Predicting Seizures with EEG Reading Time Series

Oliver Newland advised by Professor Michael Hughes

October 23, 2018

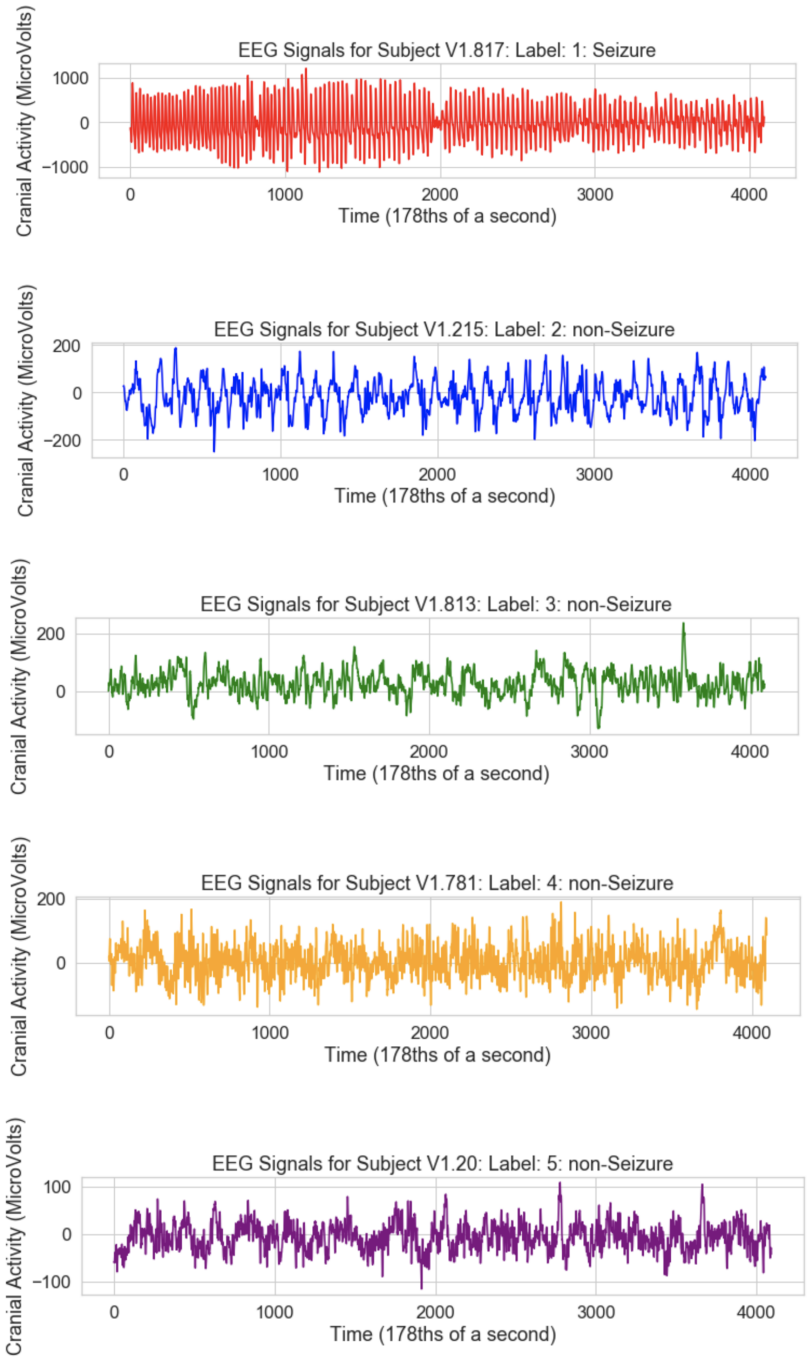
Introduction:

This project served as an introduction to classifying medical conditions with time series data, specifically predicting whether a one second chunk of an EEG reading could be classified as seizure or non-seizure cranial activity. The bulk of the project focused on getting used to working with time series data and the process of selecting features that best separated the different patterns of EEG signals. The data was sourced from a 2001 study in The American Physical Society's *Physical Review*, and featured the EEG readings of 500 patients split into 23 one second long readings, divided into 5 evenly sized categories of activity patterns (our labels), for a total of 11500 chunks of data. Typical descriptive statistics for time-series like mean, median, slope, and variance proved very productive as model features, especially when paired with more specific figures that measured frequency characteristics like power spectral density (Welch's method) and the number of times two adjacent readings are on opposite sides of the mean. Data was validated across 5 folds, each with equal representation of labels and all chunks from each subject grouped together, scikit-learn's LogisticRegression was able to achieve an average accuracy of 93.12% with average training loss of 6.51% and ExtraTreesClassifier was able to achieve an average accuracy of 96.52% with average training loss of 0.00%. Unfortunately I did not have time to do hyperparameter tuning, but that would be an excellent next step. However, the results are promising and the default hyperparameters performed admirably.

Part 1: Descriptive Statistics

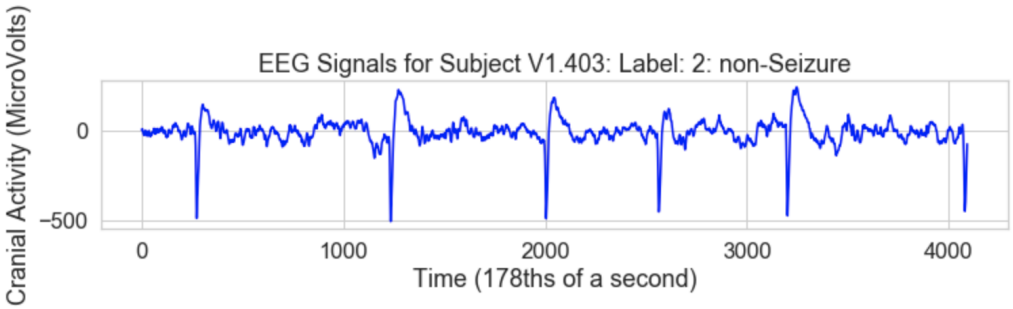
1.1 Individual Time Series

As an initial investigation, rather than pouring over all 500 time series, I opted to produce 5 randomly selected time series from each label and see what I could make of them. The typical EEG reading for each label can be found below:

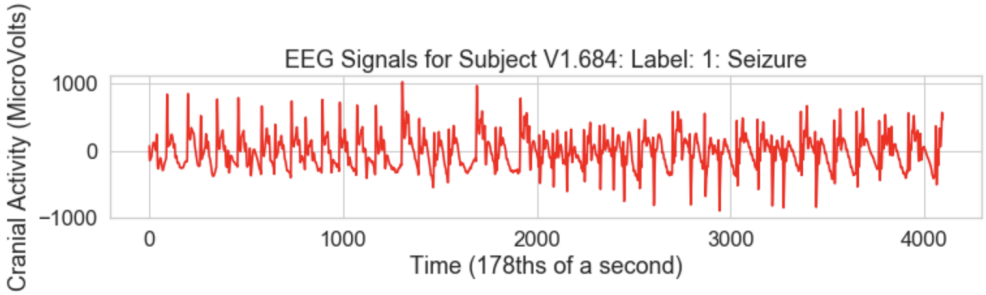


Seizures featured by far the highest frequency of spikes and valleys, and those readings frequently ranged far outside the -200 to 200 microVolt range that was typical of the other 4

patterns. The magnitude and compactness of these spikes made them easy to pick out. Pattern 2 was most similar to seizures, but lacked the frequency and magnitude, except in some rare cases like the following:



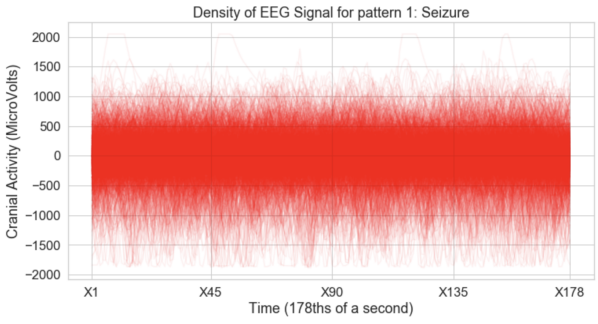
However, even this reading lacks the same kind of frequency as is common in the seizure readings. Patterns 4 and 5 also had a dense amount of peaks and valleys, but their magnitudes were at once much smaller and less consistent. Even in a seizure reading's least dense state, like the following figure, it is still notably more extreme than the other patterns.

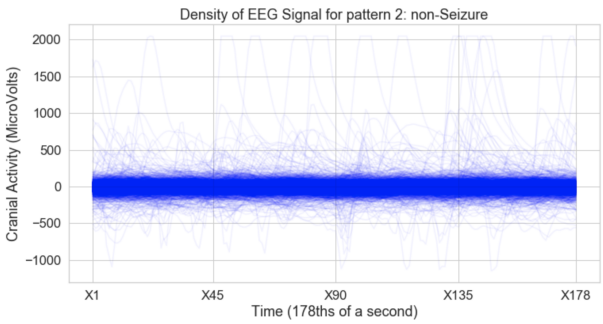


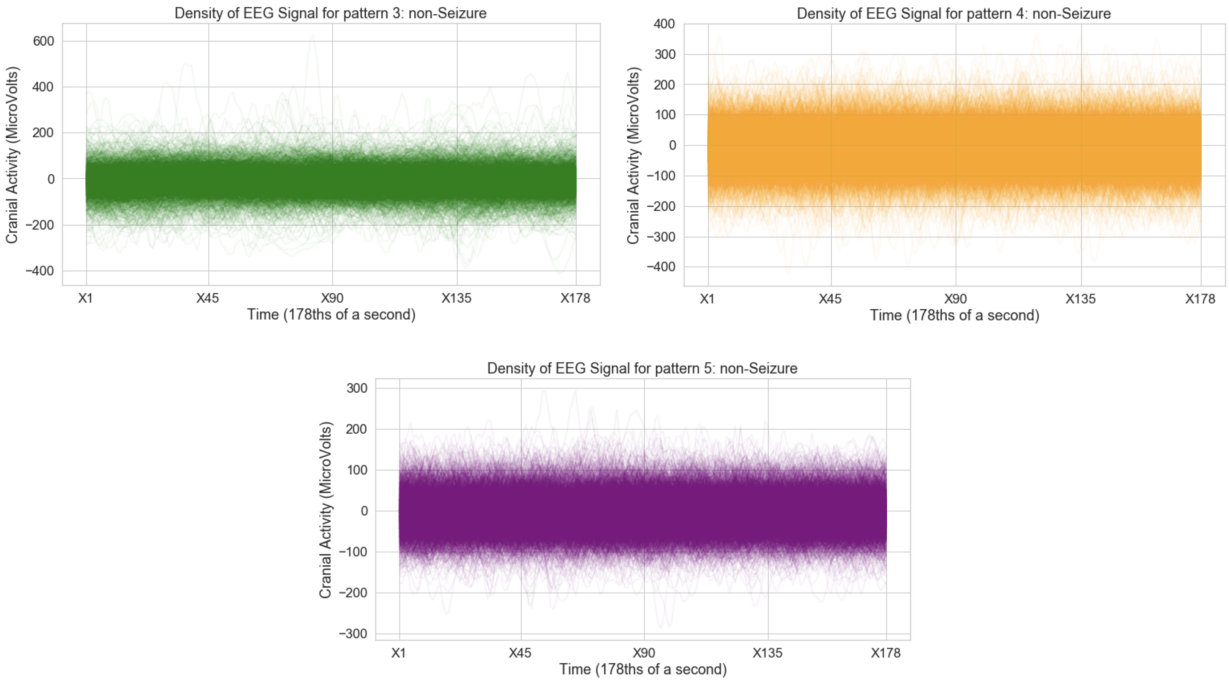
From a simple eye test, it appears that classification is very viable.

1.2 Chunks by Label

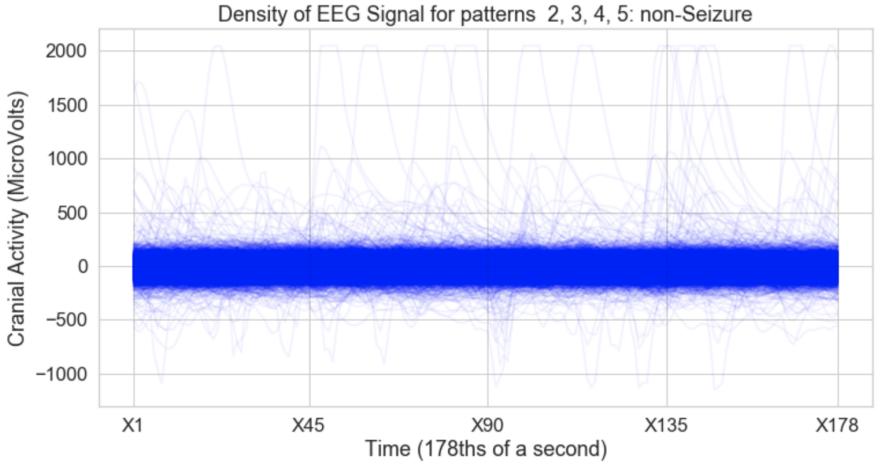
The next eye test gave a stronger sense of the typical magnitude and consistency of each of the labels. By overlaying each chunk at a low opacity with all the other chunks of a similar label, we can see that general heatmap of each chunk:





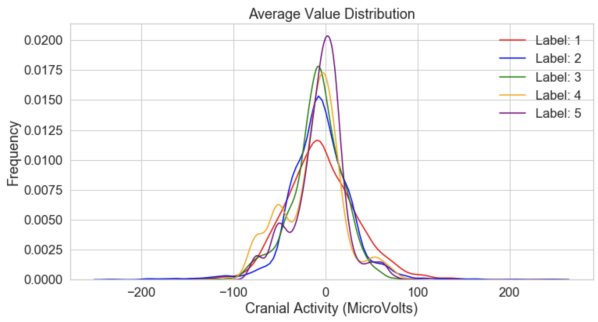


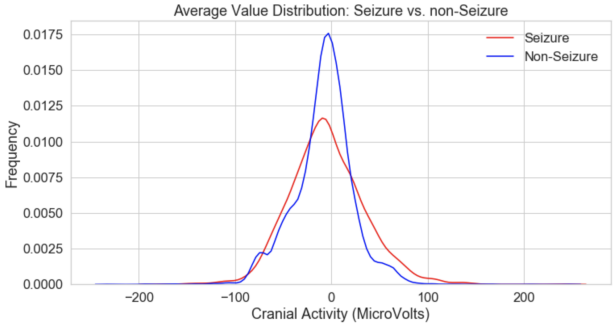
Seizure readings inhabit a much wider range and one can see how the readings diffuse slowly from the central, solid band. One can also see potential meddling outliers in the data, such as one subject's large spikes in pattern 2 that have magnitude of around 2000 microVolts. We can also group the non-seizure labels 2, 3, 4, and 5 together and compare to label 1 for similar results:



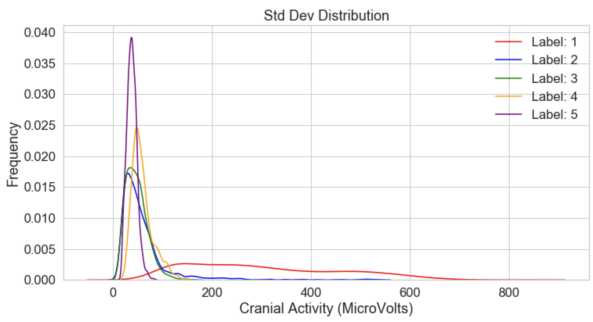
1.3 Descriptive Statistics by Label

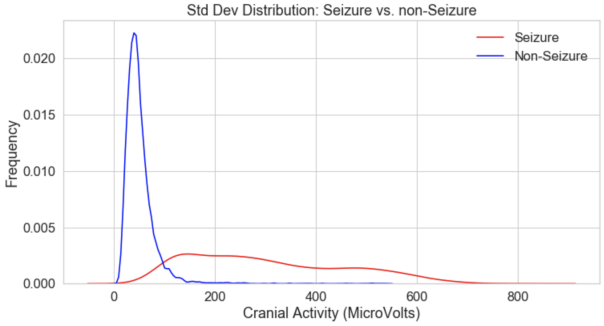
These descriptive statistics put numbers where our intuition pointed. We can examine the average value, standard deviation, the quantiles, slopes, and frequency measures to reveal contrasts in the chunks, especially when we consolidate the non-seizure labels together.



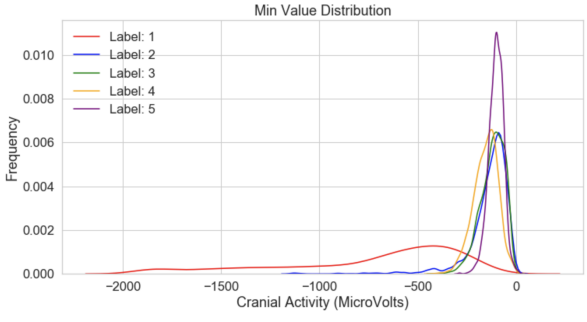


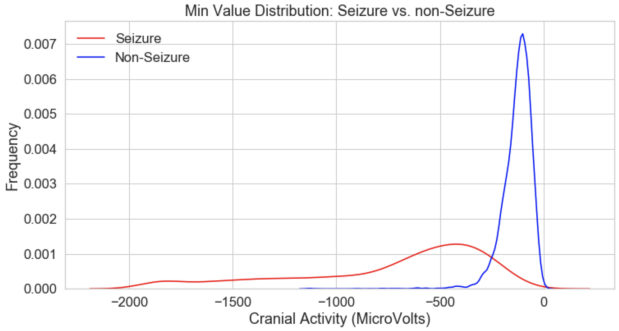
Above we can see that an average at or slightly below 0 is common in both, but more common in non-seizure subjects.



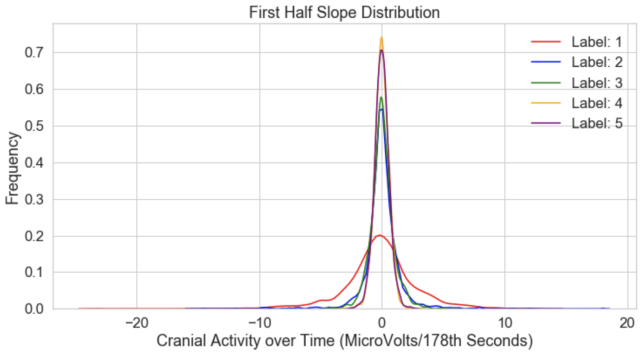


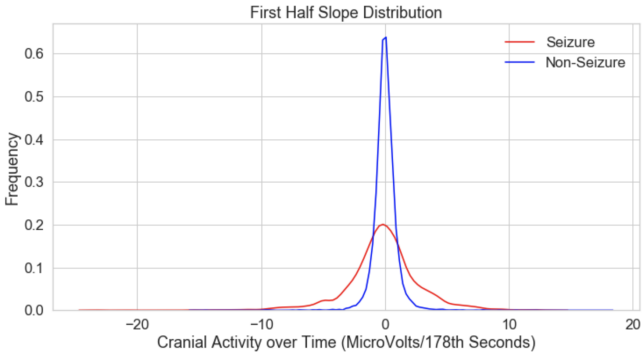
Standard deviation and variance reveal a lot, a reading with variance above around 175 microVolts is almost guaranteed to be a seizure, and a large portion of seizures fall in that category. (In fact, in a small experiment with variance as the only feature, the accuracy of our classifiers hovered around 94%).



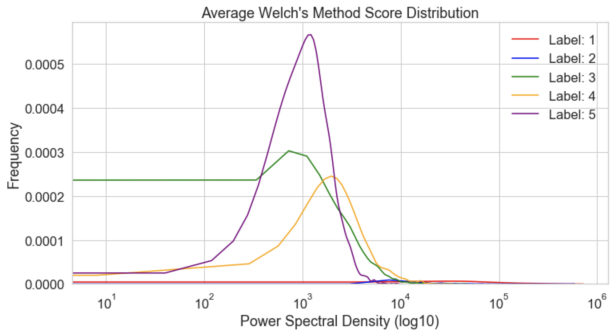


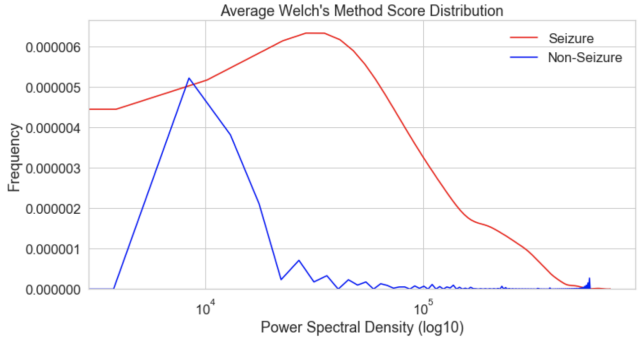
Quantile analysis reveals a pattern one might expect, with seizures generally being more extreme in both the positive and negative direction.



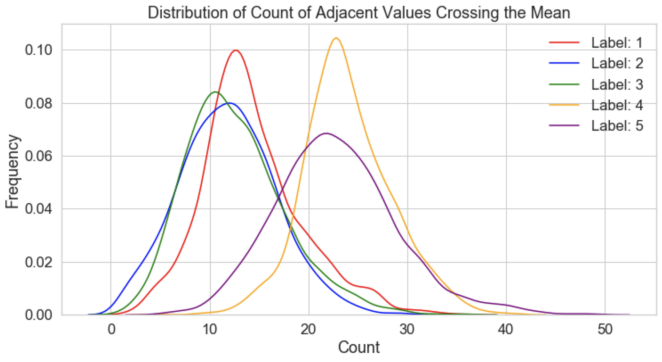


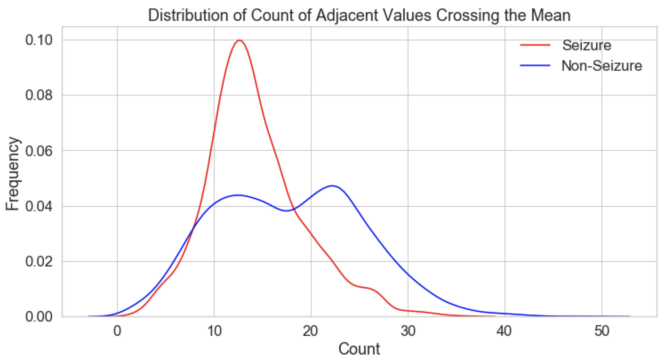
Surprisingly, seizures had a more erratic slope measurements when chunks were split in half. From just looking at the readings, I had predicted that the non-seizure measurements followed more of a wave, and would thus go and up and down more. However, the opposite seems true. This difference seems about as useful as the difference in averages: not particularly telling unless the slope is extreme.





Welch's Method maps frequency density scores (how often we encounter spikes and how extreme they are) across a time series, and when we take the average of those scores, we get a good sense of how frequently our readings are varying. From the first plot we can see that this helps differentiate label 1 from labels 3, 4, and 5 quite a bit, but not as much from label 2. In fact, the clash of label 2's frequency density with the other non-seizure labels makes the second plot harder to interpret, as those patterns are working in opposite directions. This suggests taking a one vs. all approach when training our classifiers.





Finally, our measure of how many times consecutive readings are on opposite sides of the chunks average value, another more rudimentary indicator of frequency density, shows that once again one vs. rest methodology might be most telling.

1.6 Can we tell the classes apart?

In general, the time-series for seizures have characteristics easily recognizable from visual observation. They are dense with spikes and valleys, and have large, consistent magnitudes. Large spikes are second most common in time series of label 2, however they are far less common in general and less dense when they do occur. Differentiating between 3, 4, and 5 is more difficult, but 4 and 5 seem to have denser frequencies. With information on their average Welch score one could make a reasonably educated guess. Variance seems like the most useful feature by far for picking out seizures, as it captures that density of high range readings.

1.7 What performance do we expect? Perfect classifier possible?

As far as classifying seizures vs. non-seizures, we can expect very good performance. For most of the descriptive statistics above, the other 4 labels tend to group together, and seizure chunks do not seem to overlap too heavily. However, there are cases of overlap in every descriptive statistic, which essentially rules out a perfect classifier. There is chance that a non-seizure can have many similar characteristics to a seizure, however that chance is quite small, and our classifiers should perform well.

Part 2: Training Classifiers

2.1 Feature Vector and Cross Validation

Encouraged by the discussion above, each chunk of 178 readings was refined into a feature vector with the following measurements: mean, variance, median, minimum value, maximum value, average Welch score, the number of times consecutive readings cross the mean, and the slope of each half of the chunk. These features capture a lot of information while keeping the dataset small.

Cross validation was performed to ensure the robustness of the classifiers' supposed accuracies. Five folds were created, each containing all the chunks of 20 subjects from each label, for a total of 100 subjects, or 2300 chunks. It was vital to keep any given subjects chunks in the same fold, as chunks form the same time series tend to be extremely similar, and mixing them into both the training and testing data would give us very high accuracy scores. Each fold was used to validate classifiers trained on the other four folds.

2.2 Confusion Matrices

Each time the notebook is run, it shuffles the data. The following is all the results from one run:

2.2.1 LogisticRegression (one vs. rest)

Fold 0:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	355	105
Actual: Non-Seizure	46	1794

Validation success: 2149/2300 for 93.43%
Training loss: 629/9200 for 6.84%

Fold 1:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	371	89
Actual: Non-Seizure	84	1756

Validation success: 2127/2300 for 92.48%
Training loss: 628/9200 for 6.83%

Fold 2:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	407	53
Actual: Non-Seizure	51	1789

Validation success: 2196/2300 for 95.48%
Training loss: 636/9200 for 6.91%

Fold 3:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	366	94
Actual: Non-Seizure	75	1765

Validation success: 2131/2300 for 92.65%
Training loss: 662/9200 for 7.20%

Fold 4:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	345	115
Actual: Non-Seizure	79	1761

Validation success: 2106/2300 for 91.57%
Training loss: 438/9200 for 4.77%

Overall an average accuracy of 93.12%, with training losses that closely match validation errors. A good portion of errors seem to occur when the classifier predicts a seizure as a non-seizure, ie false negatives. This form of error is undesirable as we would like to be sensitive to time series that could be seizures, so we can ensure we recognize all of them, even at the cost of recognizing some false positives.

2.2.2 ExtraTreesClassifier

Fold 0:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	423	37
Actual: Non-Seizure	28	1812

Validation success: 2235/2300 for 97.17%
Training loss: 0/9200 for 0.00%

Fold 1:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	422	38
Actual: Non-Seizure	36	1804

Validation success: 2226/2300 for 96.78%
Training loss: 0/9200 for 0.00%

Fold 2:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	413	47
Actual: Non-Seizure	12	1828

Validation success: 2241/2300 for 97.43%
Training loss: 0/9200 for 0.00%

Fold 3:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	430	30
Actual: Non-Seizure	43	1797

Validation success: 2227/2300 for 96.83%
Training loss: 0/9200 for 0.00%

Fold 4:

N = 2300	Predicted: Seizure	Predicted: Non-Seizure
Actual: Seizure	382	78
Actual: Non-Seizure	51	1789

Validation success: 2171/2300 for 94.39%
Training loss: 0/9200 for 0.00%

Overall an average accuracy of 96.52% is pretty excellent, however, our training loss is 0% for all folds, suggesting that the ExtraTreesClassifier is overfitting a bit. However, the results are still better than the LogisticRegression classifier. We still suffer the same problem with false negatives that LogisticRegression did, but on a smaller scale.

2.3 What is possible here? Is this feasible?
Without hyperparameter tuning or an added bias towards false positives, even basic classifiers seem to be up to the task of detecting seizures. With only a second of data, we can be >90% sure that our prediction is correct, which is pretty fast. With further tuning and more finely tuned feature variables, the usefulness of a classifier like this is abundantly clear.

2.4 Next Steps
Next steps should involve error and hyperparameter tuning, exploration of predicting from multiple chunks, further feature investigation, and use of more advanced classifiers.