Using ML Classifiers and Preprocessing Techniques to Diagnose Epileptic Seizures

Lycia Tran

Introduction

Epilepsy is a disorder that causes repeated and unprovoked seizures and is the fourth most common neurological disorder affecting about 3.4 million people in the US alone. Electroencephalograms (EEGs) record the electrical activity in one's brain and are the most common method used to diagnose epilepsy. Currently, trained clinicians are needed to read and evaluate EEG signals to diagnose epileptic seizures. This is both inefficient and costly, as these clinicians need to be paid and the time it takes to read and diagnose EEG signals can be quite long. Therefore, the goal of this project is to use Machine Learning (ML) classifiers and preprocessing techniques on EEG data to diagnose epileptic seizures.

Background: Dataset

The dataset used was the Epileptic Seizure Recognition dataset from Kaggle. The dataset consists of 23.6 s EEG signal recordings sampled into 4096 datapoints. Each of these recordings were split into 1 s windows resulting in 11500 EEG signal recording windows. Of these 11500 signal recording windows, 2300 were from people who were experiencing an epileptic seizure and 9200 were from people who were not experiencing an epileptic seizure. The dataset came with 5 labels: seizure activity, EEG recorded at tumor site, EEG recorded in healthy brain area, eyes closed during recording, eyes open during recording. These labels were combined into a binary label, y, to represent non-seizure activity and seizure activity.

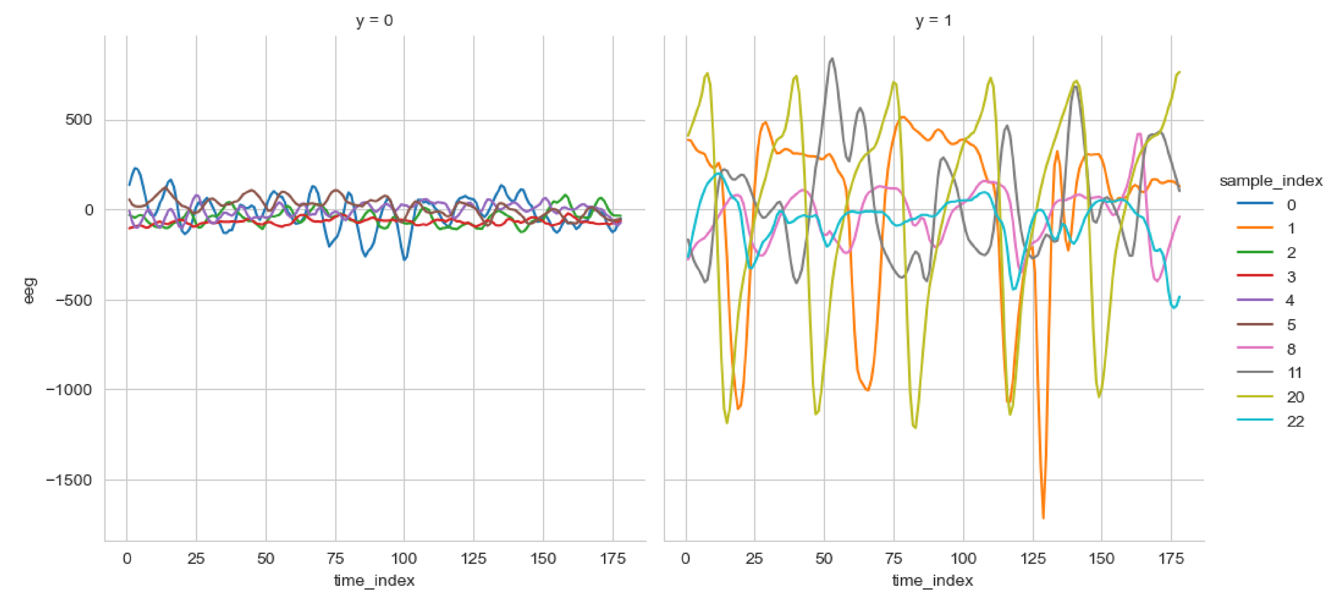

Figure 1. Samples of EEG Signal Recording from Dataset

Approach

The data was randomly split into an 80-20 training-test split. Both preprocessing techniques and ML classifiers were implemented and tested on the data.

The preprocessing techniques used were taking the Fourier Transform (FT) of the data and taking the Continuous Wavelet Transform (CWT) of the data.

The ML classifiers implemented were Linear Support Vector Machine (SVM), Polynomial SVM, Multilayer Perceptron (MLP), and k-Nearest Neighbor (k-NN) Classifier.

Linear Support Vector Machine

The best performing Linear SVM was the one using the CWT data with an accuracy of 88.91%.

- Accuracy of Linear SVM using original data: 84%
- Accuracy of Linear SVM using FT data: 62.39%
- Accuracy of Linear SVM using CWT data: 88.91%

Polynomial Support Vector Machine

I tested a Polynomial SVM with four different degree values ($d = 2, 3, 4, 5$). The best Polynomial SVM was the one with a degree of 2 trained on the CWT data with an accuracy of 92.35%.

Dataset	d = 2	d = 3	d = 4	d = 5
Original Data	96.7	87.26	94.78	86.43
FT Data	95.73	86.78	94.61	86.78
CWT Data	97.35	97.26	97.13	97.09

Table 1. Accuracy of Polynomial SVM with different degree (d) values for all the transformed datasets

Multilayer Perceptron

The best performing MLP Classifier was the one trained on the CWT data with an accuracy of 97.22

- Accuracy of MLP using original data: 86%
- Accuracy of MLP using FT data: 87.74%
- Accuracy of MLP using CWT data: 97.22%

k-Nearest Neighbors

I tested the k-NN Classifier for 10 different k values for all the transformed datasets. The best performing k-NN Classifier was the model using CWT data with an accuracy of 97.17%.

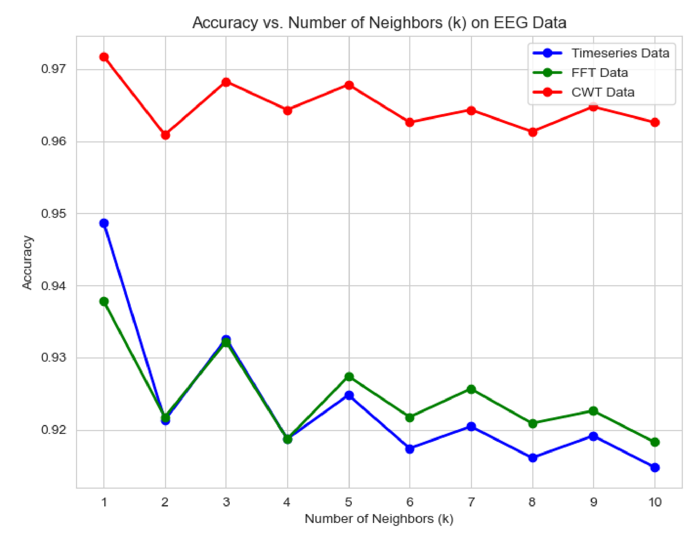

Figure 2. Accuracy values of k-NN classifiers for different k values

Results

If we take the best performing Polynomial SVM ($d = 2$) and k-NN Classifier ($k = 1$), we get the following results:

Dataset	Linear SVM	Polynomial SVM	MLP	k-NN
Original Data	84	96.7	86	94.87
FT Data	62.39	95.73	87.74	93.78
CWT Data	88.91	97.35	97.22	97.17

Table 2. Accuracy of all implemented classifiers for all the transformed datasets

The best performing classifier is the Polynomial SVM ($d = 2$) using the CWT data with an accuracy of 97.35%

Conclusion

- For all the classifiers, the model trained on the CWT data performed the best. The original dataset also performed well across the board, but taking the FT of the dataset did not often improve the performance of our classifiers. This means that frequency spectrum alone is not useful in the detection and diagnosis of epileptic seizures, but when looking at the frequency spectrum over time (taking the CWT) the information became much more useful and allowed us to get better performance.
- The other thing we notice is that for all datasets, the Polynomial SVM ($d = 2$) performed the best. This could indicate that the decision boundary between an epileptic seizure and normal brain activity is non-linear and similar to a second degree polynomial.

Future Work

- Implement principal component analysis (PCA) to reduce training and runtime
- Test different wavelets for CWT
- Test other classifiers such as a Convolutional Neural Network (CNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA)

References

Rashid, H. (2018). Epileptic Seizure Recognition, Version 1. Retrieved October 24, 2023 from https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition.

lyciat@stanford.edu