Introduction:

Epilepsy is a disorder that causes repeated and unprovoked seizures, and it is the fourth most common neurological disorder affecting about 3.4 million people in the US alone. The most common test used to diagnose epilepsy is by using electroencephalograms (EEGs) as they record the electrical activity in one’s brain. 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 different preprocessing techniques and different Machine Learning (ML) classifiers on EEG data to diagnose epileptic seizures.

Background – Dataset:

The dataset used was the Epileptic Seizure Recognition dataset [1] from Kaggle. The dataset consists of 23.6 s EEG signal recordings sampled into 4096 datapoints. Each of these recordings were split into 1s windows resulting in 11500 EEG signal recording windows. Of these 11500 signal recording windows, 2300 are 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 seizure activity and non-seizure activity.

Approach:

The data was randomly split into an 80-20 training-test split. Both preprocessing techniques and different 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 a Linear Support Vector Machine (SVM), Polynomial SVM, a Multilayer Perceptron (MLP), and a k-Nearest Neighbors (k-NN) Classifier. This gave us a total of 12 different experiments to run on the data:

1. Linear SVM trained on the original data
2. Linear SVM trained on the FT data
3. Linear SVM trained on the CWT data
4. Polynomial SVM trained on the original data
5. Polynomial SVM trained on the FT data
6. Polynomial SVM trained on the CWT data
7. MLP trained on the original data
8. MLP trained on the FT data
9. MLP trained on the CWT data
10. k-NN Classifier trained on the original data
11. k-NN Classifier trained on the FT data
12. k-NN Classifier trained on the CWT data

Linear Support Vector Machine:

I used sklearn.svm to implement the Linear SVM and tested it on all the transformed datasets (no transforms, FT, CWT). The best performing Linear SVM was the one using the CWT data with an accuracy of 88.91%.

Polynomial Support Vector Machine:

I used sklearn.svm to implement the Polynomial SVM with four different degree values (d = 2, 3, 4, 5). I tested all four degree values for each transformed dataset and found that the Polynomial SVM with degree 2 performed best on the CWT data with an accuracy of 92.35%.

Multilayer Perceptron:

I used sklearn.neural\_network to implement the MLP Classifier on all the transformed datasets. The best performing MLP Classifier was the one trained on the CWT data with an accuracy of 97.22 %.

I used sklearn.neural\_network to implement the MLP Classifier on all the transformed datasets. The best performing MLP Classifier was the one trained on the CWT data with an accuracy of 97.22 \%.

k-Nearest Neighbors:

I tested the k-NN Classifier for 10 different k values for all the transformed datasets. The k-NN Classifier performed the best on all datasets when k = 1. The best performing k-NN Classifier was the model using CWT data with an accuracy of 97.17\%.

Results:

If we take the best performing Polynomial SVM and k-NN Classifier, we get the following results:

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

Conclusions:

In general, taking the CWT of the data improved the performance of the classifiers. The original time series dataset also performed well across the board, but the taking the FT of the dataset did not often improve the performance of our classifiers. This could mean 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 becomes much more useful and allows 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: