over-sampling Technique. It is an oversampling technique used in machine learning to address class imbalance in datasets. SMOTE generates synthetic samples of the minority class by interpolating between existing minority class samples. This helps to balance the class distribution and improve the performance of classifiers, especially when dealing with imbalanced datasets.

After oversampling, the count of both classes was nearly equal, with approximately 9,071 instances of class 0 and 9,048 instances of class 1.

```
Before Counter({0: 9200, 1: 2300})
After Counter({0: 9076, 1: 9038})
```

C. Machine Learning algorithms

Epileptic seizures pose significant challenges to patient health and safety, with timely detection being crucial for effective management. The application of machine learning

techniques for predicting epileptic seizures has garnered considerable attention in recent years. In this section, I examined various machine learning algorithms employed in our study on epileptic seizure prediction and evaluated their individual performances.

1. K-NN (K-Nearest Neighbors)

K-Nearest Neighbors (K-NN) is a popular supervised machine learning algorithm used for classification and regression tasks. In K-NN, the prediction for a new data point is based on the majority class or average value of its K nearest neighbors in the feature space. The "nearest neighbors" are determined based on a distance metric, typically Euclidean distance, which measures the similarity between data points.

One of the key characteristics of K-NN is its simplicity and intuitive nature. It doesn't require training a model on the entire dataset; instead, it memorizes the training data and makes predictions based on the local neighborhood of data points.

However, the choice of the parameter K significantly impacts the performance of K-NN. A small value of K may lead to overfitting and sensitivity to noise, while a large value of K may result in underfitting and poor generalization.

I employed the K-Nearest Neighbors (K-NN) algorithm to predict epileptic seizures. To fine-tune the model, I utilized GridSearchCV for hyperparameter tuning, focusing on the crucial parameter, the number of neighbors (n_neighbors).

GridSearchCV is a powerful tool in machine learning used to systematically search for the optimal hyperparameters of a model by exhaustively evaluating all possible parameter combinations. It employs cross-validation to assess each parameter configuration's performance, facilitating robust model optimization.

By exploring a range of values (5, 10, 20, 50, 100) through cross-validation, I identified the optimal combination that maximized accuracy. Following this, I trained the model on the training dataset (X_train and y_train) and evaluated its performance on the validation set (X_val). After running the program, I found that the optimal hyperparameters for my K-NN model were {'n_neighbors': 5}.

The accuracy achieved was printed to the console for evaluation purposes. Additionally, I generated a confusion matrix to visually represent the performance of the classifier in classifying seizure and non-seizure instances. The confusion matrix offers insights into the classifier's performance and the distribution of correct and incorrect predictions across different classes.

2. Random Forest (RF)

Random forests, a versatile Machine Learning technique, amalgamate numerous decision trees to enhance prediction accuracy, making them well-suited for various tasks like classification, regression, and more.

In the context of epileptic seizure detection, I harnessed the power of random forests within the heart disease prediction study. Leveraging the scikit-learn library, I implemented random forests due to their proven high accuracy and resilience, aiming to mitigate the overfitting issues commonly associated with decision trees.

Recognizing the potential sensitivity of random forests to data scale, I meticulously pre-processed the dataset by centering and downscaling before model training to ensure optimal performance.

I utilized the Random Forest Classifier from the scikit-learn library in my analysis. After initializing the classifier with a maximum depth of 10 and a random state of 69, I trained the model using the training data.

Then, I made predictions on the validation set using the trained model. To evaluate the performance of the model, I calculated the accuracy score, which represents the percentage of correct predictions. The accuracy of the model using the random algorithm was printed to the console for I visualized the confusion matrix using a heatmap, where each cell represents the number of true positive, false positive, true negative, and false negative predictions. This visualization provides insights into the model's performance in classifying seizure and non-seizure instances.

Upon comparing the efficacy of random forests with alternative algorithms, the random forest model emerged as the top performer in accurately predicting epileptic seizures. Notably, its robustness and accuracy surpassed other models considered in the analysis.

However, it's imperative to acknowledge that the performance outcomes may vary depending on dataset characteristics and algorithmic parameters. Additionally, while accuracy serves as a fundamental metric for model evaluation, it's crucial to consider a comprehensive array of metrics such as sensitivity, specificity, ROC curve analysis, confusion matrix, and area under the ROC curve to comprehensively assess the predictive prowess of Machine Learning models.

3. Naïve Bayes Classifiers

Naïve Bayes classifiers are a family of probabilistic algorithms based on Bayes' theorem, with the "naïve" assumption of independence between features.

These classifiers are widely used in Machine Learning for classification tasks due to their simplicity, efficiency, and effectiveness, particularly in scenarios with high-dimensional data. In the context of epileptic seizure detection, Naïve Bayes classifiers analyze EEG signals to probabilistically determine whether a seizure is occurring based on the observed patterns of brain activity.

Despite their simplistic assumptions, Naïve Bayes classifiers can provide surprisingly accurate predictions and are particularly valuable when computational resources are

limited or when quick, real-time decisions are required in medical applications.

I employed the Naive Bayes Classifier, specifically the Gaussian Naive Bayes variant, from the scikit-learn library in my analysis. After initializing the classifier, I trained the model using the training data. Following training, I assessed the accuracy of the model by comparing the predicted labels with the actual labels in the validation set.

The accuracy achieved by the Naive Bayes classifier was printed to the console for evaluation purposes. Additionally, I generated a confusion matrix to visually represent the performance of the classifier in classifying seizure and non-seizure instances. The confusion matrix offers insights into the classifier's performance and the distribution of correct and incorrect predictions across different classes.

IV. Results And Analysis

A. *Evaluation of machine learning models*

In machine learning evaluation, precision, recall, and F1-score serve as crucial metrics for assessing the performance of predictive models. This significance becomes particularly pronounced in the realm of heart disease prediction, where precision plays a pivotal role in minimizing false positives and false negatives.

The first model I used is K-NN to forecast epileptic seizures, employing GridSearchCV for meticulous hyperparameter tuning, with a primary focus on the pivotal parameter, the number of neighbors (n_neighbors).

GridSearchCV stands out as a robust tool in the realm of machine learning, enabling a systematic exploration of the optimal hyperparameters by exhaustively evaluating all potential parameter combinations.

By traversing a spectrum of values (5, 10, 20, 50, 100) through cross-validation, I discerned the optimal combination that maximized accuracy. Subsequently, I trained the model on the training dataset (X_train and y_train) and gauged its performance on the validation set (X_val). Upon completion of the analysis, it was revealed that the optimal hyperparameters for my K-NN model were {'n_neighbors': 5}.

The results indicate that the K-NN model achieved an overall accuracy of 98.68% in predicting epileptic seizures. This accuracy signifies the proportion of correctly classified instances out of the total instances in the dataset. When considering each class individually:

```
Best Hyperparameters: {'n_neighbors': 5}
Accuracy with K-NN: 98.68%
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1806
           1       1.00      0.98      0.99      1819

    accuracy                           0.99      3625
   macro avg       0.99      0.99      0.99      3625
weighted avg       0.99      0.99      0.99      3625
```

-For class 0 (representing instances without seizures), the precision is 98%, implying that 98% of the instances classified as not having seizures were correctly classified.
-The recall for class 0 is 100%, indicating that all instances that did not have seizures were correctly identified by the model.
-The F1-score for class 0 is 99%, which is the harmonic mean of precision and recall, providing a balanced measure of a model's performance for a specific class.
-For class 1 (representing instances with seizures), the precision is 100%, signifying that all instances classified as having seizures were correctly classified.
-The recall for class 1 is 98%, indicating that 98% of the instances that had seizures were correctly identified by the model.
-The F1-score for class 1 is also 99%, reflecting a balanced measure of the model's performance for this class.

The second model is Random Forest. I employed the Random Forest Classifier from the scikit-learn library, specifying a maximum depth of 10 for the decision trees and setting the random state to 69 to ensure reproducibility of results.

After fitting the model to the training data (X_train and y_train), I made predictions on the validation set (X_val) using the predict method. To evaluate the model's performance, I calculated the accuracy score using the metrics module, which measures the proportion of correctly predicted instances.

Additionally, I generated a classification report using the classification_report function, which provides insights into precision, recall, and F1-score for each class. Furthermore, I constructed a confusion matrix using the confusion_matrix function to visualize the model's predictions compared to the actual labels.

```
Accuracy of the model by using the random algorithm : 96.60%
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      1789
           1       0.98      0.96      0.97      1832

    accuracy                           0.97      3621
   macro avg       0.97      0.97      0.97      3621
weighted avg       0.97      0.97      0.97      3621
```

The model achieved an accuracy of 96.60%, indicating the proportion of correctly classified instances out of the total number of instances. The classification report provides further insights into the model's performance, with precision, recall, and F1-score metrics calculated for each class (seizure and non-seizure). For class 0 (non-seizure), the model achieved a precision of 96% and a recall of 98%, while for class 1 (seizure), it achieved a precision of 98% and a recall of 96%.

The last model is Naïve Bayes. I employed the Naive Bayes Classifier, a machine learning algorithm designed to predict the occurrence of seizures based on input features.

First, I imported the necessary modules from the sklearn library, including GaussianNB for the Naive Bayes algorithm, metrics for performance evaluation, and

classification_report for generating a detailed classification report. Subsequently, I instantiated the Gaussian Naive Bayes model and trained it on the training dataset (X_train and y_train).

After fitting the model, I made predictions on the validation dataset (X_val) and calculated the accuracy of the predictions using the accuracy_score function from the metrics module. The obtained accuracy with the Naive Bayes model was 91.71%.

Precision measures the accuracy of positive predictions. For class 0 (representing non-seizure instances), the precision is 0.87, indicating that among all instances predicted as non-seizure, 87% were correctly classified. For class 1 (representing seizure instances), the precision is 0.97, indicating that among all instances predicted as seizure, 97% were correctly classified.

```
Naive Bayes
Accuracy with naive is: 91.71%.
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.97      0.92      1789
           1       0.97      0.86      0.91      1832

    accuracy                           0.92      3621
   macro avg       0.92      0.92      0.92      3621
weighted avg       0.92      0.92      0.92      3621
```

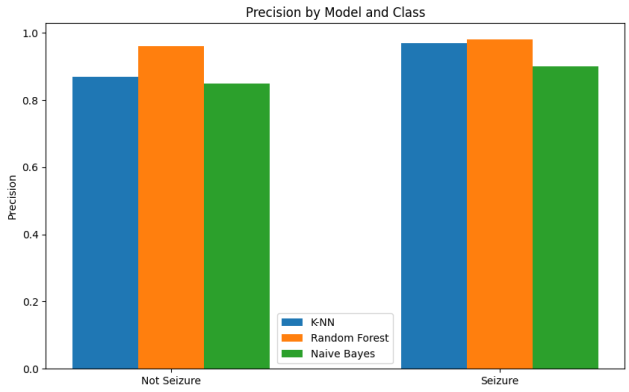
B. *Comparaison of results*

Upon completing the model training phase, I proceeded to assess its effectiveness using precision, recall, and F-measure metrics. Additionally, I utilized confusion matrices to gauge the model's performance across true positives, false positives, true negatives, and false negatives.

```
Accuracy with K-NN: 98.73%
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1789
           1       1.00      0.98      0.99      1832

    accuracy                           0.99      3621
   macro avg       0.99      0.99      0.99      3621
weighted avg       0.99      0.99      0.99      3621


Accuracy of the model by using the random algorithm : 96.60%
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      1789
           1       0.98      0.96      0.97      1832

    accuracy                           0.97      3621
   macro avg       0.97      0.97      0.97      3621
weighted avg       0.97      0.97      0.97      3621


Naive Bayes
Accuracy with naive is: 91.71%.
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.97      0.92      1789
           1       0.97      0.86      0.91      1832

    accuracy                           0.92      3621
   macro avg       0.92      0.92      0.92      3621
weighted avg       0.92      0.92      0.92      3621
```

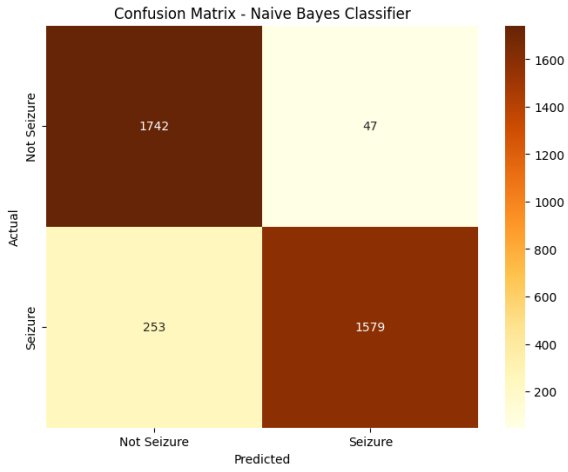
Furthermore, I conducted an evaluation of the models' performance, focusing on accuracy and precision metrics. A comprehensive summary of the results obtained from these evaluations is presented in the subsequent table.

The models, K-NN and Random Forest, are quite close in terms of accuracy, with the K-NN model slightly outperforming the Random Forest model by achieving an accuracy of 98.73% compared to 96.60%. However, the Naïve Bayes model trails slightly behind, with an accuracy of 91.71%.

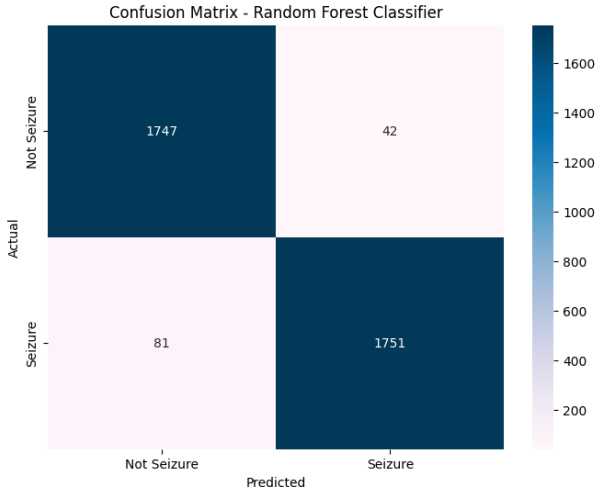


C. *Performance analysis*

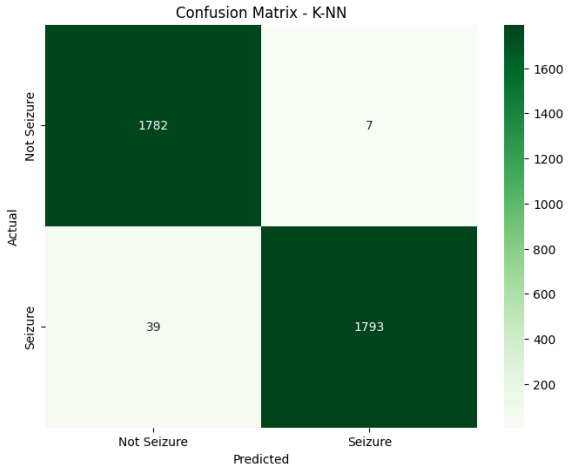
After meticulously evaluating the performance of my predictive models, I'm phappy about the results. My Naive Bayes classifier achieved an accuracy of 91.71%, demonstrating its ability to make accurate predictions. It displayed precision scores of 0.87 and 0.97 for the two classes, reflecting its capability to minimize false positives and false negatives. Moreover, its balanced F1-score of 0.92 for both classes highlights its consistency in performance.



Moving on to the Random Forest classifier, it showcased an even higher accuracy of 96.60%. My model exhibited precision scores of 0.96 and 0.98 for the respective classes, indicating its proficiency in making precise predictions. Additionally, its F1-score of 0.97 for both classes underscores its reliability in capturing the balance between precision and recall.



However, the most impressive performance came from my K-Nearest Neighbors (K-NN) algorithm, boasting an outstanding accuracy of 98.73%. With precision scores of 0.98 and 1.00 for the two classes, it demonstrated an exceptional level of precision in its predictions. This remarkable accuracy, coupled with a balanced F1-score of 0.99 for both classes, solidifies its position as the most suitable model for predicting epileptic seizures.



V. Discussion

Machine learning techniques offer promising avenues for predicting epileptic seizures, aiming to improve patient outcomes and quality of life. However, like any predictive model, there are inherent limitations that warrant consideration.

One limitation lies in the representativeness of the training data used for the model. The dataset utilized may not fully capture the diversity of epileptic seizure patterns across different populations. Thus, the model's effectiveness may vary when applied to populations with distinct demographic or clinical characteristics, potentially limiting its generalizability.

Another constraint pertains to the quality of the input data. While efforts were made to preprocess the data by addressing missing values and outliers, there remains a possibility of inaccuracies in data processing. Mishandling of missing data or outliers could introduce biases or distortions that influence the model's predictive performance.

Furthermore, employing multiple machine learning algorithms in the model presents both advantages and limitations. While the diversity of algorithms may enhance overall predictive performance, it also introduces complexity and potential conflicts in modeling approaches. Opting for a single algorithm might have mitigated these challenges and potentially yielded more streamlined and interpretable results.

Overall, acknowledging and addressing these limitations is crucial for refining the model's predictive accuracy and robustness. Future research efforts should focus on enhancing data representativeness, refining data preprocessing techniques, and optimizing model selection strategies to advance the field of epileptic seizure prediction.

VI. Conclusion

In conclusion, this article has underscored the transformative potential of artificial intelligence (AI) and machine learning (ML) in revolutionizing the prediction of epileptic seizures.

By leveraging sophisticated ML models, including k-nearest neighbors (K-NN), Random Forest (RF), and Naive Bayes classifiers, this study has demonstrated significant strides towards earlier detection and improved management of epilepsy.

Through the analysis of electroencephalographic (EEG) time series data from the CHB-MIT database, we have unveiled promising avenues for enhancing our understanding of seizure patterns and facilitating more precise predictive models.

Despite encountering limitations in data representativeness and algorithm selection, the findings highlight the critical role of AI in advancing epilepsy research and clinical practice. Moving forward, continued research efforts are warranted to address these limitations and further refine predictive models, ultimately paving the way for more effective interventions and improved quality of life for individuals living with epilepsy.

For access to the project code and further exploration, please visit our GitHub repository: https://github.com/salmaelkorchi/VTI

VII. References

[1] Athar A. Ein Shoka, Mohamed M. Dessouky, Ayman El-Sayed, Ezz El-Din Hemdan. "An efficient CNN based epileptic seizures detection framework using encrypted EEG signals for secure telemedicine applications." Faculty of Electronic Engineering, Computer Science and Engineering Department, Menoufia University, Egypt. Received 17 April 2022; revised 11 August 2022; accepted 2 October 2022.

[2] Ali Shoeb, John Guttag. "Application of Machine Learning To Epileptic Seizure Detection." Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts, 02139.

[3] Muhammad Haseeb Aslam, Syed Muhammad Usman, Shehzad Khalid, Aamir Anwar, Roobaea Alroobaea, Saddam Hussain, Jasem Almotiri, Syed Sajid Ullah, Amanullah Yasin. "Classification of EEG Signals for Prediction of Epileptic Seizures." Department of Computer Engineering, Bahria University, Islamabad 44000, Pakistan; Department of Creative Technologies, Faculty of Computing and Artificial Intelligence, Air University, Islamabad 44000, Pakistan; School of Computing and Engineering, The University of West London, London W5 5RF, UK; Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong, Seri Begawan BE1410, Brunei; Department of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway; Department of Electrical and Computer Engineering, Villanova University, Villanova, PA 19085, USA.

[4] Zhiheng Weng. "Prediction and Recognition of Epileptic Seizures Based on Artificial Intelligence." Sendelta International Academy, Shenzhen, China. Corresponding author: James.Zhiheng.Weng@student.sendelta.com.

[5] Muhammad Shoaib Farooq, Aimen Zulfiqar, Shamyla Riaz. "Epileptic Seizure Detection Using Machine Learning: Taxonomy, Opportunities, and Challenges." Department of Computer Science, University of Management and Technology, Lahore 54000, Pakistan. DOI: https://doi.org/10.3390/diagnostics13061058.

[6] Christine Rosquist Sandy Kang Lövgren. "Machine Learning Methods for EEG-based Epileptic Seizure Detection." DEGREE PROJECT IN TECHNOLOGY, FIRST CYCLE, 15 CREDITS, KTH ROYAL INSTITUTE OF TECHNOLOGY, SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, STOCKHOLM, SWEDEN 2019.

[7] Christine Rosquist Sandy Kang Lövgren. "Machine Learning Methods for EEG-based Epileptic Seizure Detection." DEGREE PROJECT IN TECHNOLOGY, FIRST CYCLE, 15 CREDITS, KTH ROYAL INSTITUTE OF TECHNOLOGY, SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, STOCKHOLM, SWEDEN 2019.

[8] Christine Rosquist Sandy Kang Lövgren. "Machine Learning Methods for EEG-based Epileptic Seizure Detection." DEGREE PROJECT IN TECHNOLOGY, FIRST CYCLE, 15 CREDITS, KTH ROYAL INSTITUTE OF TECHNOLOGY, SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, STOCKHOLM, SWEDEN 2019.

[9] Ralph G. Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, Christian E. Elger. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state." Department of Epileptology, University of Bonn, Sigmund-Freud-Strasse 25, 53105 Bonn, Germany; Institut für Strahlen- und Kernphysik, University of Bonn, Nußallee 14-16, 53115 Bonn, Germany. Received 14 May 2001; published 20 November 2001.

For access to the project code and further exploration, please visit our GitHub repository: https://github.com/salmaelkorchi/VTI