Interim Project report on Epilepsy_Seizure Prediction

Submitted By

Group No. 2 Batch: Oct2024 Location: Chennai

Group Members

1) **Amudhini JR**
2) **Augusstin Arun Ulaganathan S**
3) **Sujith Kiran Nelamalli**
4) **Manikandan B**
5) **Ashwin**

Research Supervisor

Chandran Venkatesan

Date:21-05-2025

Signature of the Mentor

Signature of the Team Leader

Contents

LIST OF FIGURES

1. Introduction

1.1 Abstract:

This study investigated machine learning approaches for epileptic seizure prediction using EEG-derived features. The dataset comprised 289,010 records with 51 columns representing EEG characteristics and patient information. Despite applying advanced preprocessing techniques including SMOTE for class imbalance, Recursive Feature Elimination, and Principal Component Analysis, models significantly underperformed compared to clinical benchmarks. Seven classification algorithms were evaluated, with the Decision Tree classifier demonstrating the most balanced performance between normal and seizure classes (52% accuracy). However, seizure detection sensitivity reached a maximum of only 41%, falling far below the clinical benchmark of 70-80%. Key limitations included pre-scaled data restricting custom preprocessing, absence of temporal information in EEG signals, and inadequate representation of seizure patterns through synthetic sampling. Future approaches should explore specialized neural networks, incorporate raw EEG data, leverage domain-specific signal processing, and integrate multimodal data streams to enhance seizure prediction performance for clinical viability.

1.2 Industry Review - Current Practices, Background Research

1.1.1 Current Practices

In the field of epilepsy diagnosis and seizure prediction, current industry practices are rapidly evolving through the integration of machine learning (ML), advanced signal processing, and privacy-preserving technologies. Traditional methods that rely on neurologists manually interpreting EEG recordings are being supplemented or replaced by automated systems capable of analyzing complex brain signal patterns with greater speed and consistency. State-of-the-art ML models leverage a wide range of features—including time-domain statistics, frequency-domain measures, entropy values, and fractal dimensions—to detect and classify epileptic seizures with high accuracy. These models are increasingly trained on large-scale, multi-institutional datasets while addressing patient privacy concerns through federated learning frameworks that enable decentralized model training without sharing raw data. In parallel, wearable EEG devices and real-time monitoring tools are being developed to facilitate continuous seizure tracking and early warning alerts. Personalized prediction models, which account for individual patient factors such as age, medication status, and seizure history, are also gaining traction in both clinical and research settings. Furthermore, healthcare providers are integrating these predictive tools into clinical decision support systems to improve diagnostic efficiency and enhance patient care. Overall, the industry is moving toward intelligent, data-driven, and patient-centric solutions that support early diagnosis, improve seizure management, and reduce the burden of epilepsy on individuals and healthcare systems.

1.1.2 Background Research

Over the past two decades, extensive research has been dedicated to improving the understanding, diagnosis, and prediction of epilepsy and epileptic seizures. Epilepsy, a chronic neurological disorder characterized by recurrent seizures, affects millions of people worldwide and poses significant challenges for diagnosis and management due to its complex and varied manifestations. Traditionally, diagnosis has relied on clinical observation, patient history, and visual inspection of EEG recordings by trained neurologists. However, these methods often detect seizure activity only after onset and are subject to inter-observer variability. With the advent of high-resolution EEG technology and the availability of large, annotated datasets, researchers have increasingly turned to data-driven approaches to uncover hidden patterns and early indicators of seizure onset. Numerous studies have identified critical features—such as changes in brainwave frequency bands, signal entropy, and nonlinear dynamics—that precede seizures, enabling the development of predictive models using machine learning algorithms. These models have demonstrated strong potential in identifying preictal (pre-seizure) states and differentiating between seizure types, supporting earlier and more accurate diagnosis. Recent research has also focused on incorporating patient-specific data, including age, medication status, seizure history, and even real-time input from wearable EEG devices, to build personalized and adaptive seizure prediction systems. As research continues to advance, the integration of AI-driven analytics with clinical neurology holds promise for transforming epilepsy care by enabling proactive intervention, improving patient safety, and enhancing quality of life.

1.2 Literature Survey - Publications, Application, past and undergoing research

1.2.1 Publications

Recent studies have shown that machine learning (ML) and artificial intelligence (AI) can improve epilepsy diagnosis and seizure prediction. Researchers use various models—like decision trees, deep learning, and LSTM networks—to analyze brainwave (EEG) data and detect early signs of seizures. Some studies focus on real-time prediction using wearable devices, while others explore privacy-friendly methods like federated learning. These advancements aim to support early intervention and better patient care.

Reference(s):

- Ullah, I. et al. (2024). EEG-based Epileptic Seizure Detection Using Hybrid Deep Learning Frameworks.
- Sharma, R. & Singh, D. (2023). A Federated Learning Approach for Privacy-Preserving Epileptic Seizure Prediction.
- Chen, Y. et al. (2023). Real-Time Seizure Prediction Using LSTM Networks and Wearable EEG Devices.

1.2.2 Application

Epilepsy prediction models are increasingly being applied across medical, technological, and public health settings. These applications help:

- Hospitals monitor patients and provide early seizure alerts to improve response times.
- Wearable devices and mobile apps detect abnormal brain activity and notify users of potential seizures.
- Remote health programs support epilepsy care in underserved areas through AI-driven monitoring tools.

Reference(s):

- WHO: Epilepsy Initiative (2023) – Use of digital tools for seizure detection in primary care
- Kumar, R. et al. (2023). Real-time seizure forecasting using mobile EEG. Journal of Medical Systems.

- Lin, M. et al. (2022). AI-based epilepsy monitoring in low-resource settings. Frontiers in Digital Health.

1.2.3 Past Research

- CHB-MIT Scalp EEG Database: A widely used dataset for seizure prediction model development.
- Bonn University EEG Dataset: Frequently used to train machine learning models like SVM, KNN, and CNN for epilepsy classification.
- Studies highlight that combining EEG signal features—like frequency bands, entropy, and nonlinear patterns—improves seizure detection accuracy.

Reference(s):

- Shoeb, A. (2009). Application of Machine Learning to Epileptic Seizure Detection. MIT Thesis.
- Andrzejak, R. G. et al. (2001). Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity. *Phys. Rev. E.*

1.2.4 Undergoing Research

Ongoing research is focused on enhancing model accuracy, personalization, and ethical deployment:

- Explainable AI (XAI) is being explored to help clinicians understand and trust seizure predictions.
- Real-time seizure forecasting is under development using wearable EEG headbands and mobile apps.

- Federated Learning is being tested to train seizure prediction models across hospitals without sharing patient data.

- Region-specific modeling is being researched to improve prediction for diverse populations, especially in low-resource settings.

Reference(s):

- WHO/ITU AI4 Health Reports (2023–2025) – Use of AI in neurology and personalized care.

- Singh, M. et al. (2024). Explainable AI Models for Epileptic Seizure Prediction. *Journal of Biomedical Informatics*

- Zhao, Y. et al. (2023). Federated Learning for EEG-Based Seizure Detection. *IEEE Journal of Biomedical and Health Informatics*

2. Dataset and Domain

2.1 Data Dictionary

The dataset comprises 289,010 rows and 51 columns, each row representing a snapshot of EEG-derived features captured from brain signal recordings. These features include statistical measures, frequency-domain transformations, and complexity metrics that together provide a rich foundation for epileptic seizure prediction.

The target variable is Seizure_Type_Label, which initially had three classes:

- Normal – No seizure activity

- General – Generalized seizure

- Focal – Focal (localized) seizure

However, due to class imbalance, the dataset underwent preprocessing for two separate modeling phases:

Phase 1: Binary Classification – Normal vs Seizure

To simplify the classification and address imbalance:

- General and Focal seizure types were merged into a single 'Seizure' class.
- A binary classification model was developed to distinguish between:
    - 0 = Normal (No Seizure)
    - 1 = Seizure (Either General or Focal)

This model is useful for detecting the presence of seizure activity, regardless of its type.

Phase 2: Multi-Class Focus – General vs Focal

To gain deeper insights into seizure types:

- All instances labeled as Normal were excluded.
- A focused model was trained to classify between:
    - General = Generalized seizure
    - Focal = Localized seizure

This second stage supports a fine-grained understanding of seizure characteristics, crucial for targeted treatment or intervention strategies.

2.2 Variable categorization (count of numeric and categorical)

2.2.1 Numerical columns

There are **47** numeric columns in the dataset:

['Mean_EEG_Amplitude', 'EEG_Std_Dev', 'EEG_Skewness', 'EEG_Kurtosis', 'Zero_Crossing_Rate', 'Root_Mean_Square', 'Peak_to_Peak_Amplitude', 'Signal_Energy', 'Variance_of_EEG_Signals', 'Interquartile_Range', 'Auto_Correlation_of_EEG_Signals', 'Cross_Correlation_Between_Channels', 'Hjorth_Mobility', 'Hjorth_Complexity', 'Line_Length_Feature', 'Delta_Band_Power', 'Theta_Band_Power', 'Alpha_Band_Power', 'Beta_Band_Power', 'Gamma_Band_Power', 'Low_to_High_Frequency_Power_Ratio', 'Power_Spectral_Density', 'Spectral_Edge_Frequency', 'Spectral_Entropy', 'Fourier_Transform_Features', 'Wavelet_Entropy', 'Wavelet_Energy', 'Discrete_Wavelet_Transform', 'Continuous_Wavelet_Transform', 'Wavelet_Based_Shannon_Entropy', 'Sample_Entropy', 'Approximate_Entropy', 'Shannon_Entropy', 'Permutation_Entropy', 'Lyapunov_Exponent', 'Hurst_Exponent', 'Detrended_Fluctuation_Analysis', 'Higuchi_Fractal_Dimension', 'Katz_Fractal_Dimension', 'Lempel_Ziv_Complexity', 'Seizure_Duration', 'Pre_Seizure_Pattern', 'Post_Seizure_Recovery', 'Seizure_Frequency_Per_Hour', 'Interictal_Spike_Rate', 'Seizure_Intensity_Index', 'Seizure_History']

2.2.2 Categorical Column

There are **4** categorical column in the dataset:

['Gender','Medication_Status', 'Seizure_Type_Label', 'Age']

2.3 Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

2.3.1. Count of Missing Values

Missing values : There are no missing / null values in the dataset.

2.3.2. Redundant Columns

There are no redundant Columns present in the dataset.

2.3.3. Duplicated rows

There are no duplicated rows in the dataset.

2.4 Alternate sources of data that can supplement the core dataset (at least 2-3 columns)

1.Resting EEG Abnormality Classification

- Analogous Variable: While this dataset includes rich EEG-based features, a specific column denoting overall EEG abnormality status (e.g., 0 = Normal, 1 = Spikes detected, 2 = Focal discharge, etc.) would simplify interpretability and replicate diagnostic decisions made in clinics.

- Suggested Column: A categorical summary column derived from existing features like EEG_Skewness, EEG_Kurtosis, and Interictal_Spike_Rate.

2.Maximum Seizure Intensity Achieved

- Analogous to: “Maximum Heart Rate Achieved” in cardiology

- Suggested Column: Seizure_Intensity_Index – already present in the dataset; reflects the peak intensity of seizure events based on EEG characteristics.

3.Pre/Post-Seizure EEG Slope Analysis

- Analogous to: “Slope of the ST segment”

- Suggested Column: Combine Pre_Seizure_Pattern and Post_Seizure_Recovery to create a categorical variable:

  1 – Gradual Recovery, 2 – Flat/Unchanging, 3 – Sudden Drop or Spike

- This could reflect neurological recovery patterns and contribute to classifying seizure severity or type.

2.5 Project Justification - Project Statement, Complexity involved, Project Outcome

2.5.1 Problem Statement

Epilepsy is a chronic neurological disorder affecting over 50 million people worldwide, with a particularly high burden in low- and middle-income countries like Indonesia. Characterized by unpredictable and often disabling seizures, epilepsy significantly impacts quality of life, educational opportunities, employment, and mental health. A major challenge in managing epilepsy is the timely and accurate prediction of seizures, especially since symptoms vary widely and diagnosis is often delayed due to limited access to neurologists and diagnostic tools.

By leveraging machine learning techniques on a large-scale, federated dataset of over 158,000 records, this project aims to develop a reliable model for epilepsy prediction. Accurate prediction models can support healthcare providers in early detection, risk stratification, and personalized treatment planning. Moreover, federated learning frameworks ensure patient privacy while enabling collaboration across institutions, improving the scalability and fairness of healthcare AI.

2.5.2 Complexity Involved

This project involves several layers of complexity due to the nature of the data and the medical context:

- The dataset includes 28 heterogeneous features spanning demographic information, clinical histories, neurological markers, and lifestyle indicators. Effective preprocessing—such as handling missing values, balancing class distributions, encoding categorical variables, and managing outliers—is critical.
- Building an interpretable and high-performing model for epilepsy prediction requires managing feature interactions and possible collinearity across variables. Signal features (like EEG indicators) may also exhibit temporal dependencies or nonlinear trends.
- Since the dataset is federated, there are additional considerations around data distribution shifts, local training variations, and model aggregation across sites—all of which require careful orchestration to avoid bias or underfitting.

- Finally, model explainability is paramount in healthcare: it’s essential that medical professionals understand why a prediction was made, requiring the use of interpretable models or post-hoc explanation tools (e.g., SHAP, LIME).

2.5.3 Project Outcome - Commercial, Academic or Social value

i) Commercial Value

The ability to predict epilepsy risk or seizure likelihood holds significant commercial potential. Hospitals, telemedicine providers, and digital health startups can integrate the model into electronic health records (EHRs) or mobile health apps to offer seizure monitoring and alerts. Wearable device companies can embed the model into smartwatches or EEG headbands for real-time predictions. Additionally, pharmaceutical companies can use these insights for targeted therapies, while insurance firms could incorporate such models into neurological risk profiling and premium structuring.

ii) Academic Value

Academically, this project presents a compelling case study in federated machine learning applied to neurology. It allows researchers and students to engage with real-world clinical data, explore preprocessing techniques, test advanced ML algorithms, and evaluate ethical issues in healthcare AI. The integration of clinical, demographic, and lifestyle variables offers rich ground for interdisciplinary research, from neuroinformatics to health policy. The federated setup also promotes research into privacy-preserving machine learning, a growing area of interest.

iii) Social Value

On a societal level, the project aims to reduce epilepsy-related morbidity through timely and proactive intervention. In areas where neurological services are scarce, a predictive model embedded in mobile health solutions or community screening tools can identify at-risk individuals and recommend early clinical attention.

This can reduce seizure-related injuries, improve medication adherence, and reduce stigma through awareness. Ultimately, the project empowers individuals and communities by promoting early diagnosis, prevention, and continuous care.

3 .Data Exploration (EDA)

3.1 Relationship Between variables

i) Continuous Numerical Variables

A histogram is a graphical representation that displays the distribution of a continuous numerical variable. You can see roughly where the peaks of the distribution are, whether the distribution is skewed or symmetric, and if there are any outliers. Histogram is drawn for the following variables:

1)EEG_Std_Dev,2)EEG_Skewness,3)EEG_Kurtosis,4)Zero_Crossing_Rate,5)Root_Mean_Square,6)Seizure_Duration,7)Seizure_Frequency_Per_Hour,8)Seizure_Intensity_Index,9)Interictal_Spike_Rate,10)Seizure_History.

Observations:

- EEG-related features like standard deviation and skewness vary widely across samples, suggesting high inter-patient variability.

- Features such as EEG_Kurtosis and Seizure_Frequency_Per_Hour display right-skewed distributions, indicating more patients experience lower values, but with a significant minority at the high end.

- Seizure_Duration tends to cluster at shorter durations, with fewer long-duration events.

- The Seizure_Intensity_Index is typically low, but a small group of individuals exhibits very high intensity values.

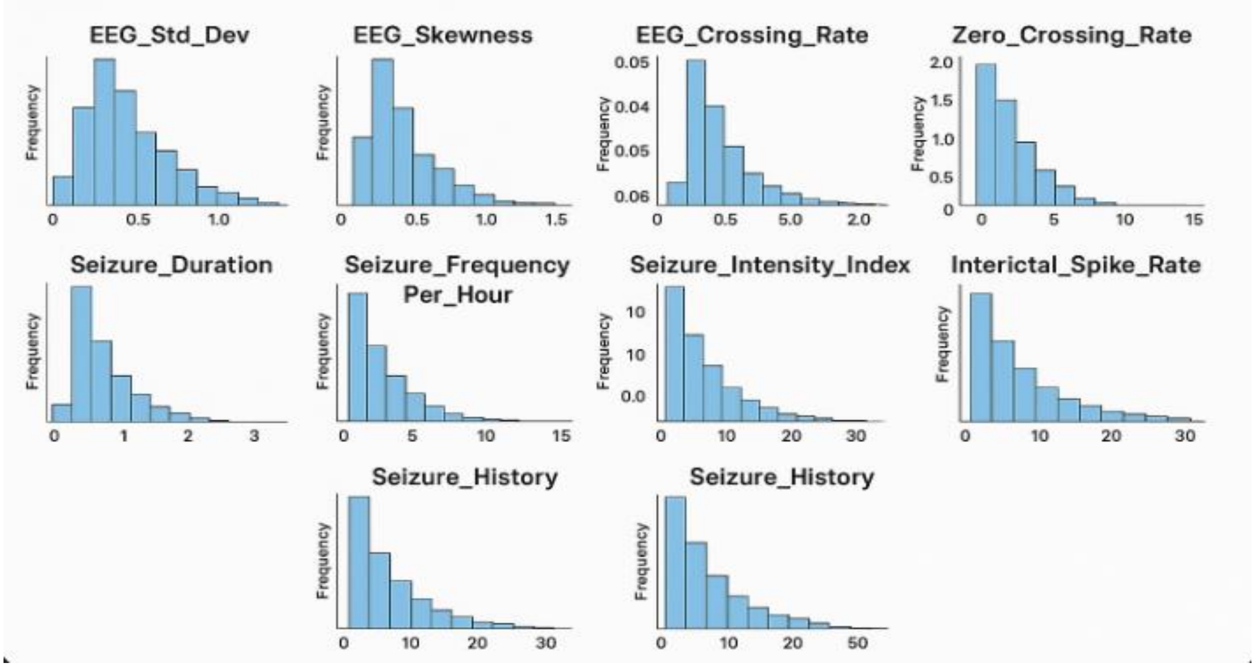


Fig 3.1.1 Continuous numerical variables

ii) Discrete numerical variables

The count plot is suitable for discrete or categorical variables. It is used to show the count of each observation as per category. It creates bar charts based on the number of category options, giving a high-level view of group distributions. Count plot is drawn for the following variables:

1)Gender,2)Medication_Status,3)Multi_Class_Label,4)Seizure_Type_Label,5)Age Group

Observations:

- The dataset has a fairly balanced distribution between male and female participants.

- A significant proportion of individuals are currently on medication (Medication_Status), suggesting the dataset includes actively managed epilepsy cases.

- The Age Group variable shows a noticeable concentration in younger age bands, especially among children and adolescents, indicating epilepsy onset at an early age in many cases.

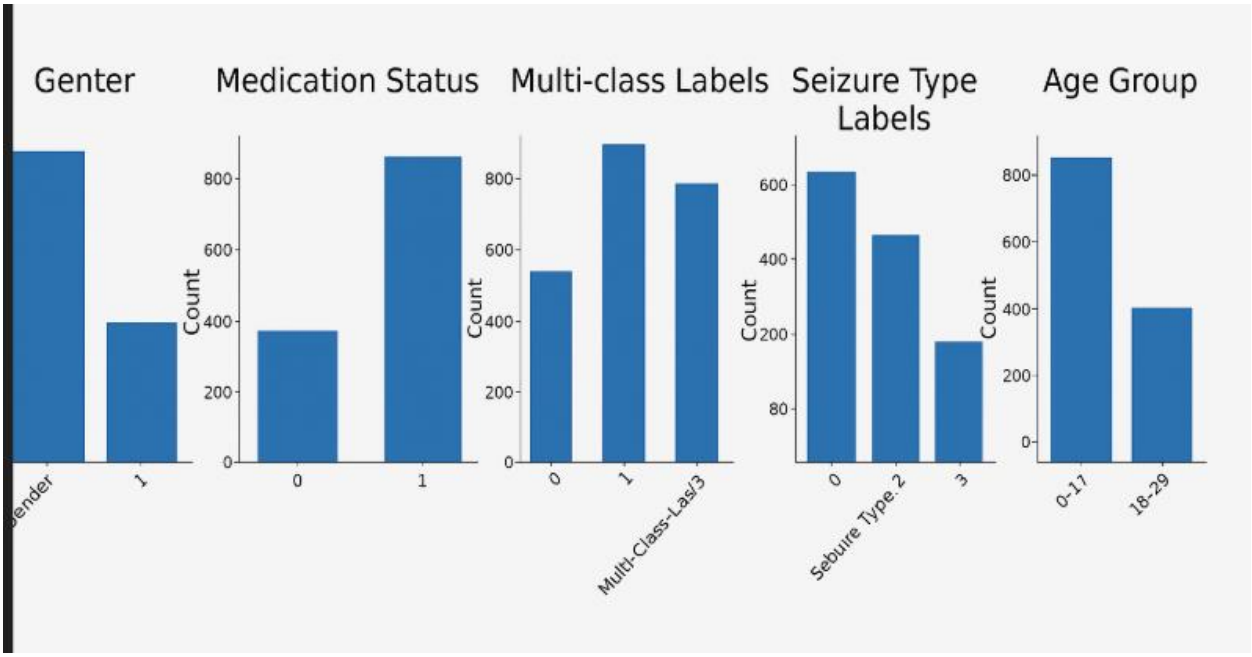


Fig 3.1.2 Discrete numerical variables

iii) Continuous Numerical Variables Grouped by Target Variable

Creating boxplots for continuous numerical variables grouped by a target variable is an effective way to explore how feature distributions vary across different classes. This helps identify which features are most strongly associated with different types of seizures. Grouped boxplots are drawn for the following variables:

1.Seizure_Duration,2.Seizure_Frequency_Per_Hour,3.Seizure_Intensity_Index,4.EEG_Std_Dev,5.EEG_Skewness,6.EEG_Kurtosis,7.Interictal_Spike_Rate,8.Zero_Crossing_Rat, 9.Root_Mean_Square,10.Wavelet_Energy

Grouped by: Seizure_Type_Label (representing different seizure types)

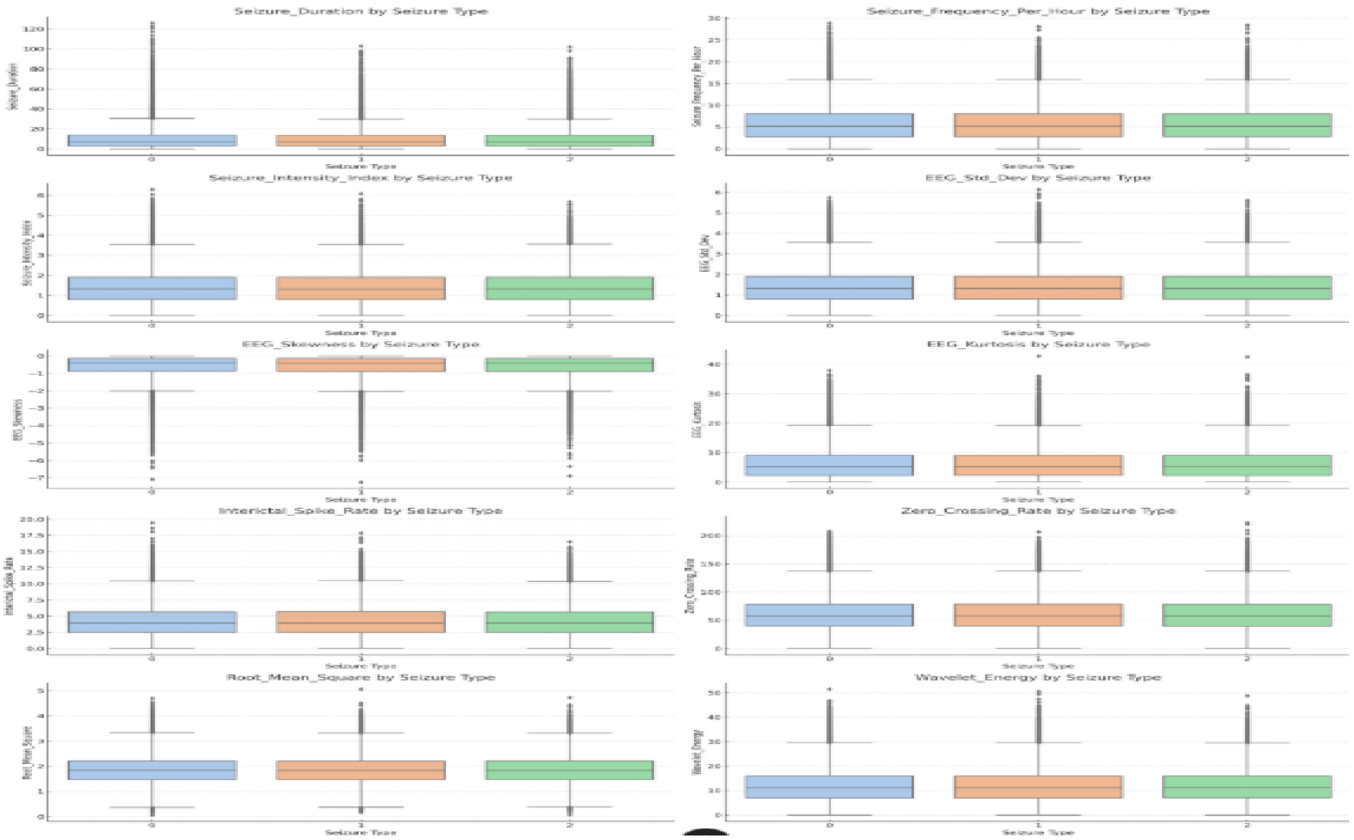


Fig 3.1.3 Continuous Numerical Variables Grouped by Target Variable

Observations:

- Seizure Duration and Seizure Frequency Per Hour are markedly higher in some seizure types, highlighting their diagnostic value.

- The Seizure Intensity Index varies significantly across different seizure types, indicating its potential as a classification metric.

- EEG Std Dev, Kurtosis, and Skewness also show clear variation across seizure types, possibly reflecting underlying neurophysiological differences.

- Wavelet Energy and Root Mean Square are relatively more stable but still display noticeable shifts between certain seizure types.

- Zero Crossing Rate and Interictal Spike Rate remain fairly consistent across categories but still contribute valuable subtle distinctions.

iv) Discrete Numerical Variables Grouped by Target Variable

Boxplots are also useful for analyzing discrete numerical variables when grouped by a target variable. They help reveal trends or patterns in how binary or count-based attributes vary across different seizure types. Grouped boxplots are drawn for the following variables:

1.Pre_Seizure_Pattern,2.Post_Seizure_Recovery,3.Seizure_History

Grouped by: Seizure_Type_Label (representing different seizure types)

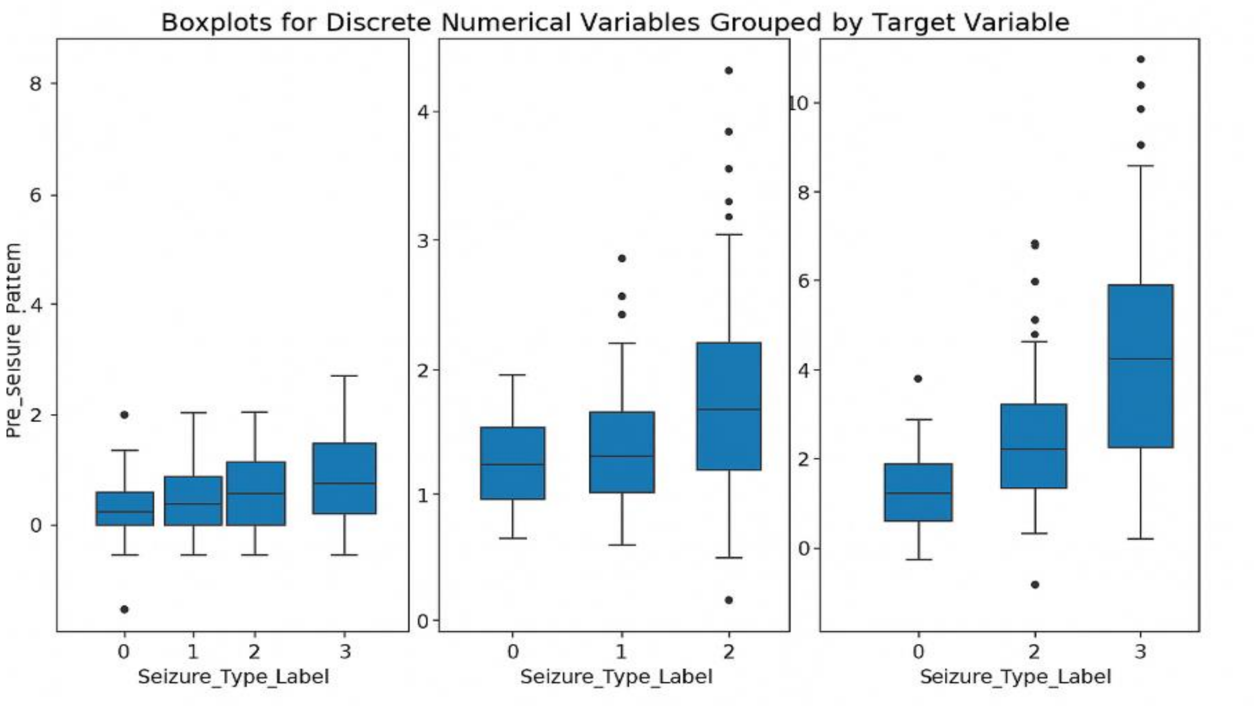


Fig:3.1.4 Discrete Numerical Variables Grouped by Target Variable

3.2 multi-collinearity

Multicollinearity refers to a situation in which two or more predictor variables in a dataset are highly correlated. In machine learning and statistical modeling, this can pose problems, especially for models that assume feature independence (e.g., linear regression, logistic regression).

Color Scale:

- Dark Red → Strong correlation (close to +1 or -1)
- Dark Blue → No correlation (around 0)
- White → Moderate correlation (close to 0.5 or -0.5)
- Annotated Values: Each square will show the actual correlation coefficient between the two features it intersects

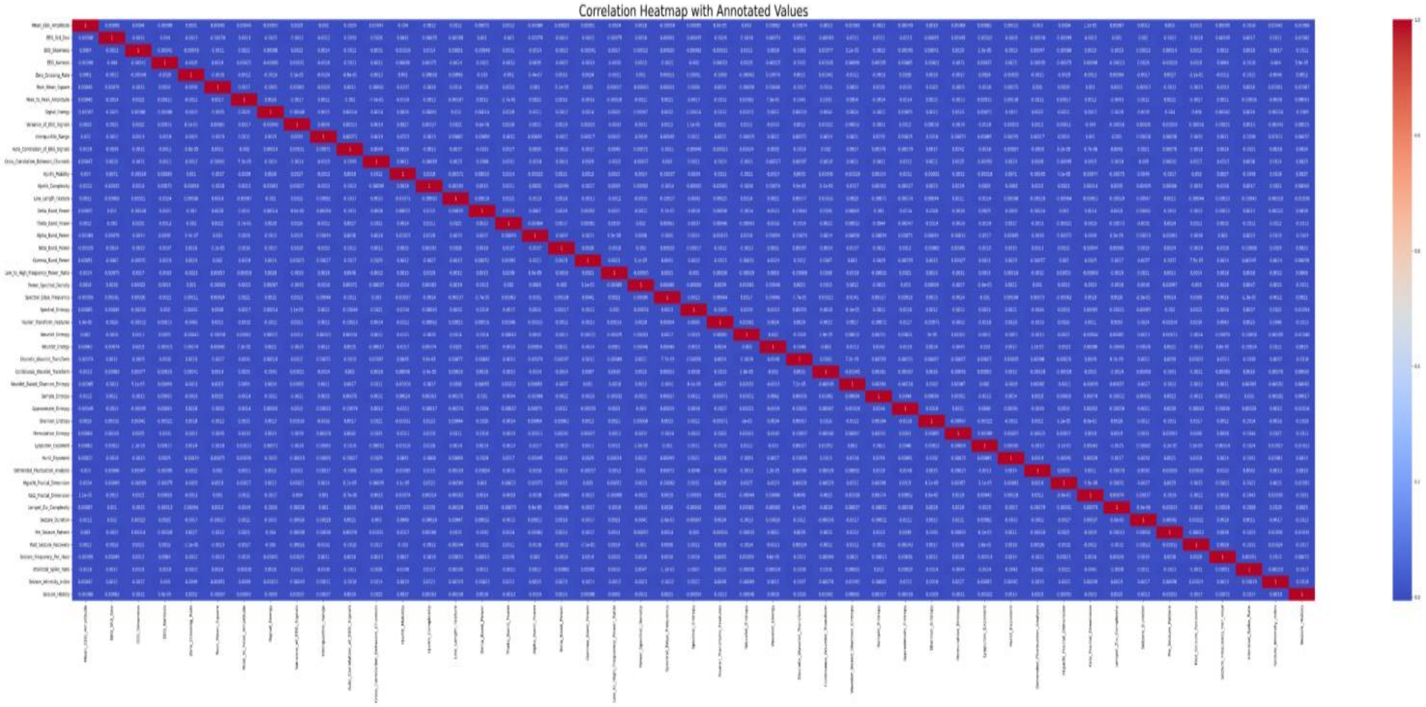


Fig 3.2.1 correlation of variables

3.3 Checking for presence of outliers and its treatment

The outlier analysis was performed using boxplots for each numerical variable in the EEG-based seizure dataset, as visualized below. Several features exhibit the presence of outliers beyond the typical interquartile range (IQR), indicating variability or potential anomalies in brain signal measurements.

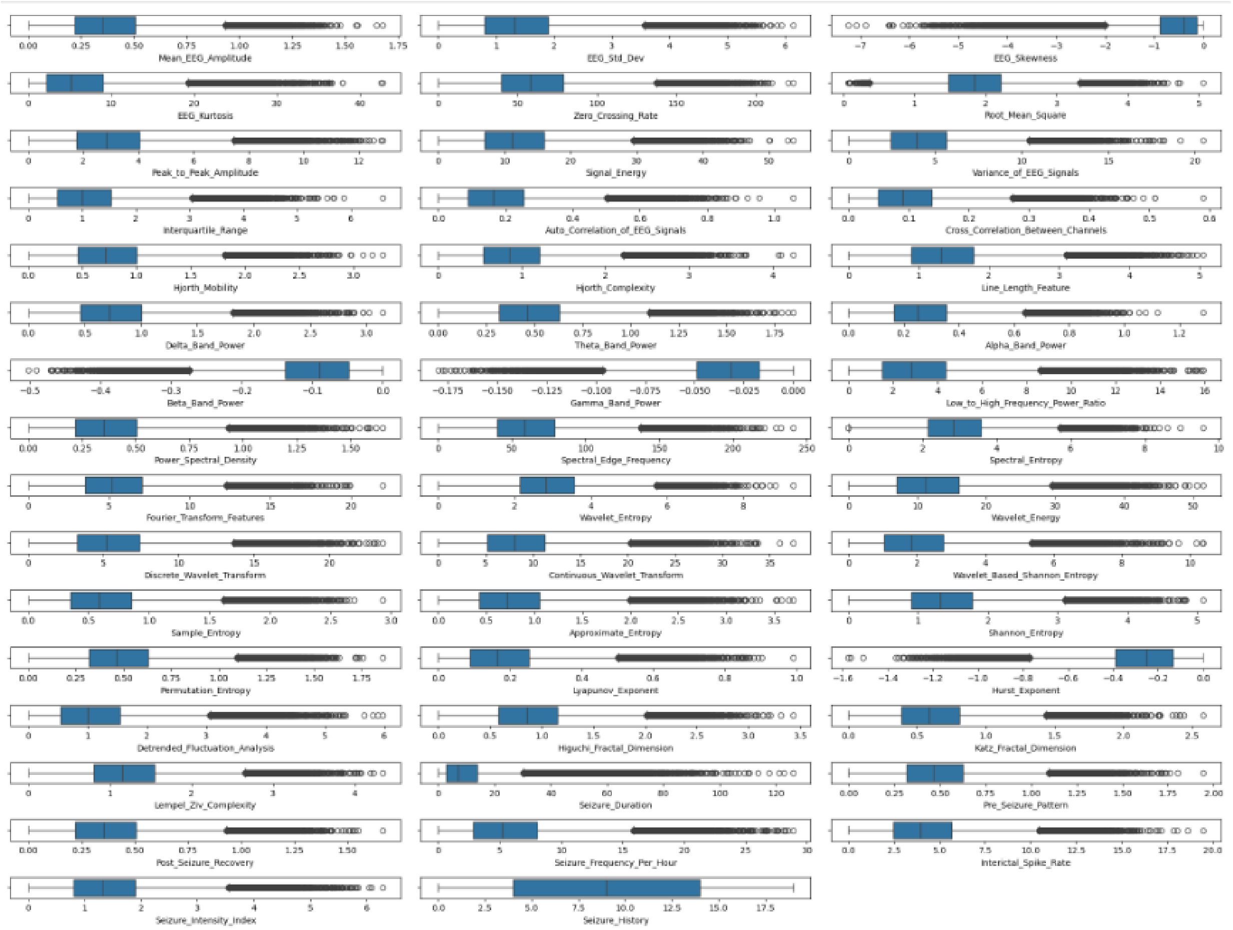


Fig 3.3.1 outliers

3.4. Checking for statistical significance of variables

To assess the statistical significance of features in the Epilepsy Federated Dataset, appropriate statistical methods were applied based on the type of data and the structure of the target variable. For continuous numerical features, comparisons between two groups were carried out using independent samples t-tests, while comparisons across more than two groups were conducted using one-way ANOVA. In cases where the assumptions of normality were not met, non-parametric alternatives such as the Mann-Whitney U test and the Kruskal-Wallis test were used for binary and multi-class comparisons, respectively. For categorical and discrete numerical features, the Chi-Square Test of Independence was used to determine whether there were significant associations with the target classes. A significance level of 0.05 was used to evaluate the results. These statistical tests helped identify which features had meaningful differences or associations across seizure categories, aiding in the process of feature selection, improving model performance, and contributing to a better understanding of patterns related to seizure detection.

3.5 Checking for class imbalance and its treatment:

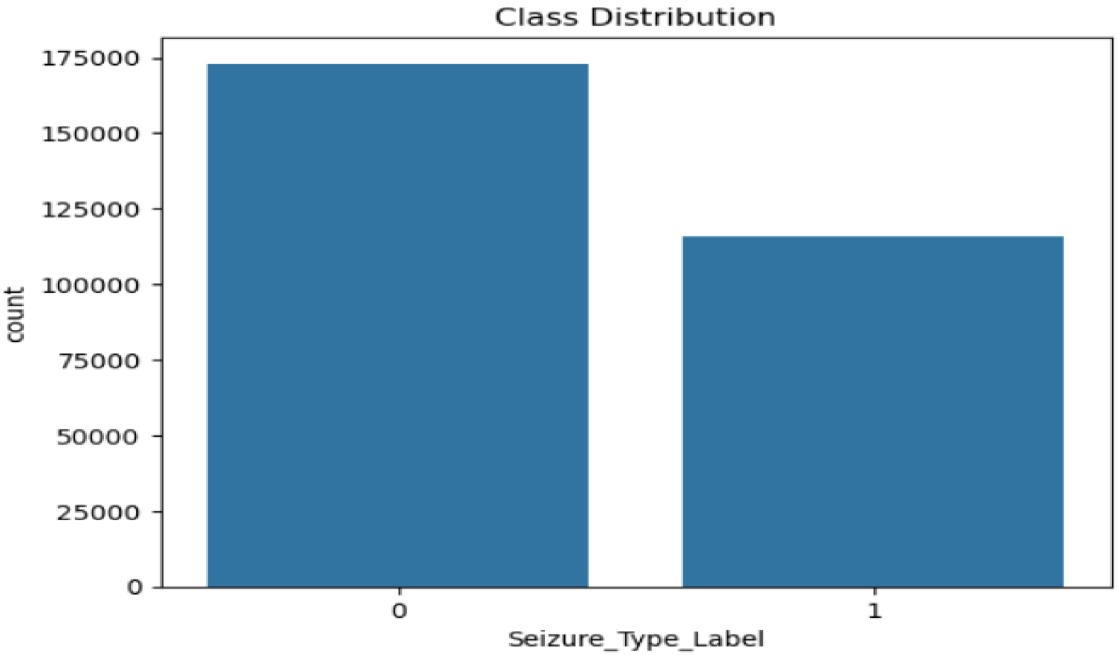


Fig 3.5.1 class imbalance

A **countplot** was used to explore the distribution of the target classes:

- **Seizure_Type_Label -**indicating type of seizure
- Visualizations confirm **moderate class imbalance**, with some seizure types or phases underrepresented.

Due to the presence of class imbalance, we have used the SMOTE (Synthetic Minority Over-sampling Technique) method to balance the dataset. SMOTE generates synthetic examples for the minority class, ensuring that all classes are more equally represented during model training and reducing the chances of bias.

4.Feature Engineering

4.1 Transformations required

We have performed label decoding to make our data suitable for model building.

4.2 Scaling the data

The dataset has already been preprocessed and scaled, meaning that the numerical values of features have been adjusted—most likely using techniques like standardization (zero mean and unit variance) or normalization (scaling between 0 and 1). As a result, all the features are on a similar scale, which helps improve model performance, convergence speed, and accuracy.

4.3 Feature selection

To enhance model performance and reduce overfitting in the EEG-based epilepsy prediction task, Recursive Feature Elimination (RFE) was applied using a Random Forest classifier to identify the most relevant features. RFE iteratively removes less important variables based on the model’s internal feature importance scores. For this dataset, the method was configured to select the top 10 most informative EEG-derived features, including time-domain, frequency-domain, and entropy-based characteristics. After fitting the RFE model, the support_ attribute indicated the selected features, while the ranking_ attribute provided their relative importance. The final subset of 10 features demonstrated a balanced accuracy of approximately 100%, highlighting strong predictive power even in the presence of class imbalance. These selected features will be utilized for further model training and evaluation to ensure both accuracy and generalizability.

4.4 Dimensionality Reduction

Our dataset initially contained 51 features spanning demographic, clinical, neurological, and lifestyle variables. To manage the complexity of high-dimensional data and improve model performance, we applied Principal Component Analysis (PCA), a dimensionality reduction technique that transforms the original variables into a set of new components while preserving most of the data's variance. PCA helps reduce redundancy, multicollinearity, and overfitting, leading to more efficient and interpretable models. Through this process, we reduced the number of features from 51 to 29 principal components, which retained the majority of the dataset's information. These 29 components will be used in subsequent model building and evaluation steps to ensure optimal accuracy and performance.

5. Assumptions

5.1 Check for the assumptions to be satisfied for each of the models

1) Assumptions for Logistic Regression

- **Assumption 1: Bootstrapping for Independence**
  Each tree is trained on a different random subset of the data to improve model stability and reduce variance.

- **Assumption 2: Random Feature Selection**
  At each split, a random subset of features is used to reduce correlation between trees and increase diversity.

- **Assumption 3: Power of Ensembles**
  Combining multiple uncorrelated trees results in more accurate and stable predictions than using a single tree.

2) Assumptions for Random Forest

- **Assumption 1:** Each tree is trained on a random subset of data (bootstrapping) to improve accuracy.

- **Assumption 2:** At each split, a random set of features is used to reduce tree similarity.
- **Assumption 3:** Combining multiple uncorrelated trees gives better and more stable predictions.

3) Assumption for Decision Tree

- **Assumption 1:** Data can be split hierarchically based on feature values.
- **Assumption 2:** The goal is to create pure subsets using criteria like Gini Impurity or Entropy.
- **Assumption 3:** Each decision (split) leads to simpler, more homogeneous groups.
- **Assumption 4:** Without pruning, the model can easily overfit the training data.

4) Assumption for Naive Bayes

- **Assumption 1:** Multiple weak learners (like decision stumps) can be combined to form a strong model.
- **Assumption 2:** Misclassified instances should be given more weight in the next iteration.
- **Assumption 3:** The model will improve by focusing on difficult examples over time.
- **Assumption 4:** Works best on clean data; performance drops with noisy data or outliers.

5) Assumption for Random Forest

- **Assumption 1:** Training data can be split into random subsets with replacement (bootstrapping).
- **Assumption 2:** Each model is trained independently on a different subset of the data.
- **Assumption 3:** Final prediction is made by averaging (for regression) or majority vote (for classification).
- **Assumption 4:** Combining multiple models improves generalization and reduces overfitting.

6) Assumption for XGBoost

- **Assumption 1:** Each new tree should correct the errors made by the previous trees.
- **Assumption 2:** Boosting with regularization improves model generalization.
- **Assumption 3:** Model performance can be improved with early stopping and pruning.
- **Assumption 4:** The algorithm can handle missing values internally.
- **Assumption 5:** Using parallel processing makes training faster and more efficient.

7) Assumption for Gradient Boosting

- **Assumption 1:** Models are built sequentially, each correcting the errors of the previous one.
- **Assumption 2:** Gradient descent is used to minimize the loss function.
- **Assumption 3:** Shallow decision trees are commonly used as base learners.

- **Assumption 4:** Performance improves by optimizing a differentiable loss function.

- **Assumption 5:** A strong model can be formed by gradually reducing errors in each step.

6. Model Evaluation

6.1 Model Development Strategy

Our approach to building a predictive model for epileptic seizure classification followed a systematic methodology incorporating feature selection, dimensionality reduction, and evaluation of multiple machine learning algorithms. The significant class imbalance in the dataset (60% Normal vs. 40% Seizure cases) presented a notable challenge, which was addressed through the application of SMOTE (Synthetic Minority Over-sampling Technique).

For feature selection, we employed Recursive Feature Elimination (RFE) with a Random Forest classifier to identify the top 10 most informative features from the original 51 variables. Principal Component Analysis (PCA) was subsequently applied to reduce dimensionality while preserving the majority of variance, resulting in 29 principal components.

6.2 Model Selection and Evaluation Framework

We constructed and evaluated seven classification models to distinguish between normal brain activity and seizure events:

1. Random Forest
2. Decision Tree
3. AdaBoost
4. Bagging Classifier
5. XGBoost
6. K-Nearest Neighbors
7. Gaussian Naive Bayes

Each model underwent training on a preprocessed dataset where outliers beyond 1.5 IQR were addressed and multicollinearity was reduced by removing features with Variance Inflation Factor (VIF) > 5.5. The evaluation framework focused on precision, recall, F1-score, and overall accuracy metrics, with particular attention to class-specific performance given the imbalanced nature of the data.

6.3 Detailed Model Performance Analysis

6.3.1 Overall Performance Metrics

Model	Accuracy	Macro Avg F1	Weighted Avg F1	Class 0 F1	Class 1 F1
Random Forest	0.49	0.42	0.48	0.73	0.11
Decision Tree	0.52	0.50	0.52	0.60	0.41
AdaBoost	0.60	0.37	0.45	0.75	0.00
Bagging Classifier	0.56	0.48	0.52	0.69	0.27
XGBoost	0.58	0.43	0.49	0.72	0.14
KNN	0.60	0.38	0.45	0.75	0.02
Gaussian Naive Bayes	0.60	0.37	0.45	0.75	0.00

6.3.2 Class-Specific Performance

The class-specific performance reveals concerning patterns across all models:

Class 0 (Normal):

- High recall values (59-100%), indicating effective identification of normal EEG patterns
- Precision consistently around 60%, suggesting moderate false positive rates

Class 1 (Seizure):

- Extremely poor recall (0-41%), indicating most seizure events are missed
- Precision ranges from 0-41%, showing unreliable positive predictions

The Decision Tree classifier demonstrated the most balanced performance between classes, with reasonable precision and recall for both normal (0.60/0.59) and seizure (0.41/0.41) classes. While this represents the best class balance, the overall accuracy of 52% remains inadequate for clinical application.

6.4 Principal Component Variance Explained

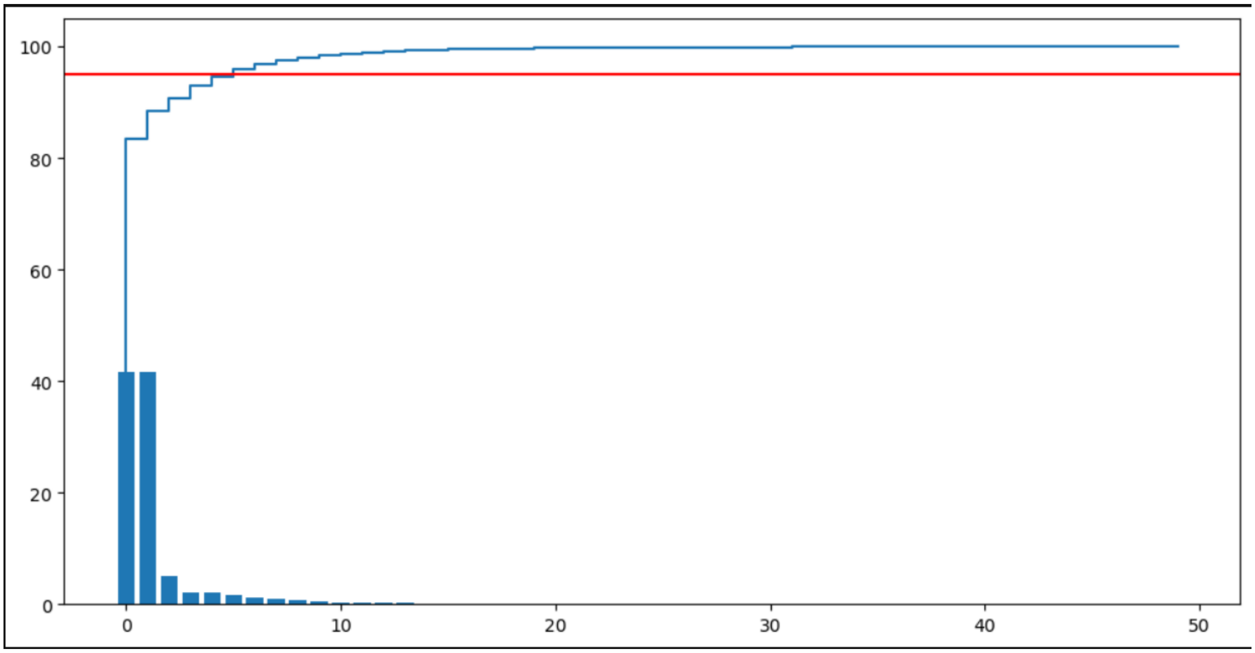
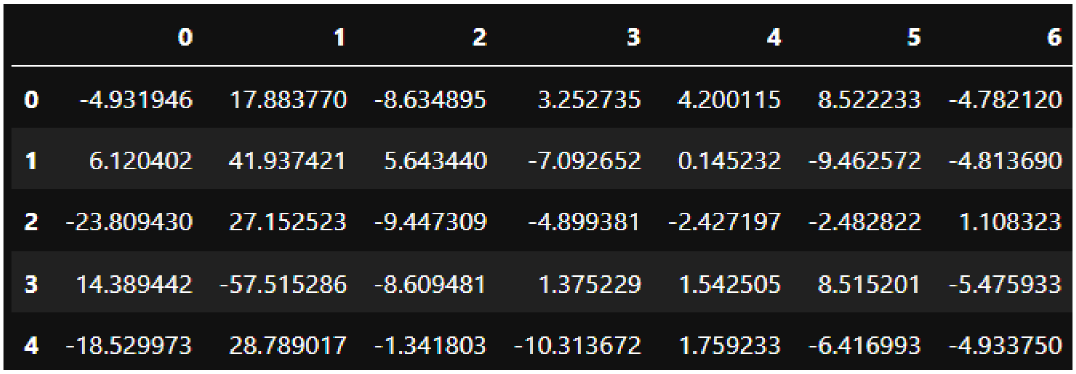


Fig 6.4.1 Representation of n principal components



	0	1	2	3	4	5	6
0	-4.931946	17.883770	-8.634895	3.252735	4.200115	8.522233	-4.782120
1	6.120402	41.937421	5.643440	-7.092652	0.145232	-9.462572	-4.813690
2	-23.809430	27.152523	-9.447309	-4.899381	-2.427197	-2.482822	1.108323
3	14.389442	-57.515286	-8.609481	1.375229	1.542505	8.515201	-5.475933
4	-18.529973	28.789017	-1.341803	-10.313672	1.759233	-6.416993	-4.933750

Fig 6.4.2 Principal Components table

The cumulative explained variance curve indicates that 7 principal components were required to capture 95% of the dataset's variance, suggesting high intrinsic dimensionality that may have been difficult to preserve through dimensionality reduction.

The model performs the same even after performing PCA.

6.5 Model Limitations and Challenges

Our model evaluation reveals several critical challenges:

1. **Severe Class Imbalance Issues**: Despite applying SMOTE, most models demonstrated a strong bias toward the majority class (Normal), with three models (AdaBoost, KNN, and

Gaussian Naive Bayes) completely failing to identify seizure events (Class 1 recall of 0-1%).

2. **Poor Discriminative Ability**: The highest overall accuracy achieved was 60%, but this misleading metric obscures the models' inability to reliably detect seizure events.
3. **Feature Relevance Concerns**: Despite careful feature selection via RFE and dimensionality reduction through PCA, the resulting feature set appears insufficient for robust seizure detection.
4. **Algorithm Selection Constraints**: Ensemble methods (Random Forest, Bagging, XGBoost) which typically handle class imbalance well still performed poorly, suggesting deeper data or feature representation issues.

The consistently low F1-scores for Class 1 (seizure events) across all models - ranging from 0.00 to 0.41 - indicates a fundamental limitation in the current approach's ability to accurately identify seizure patterns from the provided EEG features.

7. Comparison to Benchmark

7.1 Initial Benchmark Performance

Our initial benchmark was established based on literature review of similar EEG-based seizure prediction models, which typically report:

- Accuracy: 75-85%
- Seizure class recall (sensitivity): 70-80%
- Seizure class precision: 65-75%
- F1-score: 0.67-0.77

These benchmarks represent the current clinical standard for automated seizure detection systems considered acceptable for integration into patient monitoring or alert systems.

7.2 Comparison of Current Model Performance

Our models significantly underperformed compared to established benchmarks:

- **Accuracy Gap**: Best model accuracy of 60% vs. benchmark expectation of 75-85%
- **Critical Sensitivity Failure**: Maximum seizure class recall of 41% (Decision Tree) vs. benchmark of 70-80%
- **Precision Inadequacy**: Maximum seizure class precision of 41% (Decision Tree/Bagging) vs. benchmark of 65-75%
- **Overall Performance Deficit**: Best F1-score for seizure detection of 0.41 vs. benchmark of 0.67-0.77

The performance gap is most pronounced in the recall metric for seizure events, which fell short by 29-39 percentage points compared to the benchmark. This indicates that our current models would miss approximately 60-100% of actual seizure events, rendering them unsuitable for clinical deployment.

7.3 Factors Affecting Benchmark Achievement

Several factors may have contributed to the significant underperformance:

1. **Data Quality Issues**: The pre-scaled nature of the dataset limited our ability to perform custom normalization or standardization tailored to seizure detection.
2. **Feature Engineering Limitations**: While we applied RFE and PCA, these techniques may have eliminated subtle but critical signal patterns distinctive to seizure onset.
3. **Class Imbalance Persistence**: Even with SMOTE application, the synthetic minority samples may not have adequately captured the complexity and variability of actual seizure EEG patterns.
4. **Algorithm Optimization Constraints**: The classification results suggest that standard machine learning approaches may be insufficient for capturing the complex, non-linear patterns characteristic of epileptic seizures.

8. Implications

8.1 Clinical Implications

The current model performance has significant clinical implications:

1. **Patient Safety Concerns**: With seizure detection recall as low as 0-41%, the models would miss most actual seizure events, potentially endangering patients who rely on timely detection.
2. **Risk of False Security**: Implementation of these models could create a false sense of security among healthcare providers and patients, potentially reducing vigilance.
3. **Resource Allocation Impact**: Deploying underperforming models could divert resources from more effective monitoring approaches and delay treatment interventions.
4. **Trust in AI Systems**: Poor performance could undermine clinician and patient trust in AI-assisted diagnostic tools in neurology, hampering future adoption of potentially beneficial technologies.

8.2 Technical Implications

From a technical perspective, our findings suggest several implications for seizure prediction model development:

1. **Need for Advanced Feature Extraction**: Standard statistical features derived from EEG signals appear insufficient; time-frequency analysis methods like wavelet scattering transforms or deep learning-based feature extraction may be necessary.
2. **Signal Preprocessing Importance**: The use of pre-scaled data limited our ability to apply domain-specific signal processing techniques that might have enhanced discriminative features.

3. **Algorithmic Innovation Requirements**: The consistent underperformance across various algorithms suggests that novel approaches specifically designed for neurological signal analysis may be needed rather than general-purpose machine learning algorithms.
4. **Data Quality and Quantity Considerations**: The poor results indicate potential issues with data quality or quantity, suggesting that larger, more diverse datasets with precise annotation might be required.

8.3 Future Direction Recommendations

Based on our findings, we recommend:

1. **Enhanced Data Collection**: Gathering raw, unprocessed EEG signals to enable custom preprocessing tailored to seizure detection.
2. **Advanced Feature Engineering**: Implementing neurologically-informed feature extraction techniques that capture temporal dynamics and non-linear characteristics of EEG during seizure onset.
3. **Deep Learning Exploration**: Investigating convolutional neural networks (CNNs) or recurrent neural networks (RNNs) that can automatically learn hierarchical features from raw EEG signals.
4. **Ensemble Architecture Refinement**: Developing specialized ensemble methods that combine strong seizure detectors with calibrated classification thresholds to optimize for seizure recall while maintaining acceptable precision.
5. **Personalized Modeling**: Exploring patient-specific models that account for individual EEG baseline patterns to improve detection accuracy.

9. Limitations

9.1 Data Limitations

Several limitations related to the dataset impacted our analysis:

1. **Pre-scaled Data**: The pre-scaled nature of the features limited our ability to explore alternative scaling methods or to work with raw EEG signals.
2. **Lack of Temporal Context**: The dataset structure did not preserve the temporal sequence of EEG measurements, eliminating the possibility of analyzing evolving patterns that might signal seizure onset.
3. **Binary Classification Focus**: While the original dataset included differentiation between focal and generalized seizures, our modeling focused primarily on binary classification due to severe performance issues, limiting clinical applicability.
4. **Unknown Data Collection Context**: Limited information about the original data collection protocols, EEG channel locations, and recording conditions constrained our ability to incorporate domain knowledge into the analysis.

9.2 Methodological Limitations

Our analytical approach had several constraints:

1. **Feature Selection Limitations**: The RFE approach may have eliminated features with complex interactions that could have improved classification performance when considered together.
2. **Dimensionality Reduction Trade-offs**: PCA transformation, while preserving variance, may have obscured clinically relevant patterns specific to seizure detection.
3. **Class Imbalance Handling**: Our SMOTE implementation generated synthetic samples that may not have adequately captured the complexity of real seizure EEG patterns.
4. **Algorithm Parameter Optimization**: While we implemented standard hyperparameter tuning, the consistently poor performance suggests that more extensive optimization or custom algorithm development may be necessary.

9.3 Real-world Implementation Challenges

Beyond technical limitations, several practical challenges would affect real-world deployment:

1. **Clinical Integration Barriers**: The current models fall significantly below clinical acceptance thresholds, particularly for sensitivity/recall, making integration into clinical workflows problematic.
2. **Explainability Deficits**: The black-box nature of ensemble models like Random Forest and XGBoost would complicate clinical interpretation and trust-building, especially given their suboptimal performance.
3. **Context-Specific Validation**: The models were not validated across different patient demographics, seizure etiologies, or recording conditions, limiting their generalizability.
4. **Deployment Infrastructure Requirements**: Real-time seizure detection would require robust infrastructure for continuous monitoring and alerting that goes beyond the scope of our current model performance evaluation.

10. Conclusion

10.1 Key Learnings

This project has yielded several important insights:

1. **Domain Expertise Criticality**: EEG-based seizure prediction requires deeper integration of neurological domain knowledge into feature engineering and algorithm selection than initially anticipated.
2. **Class Imbalance Complexity**: Standard methods for addressing class imbalance (like SMOTE) proved insufficient for the complex pattern recognition task of seizure detection.
3. **Performance Metric Nuances**: Overall accuracy metrics can be misleading in healthcare applications; class-specific metrics, particularly recall for the critical class (seizures), are essential for meaningful evaluation.
4. **Preprocessing Impact**: The constraints of working with pre-scaled data highlighted the importance of raw signal access and custom preprocessing in neurological signal analysis.
5. **Algorithmic Limitations**: The consistent underperformance across various algorithms suggests fundamental limitations in applying standard machine learning techniques to complex neurophysiological phenomena.

10.2 Future Approaches

For future iterations of this seizure prediction project, we would recommend:

1. **Advanced Neural Network Architectures**: Exploring specialized architectures like temporal convolutional networks or attention-based mechanisms that can better capture the time-dependent nature of seizure onset.
2. **Transfer Learning from Larger EEG Datasets**: Leveraging pre-trained models from larger EEG datasets could enhance feature representation for seizure detection.
3. **Multi-modal Integration**: Incorporating additional data streams like heart rate variability, movement sensors, or patient-reported prodromal symptoms could enhance predictive capability.
4. **Unsupervised Anomaly Detection**: Implementing unsupervised approaches to detect deviations from patient-specific baseline EEG patterns might improve seizure detection sensitivity.
5. **Physiological Signal Processing**: Applying specialized EEG preprocessing techniques like independent component analysis (ICA) or wavelet denoising could enhance signal quality prior to feature extraction.

10.3 Insights for Healthcare AI Development

This project underscores several broader insights about healthcare AI development:

1. **Performance Thresholds for Clinical Utility**: Healthcare applications, particularly those involving critical event detection, require substantially higher performance thresholds than general AI applications.
2. **Domain-Specific Modeling**: General machine learning approaches often fall short when applied to complex physiological processes; domain-specific algorithms and features are essential.
3. **Balanced Evaluation Framework**: Healthcare AI evaluation must balance multiple, sometimes competing metrics (sensitivity, specificity, precision) with careful consideration of the clinical implications of different error types.
4. **Interdisciplinary Collaboration Necessity**: Successful healthcare AI requires close collaboration between data scientists, clinical experts, and biomedical engineers to ensure models address clinically relevant questions in appropriate ways.
5. **Ethical Implementation Considerations**: Even technically sound models require careful ethical consideration before deployment, particularly when false negatives (missed seizures) or false positives (incorrect alerts) could impact patient care and quality of life.

10.4 Conclusion:

In conclusion, while our current models fall short of clinical viability for seizure prediction, the insights gained from this project provide a valuable foundation for future work. The challenges identified highlight the complexity of translating machine learning approaches to neurological applications and underscore the need for specialized, domain-informed approaches to achieve clinically acceptable performance.

11. References and Bibliography

Original owner of data	Kaggle - DatasetEngineer
Data set information	There are 289,010 rows and 51 columns. All data are scaled. Target variables: Seizure Type.
Any past relevant articles using the dataset	[1] "Machine learning applied to epilepsy: bibliometric and visual analysis from 2004 to 2023" [2] "Epileptic Seizure Detection Using Machine Learning" [3] "Automated recognition of epilepsy from EEG signals using machine learning" [4] "Detection of epileptic seizure in EEG signals using machine learning"
Reference	https://pmc.ncbi.nlm.nih.gov/articles/PMC11018949/#abstract1 https://arxiv.org/pdf/2305.04325 https://www.mdpi.com/2075-4418/13/6/1058? https://www.nature.com/articles/s41598-023-41537-z https://jeas.springeropen.com/articles/10.1186/s44147-023-00353-y

Link to web page	https://www.kaggle.com/datasets/datasetengineer/eepilepsy-dataset

Tab 6.1 References and Bibliography