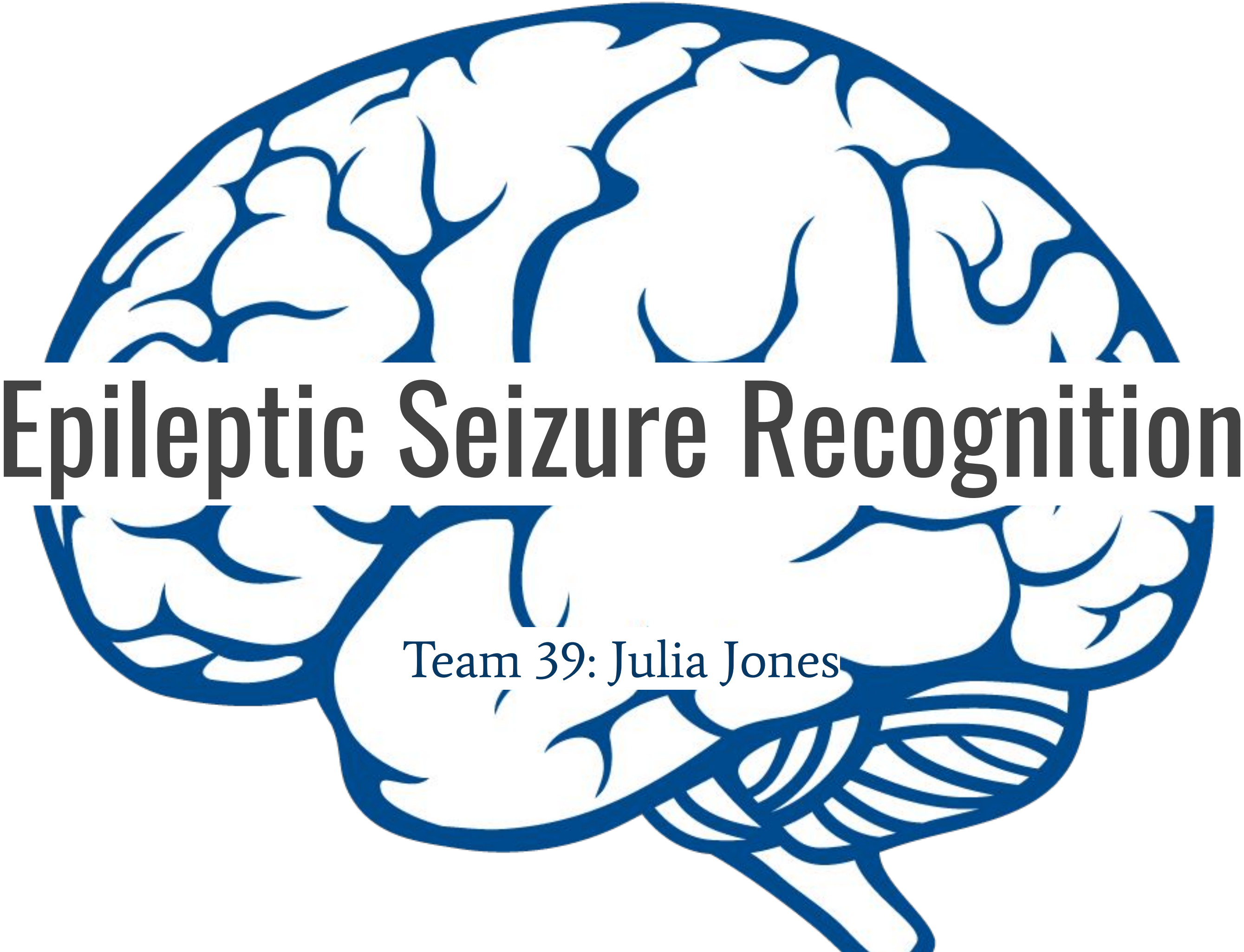
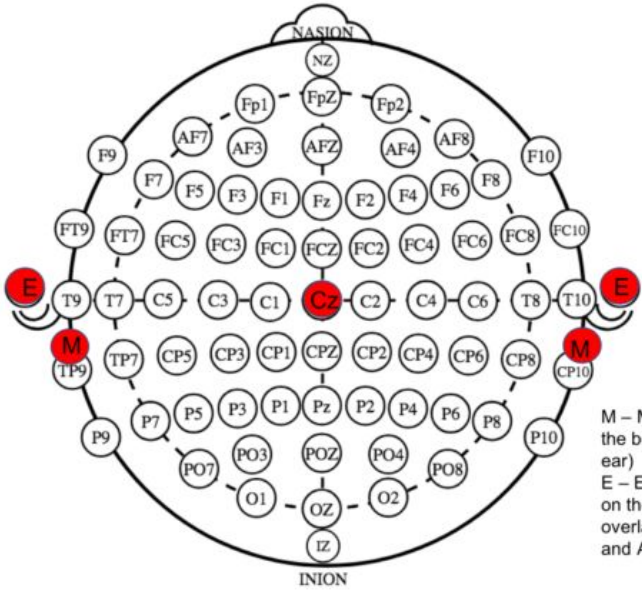


Epileptic Seizure Recognition

Team 39: Julia Jones

Project Description

The goal of this project is to classify EEG data on epileptic seizures. I accomplished this through classification of both feature reduced data and unreduced feature data.



M – Mastoids (on the bone behind the ear)
E – Ears (typically on the earlobe, overlap with A1 (L) and A2 (R).

Importance/Motivation

Being able to classify EEG data for epileptic seizures is important for the advancement of our understanding on them.

- Understanding what signals precede and compose an episode could improve diagnosis and treatment of these seizures.

Related Works: Literature

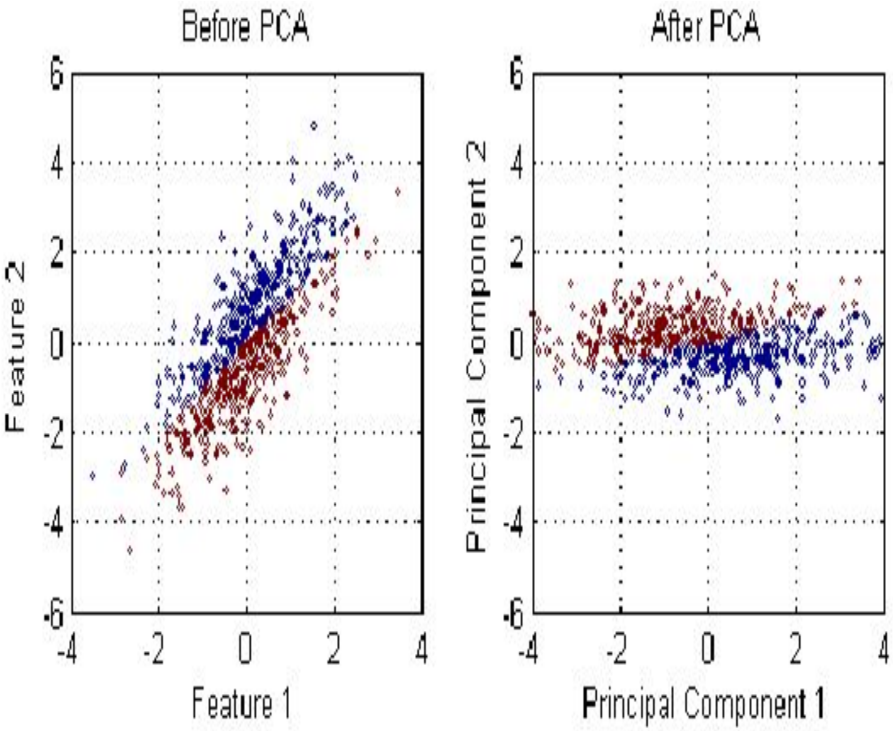
“Wavelet-based feature extraction for classification of epileptic seizure EEG signal:”

- This is a 2017 paper from the Journal of Medical Engineering & Technology. The authors used binary classification of 5-class seizure data. We also used some of the same feature reducers and classifiers.

Related Works: Course discussion

Principal component analysis (PCA).

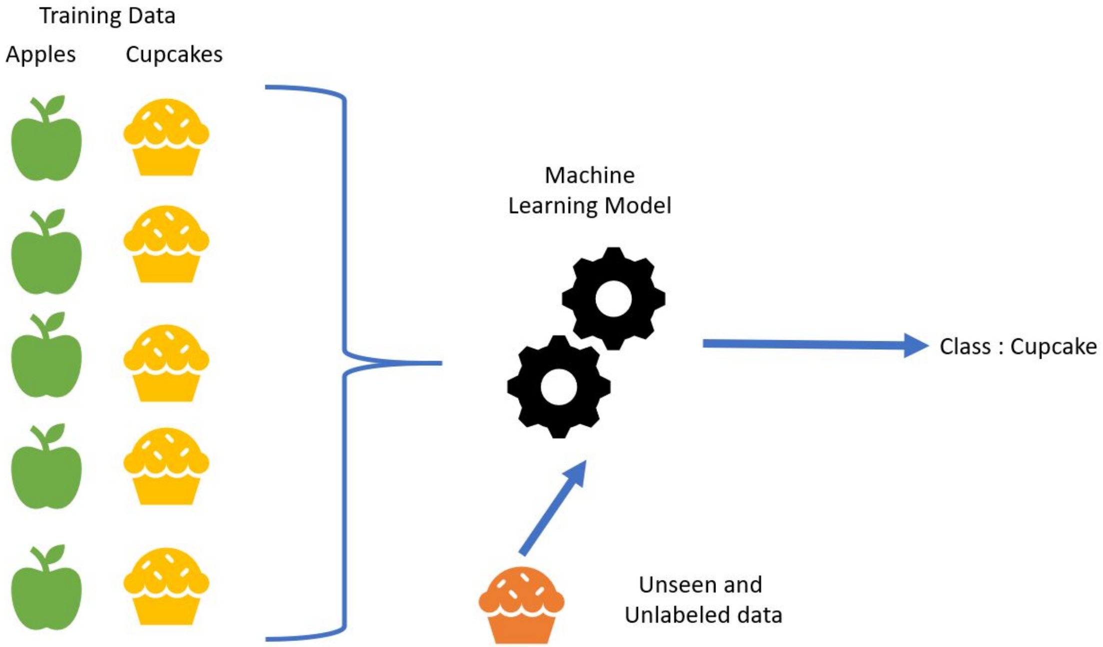
Principal component analysis was first covered in the course in the BCI Review paper in section 5.1 *Principal Component Analysis (PCA)*



Related Works: Course discussion

Classification

Classification was first covered in the BCI Review paper in section 5 *Features Extraction and Selection*



Data Explanation

In the original dataset there were 5 folders. Each folder represented one of the 5 states that was being studied:

1. Recording of seizure activity
2. Recording at location of brain tumor (no seizure)
3. Yes they identify where the region of the tumor was in the brain and record the EEG activity from the healthy brain area
4. Eyes closed, means when they were recording the EEG signal the patient had their eyes closed
5. Eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

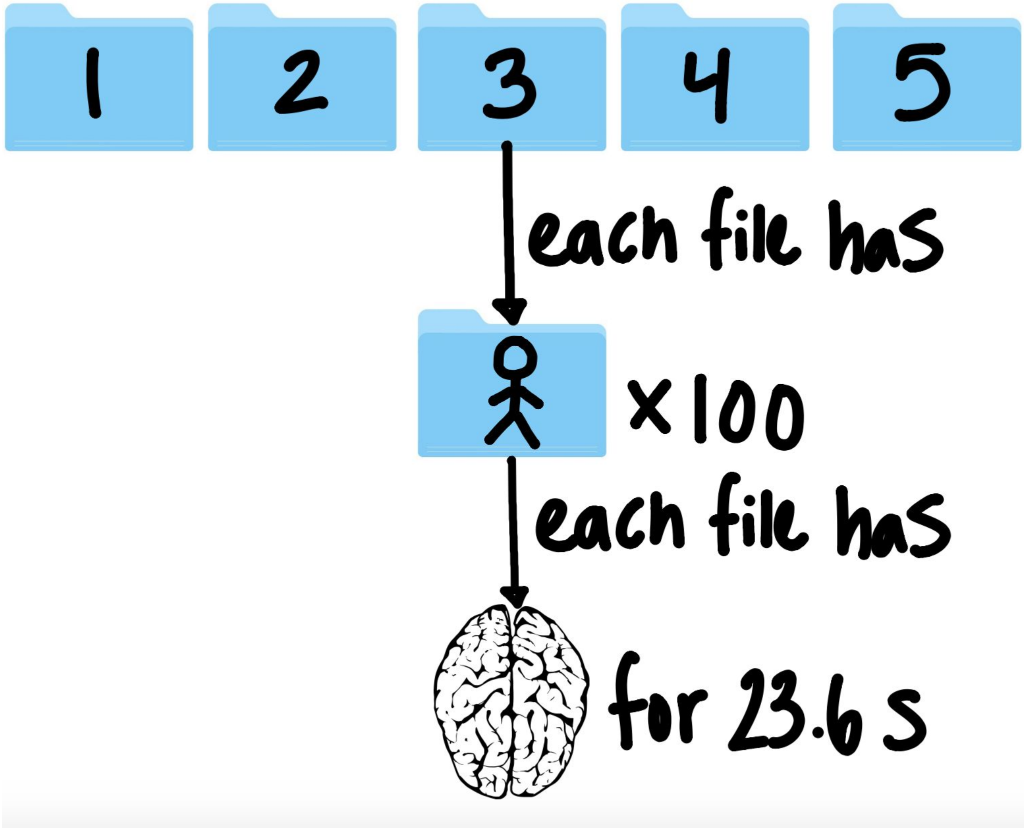
Data Explanation cont.

Each of the previously explained 5 folders contained 100 files. Every files represents a one subject/person.

Each subject file is a recording of brain activity for 23.6 seconds. This time-series is sampled into 4097 data points. A data point is the value of the EEG recording at a different point in time.

These 4097 data points were divided and shuffled into 23 chunks. Each chunk contains 178 data points for 1 second. So now we have 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column), the last column represents the label y {1,2,3,4,5}.

Data Explanation cont.

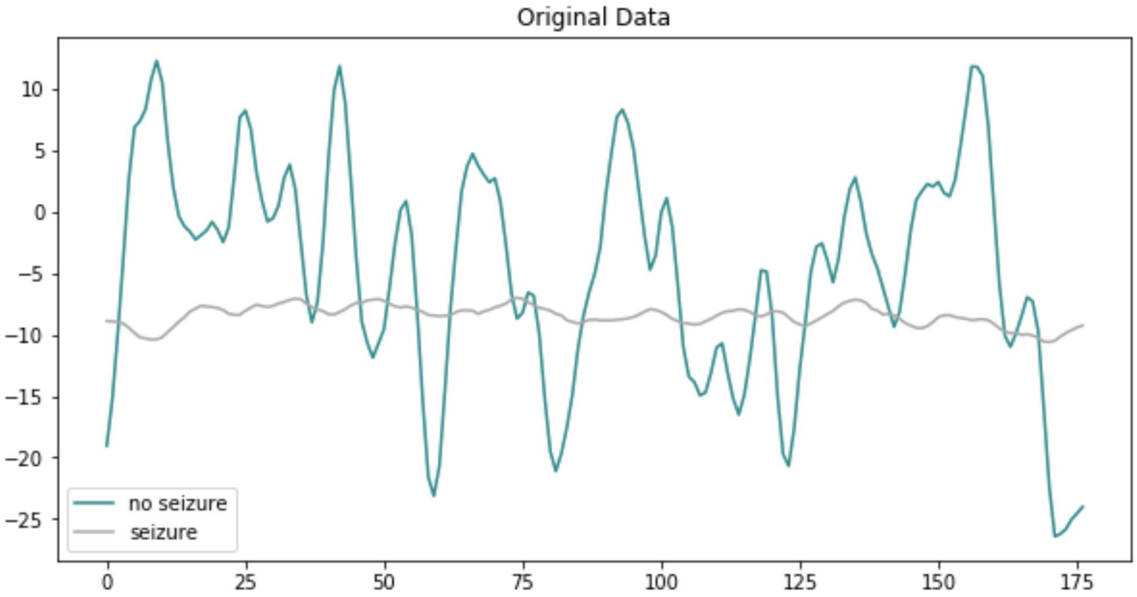


Methods: Overview

- Clean up data (make data binary classifiable)
- Determine train and test data (non-random)
- Perform feature reduction (PCA)
- Perform classification (RandomForest, KNN, LDA, SVC) on PCA data
- Perform classification (RandomForest, KNN, LDA, SVC) on unreduced data
- Analyze results

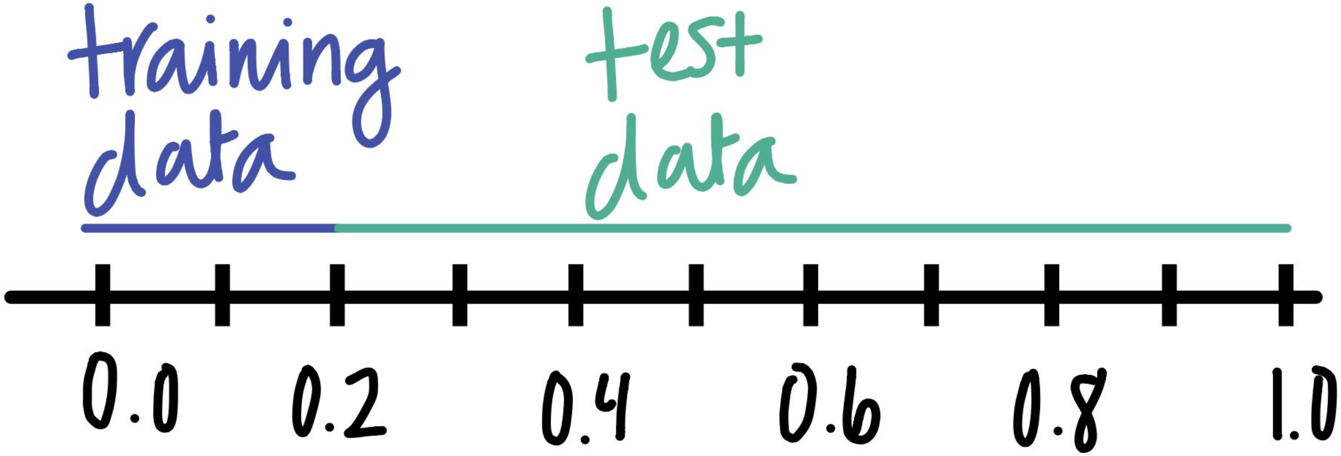
Methods: making data binarily classifiable

In the original data classes 2-5 represent EEG data from subjects not having a seizure and class 1 represents all subjects having a seizure, the data can be binarily classified. Class 1 stayed class 1 and classes 2-5 were combined into Class 0 (no seizure).



Methods: Determine training and test data

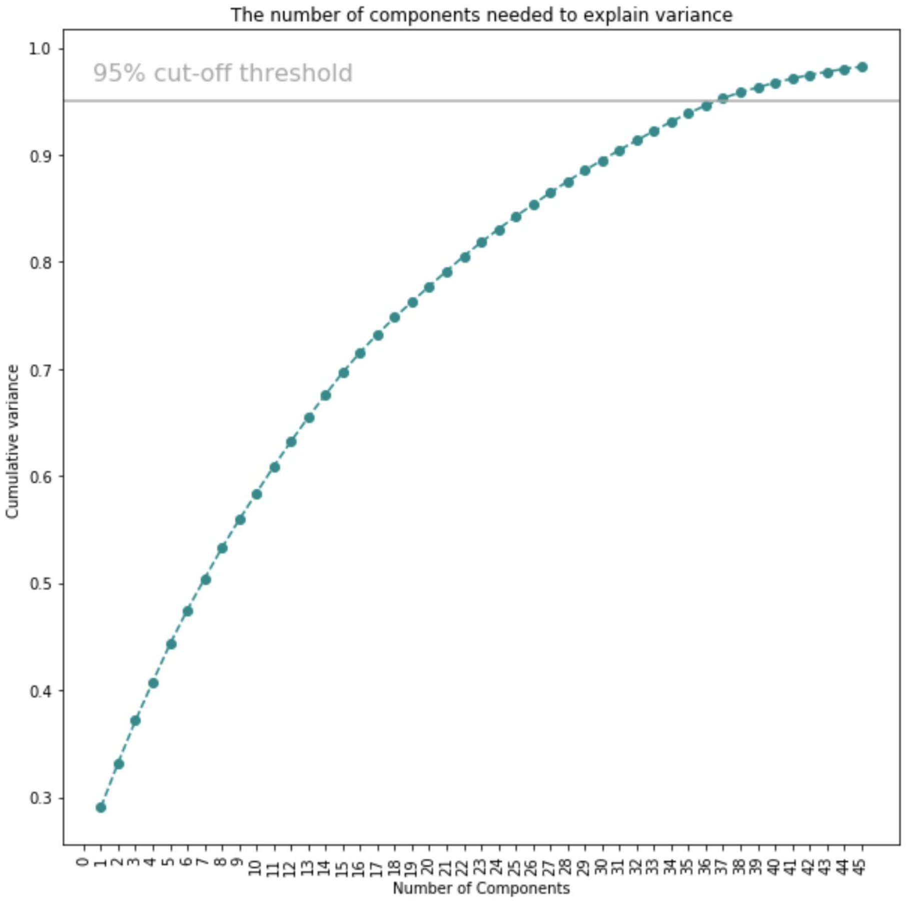
Since this data was taken continuously for each subject, I did not randomly select training and test data. Instead training and test data were taken from distinct time segments. This was an attempt to reduce the time bias on my classifier.



Methods: PCA

Next I performed PCA as a feature reduction method on the data.

- To determine how many features to reduce to I computed the cumulative variance. I graphed it to visually determine the feature cut-off point.
- After examining this figure, I chose to reduce the data to 36 features.



Methods: PCA+Classification

After reducing the data, I then performed classification on the PCA data. I ran four different classifiers to compare the results.

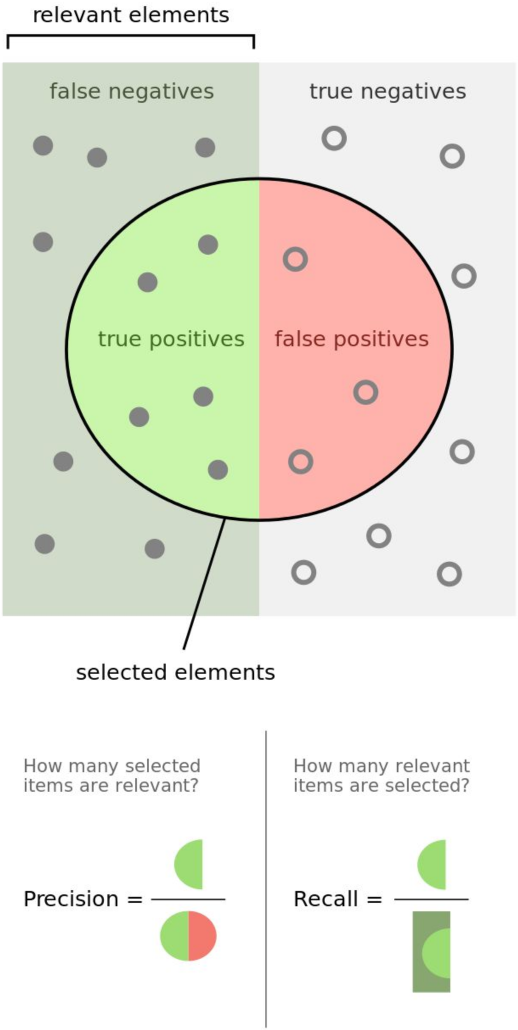
- I ran KNN since it was used in the related works paper and worked well
- I ran LDA and SVC they were covered in the BCI Review paper and seemed to fit the bill
- I ran RandomForest because it performed the best naively out of all the other classifiers I tested
- For all of these classifiers I utilized GridSearchCV to refine my input parameters

Methods: Classification with no feature reduction

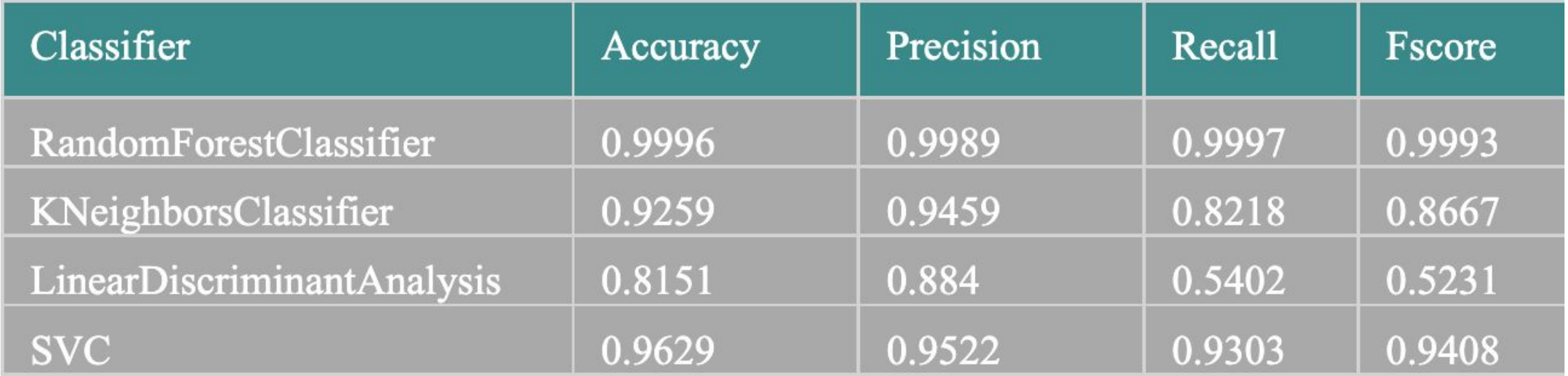
I then performed classification on the unreduced data. I ran the same four classifiers for then the results could truly be comparable.

Methods: Analysis

After running the classifiers, I ran some cursory data analysis to be able to properly interpret my results. I calculated the accuracy, precision, recall, and f score on all of my data.



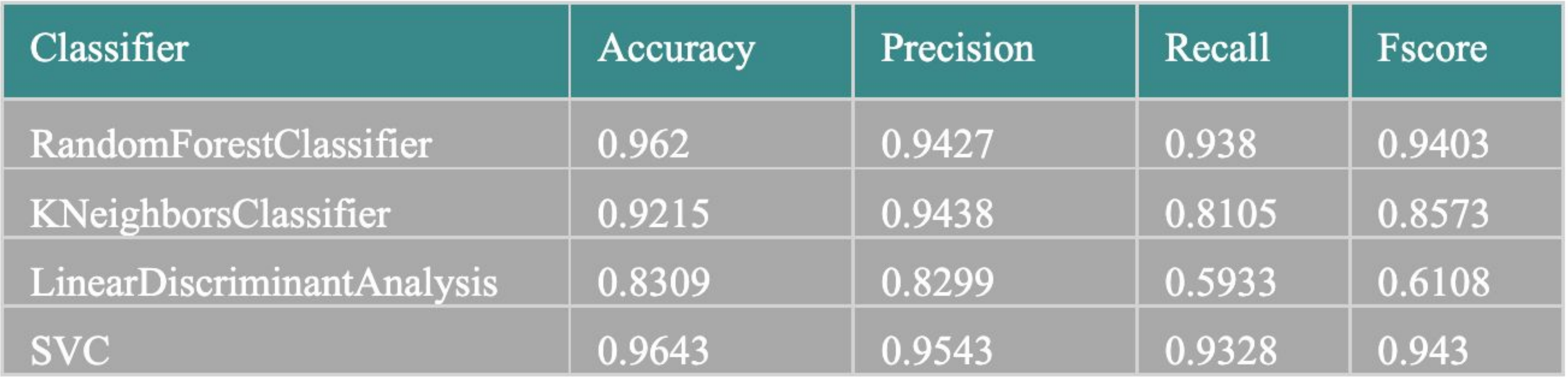
Results: PCA+Classification



Classifier	Accuracy	Precision	Recall	Fscore
RandomForestClassifier	0.9996	0.9989	0.9997	0.9993
KNeighborsClassifier	0.9259	0.9459	0.8218	0.8667
LinearDiscriminantAnalysis	0.8151	0.884	0.5402	0.5231
SVC	0.9629	0.9522	0.9303	0.9408

RandomForest performed the best, with K-nn and SVC close behind, while LDA by far performed the worst

Results: Classification with no feature reduction



Classifier	Accuracy	Precision	Recall	Fscore
RandomForestClassifier	0.962	0.9427	0.938	0.9403
KNeighborsClassifier	0.9215	0.9438	0.8105	0.8573
LinearDiscriminantAnalysis	0.8309	0.8299	0.5933	0.6108
SVC	0.9643	0.9543	0.9328	0.943

SVC performed the best, with RandomForest and K-nn as close seconds

Results: Summary

PCA

- RandomForest
  - Accuracy: 0.9996
  - Fscore: 0.9993
- KNeighbors
  - Accuracy: 0.9259
  - Fscore: 0.8667
- LDA
  - Accuracy: 0.8151
  - Fscore: 0.5231
- SVC
  - Accuracy: 0.9629
  - Fscore: 0.9408

Unreduced

- RandomForest
  - Accuracy: 0.962
  - Fscore: 0.9403
- KNeighbors
  - Accuracy: 0.9215
  - Fscore: 0.8573
- LDA
  - Accuracy: 0.8309
  - Fscore: 0.6108
- SVC
  - Accuracy: 0.9643
  - Fscore: 0.943

Results: What went wrong

Nothing went horribly wrong with my project, but I did have some issues with overfitting, especially with RandomForest.

- Even after utilizing GridSearchCV, to try and reduce the depth of the decision tree RandomForest was still giving near perfect scores with the PCA data. This is almost certainly due to overfitting.
- In future machine learning projects I'll take extra precautions to allay this issue.

Discussion: Possible Improvements: EEG Recording

Since subjects in this experiment were recorded continuously, a time dependency could be present (ie that my classifiers are learning which time window the recording is from).

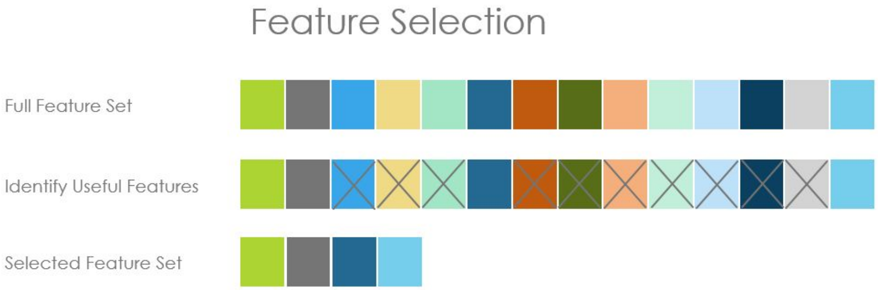
In an attempt to allay this I did not randomly select train and test data: I took training and test data from different segments of the recordings.

This helped allay the time issue, but this is not a perfect fix. They best way to fix this issue is to record the data properly

Possible Improvements: Further Algorithmic exploration

Since I only tested four different classifiers, a possible improvement to this project would be to examine more classifiers effectiveness with classifying feature data.

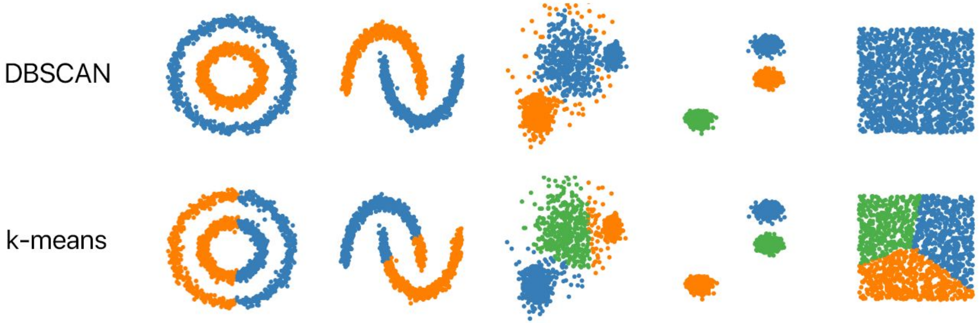
A similar improvement would to be to try more feature reducers


Feature Selection

Full Feature Set

Identify Useful Features

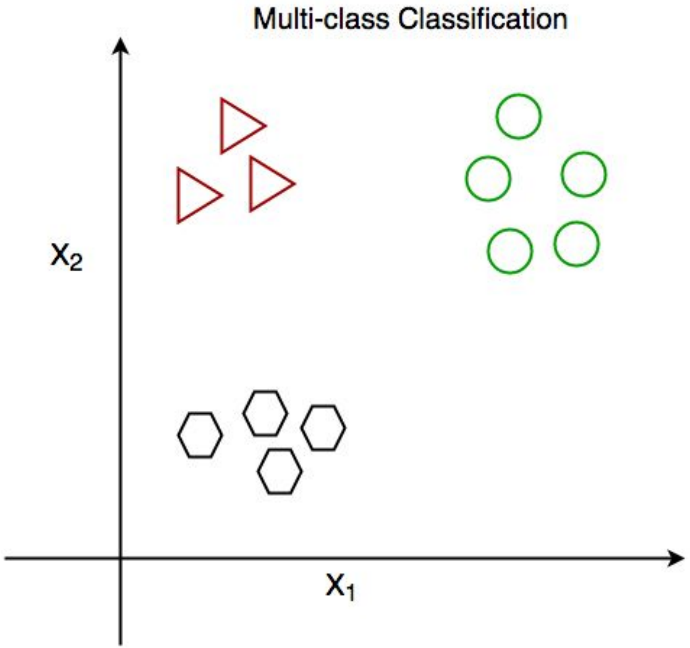
Selected Feature Set



Possible improvements: Multiclass Classification

Since this dataset did originally include five classes, training a classifier to properly classify seizure data in a multiclass setting as a natural next step.

- Testing out the performance of multi-class EEG data would also open up experimentation with other classifiers covered in class such as ANN



Discussion: What did I learn

What I learned about classifying seizure data:

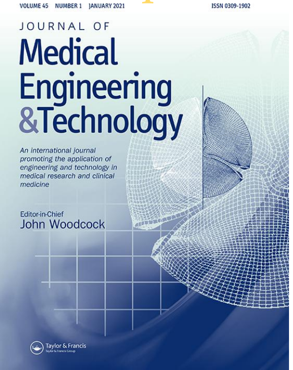
- While some classifiers performed better than others, there is a clear distinction between EEG signals from a subject experiencing seizures vs. a subject not experiencing seizures. All classifiers performed above chance though LDA had a recall and f score close enough to chance that I would discard them from future analysis
- Overfitting on such data can be a real issue, especially when using tree based classifiers such as RandomForest

Sources

Related literature paper: https://www.tandfonline.com/doi/full/10.1080/03091902.2017.1394388

Raw EEG data (in csv format): UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

Code I used to determine binary classes @jaketuricchi: https://www.kaggle.com/jaketuricchi/using-pca-and-clustering-to-improve-classification



VOLUME 45 NUMBER 1 JANUARY 2021 ISSN 0309-1902

JOURNAL OF Medical Engineering & Technology

An international journal promoting the application of engineering and technology in medical research and clinical medicine

Editor-in-Chief
John Woodcock

Taylor & Francis

UCI Machine Learning Repository

kaggle

Thank you for listening!

I hope you enjoyed my presentation.