Machine Learning Analysis of Schizophrenic EEG Records

Christian Chitty, BME Undergrad '23

Amadu Toronka, M.S – Medical Physics '22

Maria Williams, MEng – AIPI ‘22

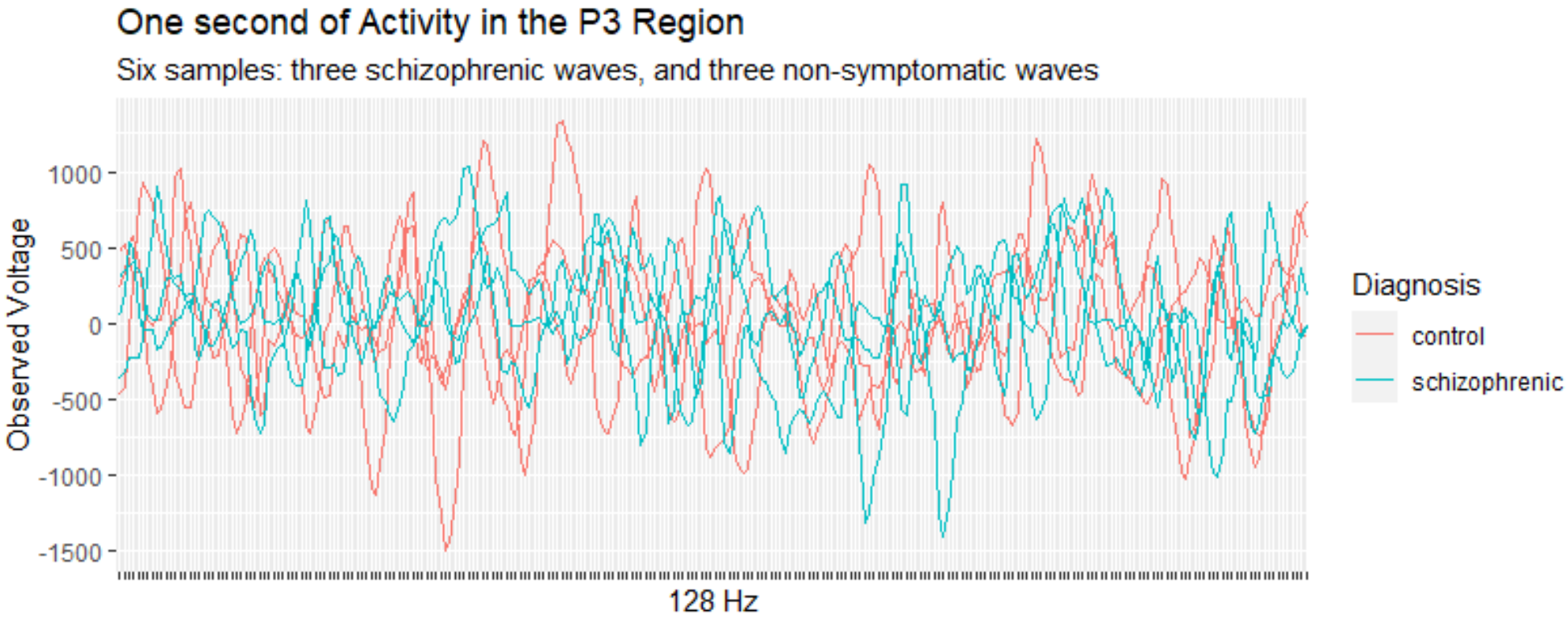
Problem Description

The Problem:

- Schizophrenia is a chronic brain disorder characterized by disruption in cognitive function.
- It is important that it is diagnosed and treated as early as possible to mitigate its damage.
- It has been difficult to find reliable biomarkers to help diagnose and treat this disorder.

Our Goals:

- Support the finding that schizophrenia involves abnormal relationships between different regions of the brain
- Identify which areas of the brain are more highly correlated in schizophrenic brains
- Examine waveform patterns for other indicators of the disorder



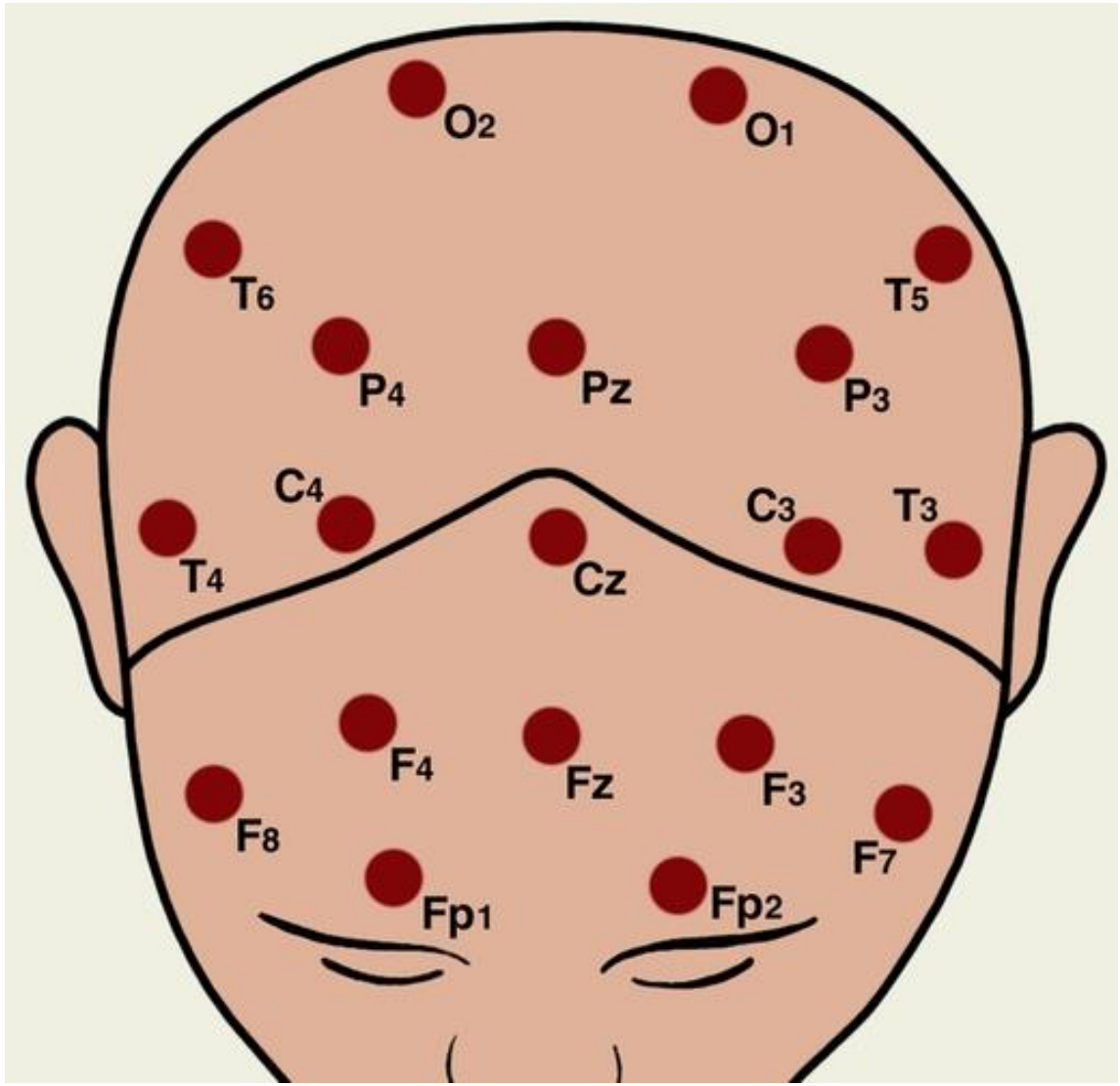
Data

- EEG readings from 84 adolescent subjects: 45 with schizophrenia and 39 in a control group
- A full minute of data at 128Hz (128 readings per second) from 16 channel sensors (122881 factors)
- Originally obtained and pre-processed by researchers at Lomonosov Moscow State University
- Chosen for its cleanliness and for its subjects: the age and stage that schizophrenia is most often first diagnosed

Feature Engineering

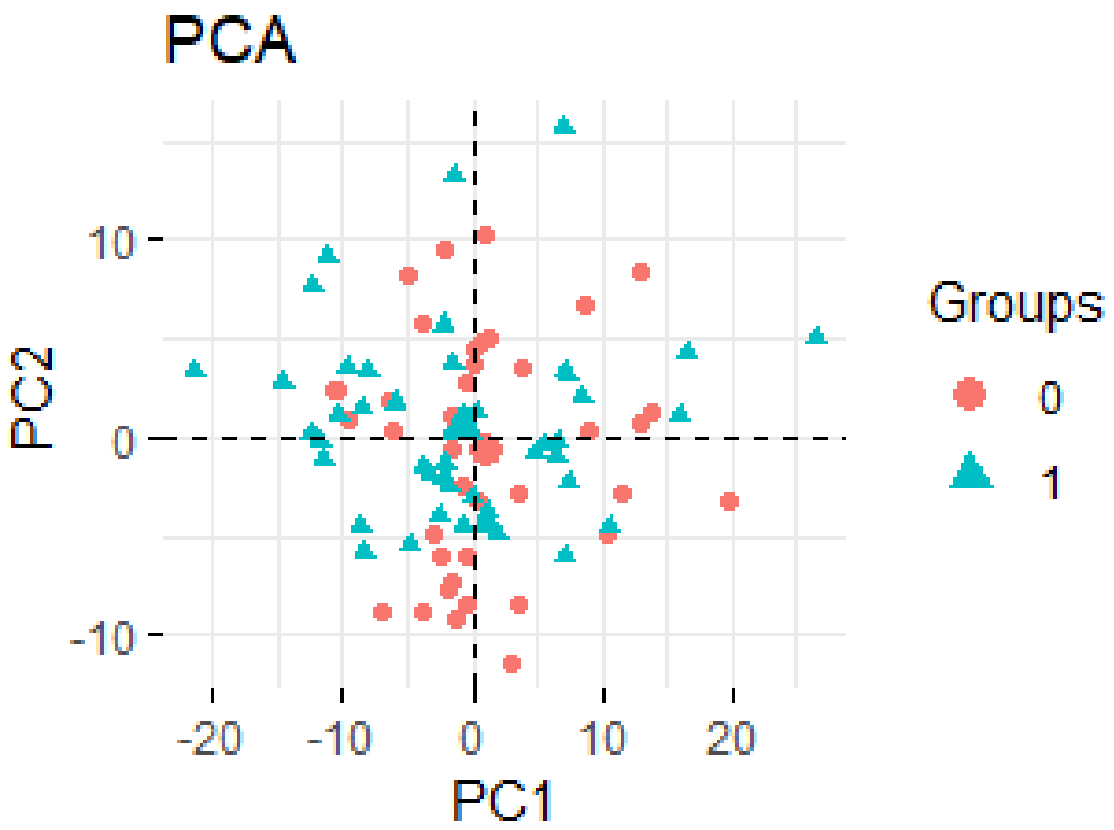
We compiled a dataset of 84 observations with 281 factors:

- a Boolean diagnosis flag
- 64 sensor statistics (4 stats for each of the 16 sensors)
  - Average voltage
  - Minimum voltage
  - Maximum voltage
  - Range of voltage values
- 96 waveform statistics (6 stats for each of the 16 sensors)
  - Number of peaks
  - Number of valleys
  - Average peak voltage
  - Average valley voltage
  - Range in peak voltage
  - Range in valley voltage
- 120 correlation values (for each pair of 16 sensors)
  - correlation values less the .5 were represented as 0



Exploratory Data Analysis

- We looked at the data in many ways including calculating p-values
- The most valuable aspect of our EDA was PCA
  - Expected correlation
  - Took loadings explaining top 95% of variance
  - Combined and maintained these



Modeling

What

- K-Nearest Neighbors on entire dataset
- Regularized Logistic Regression (LASSO) on top 30% of features as determined by PCA

Why

- Classification problem with actual data
- Many more features than observations

Challenges

- Different results due to few observations: solved by using k-folds cross validation

Evaluation

- Accuracy across 10 k-folds

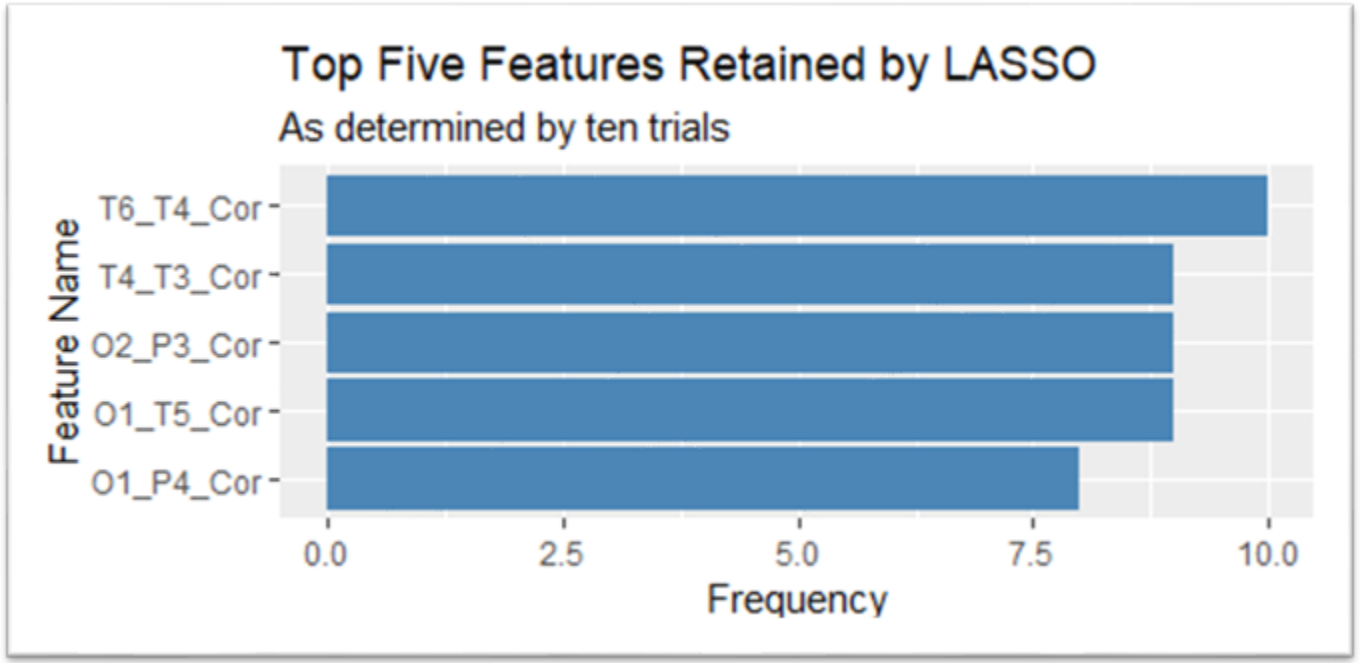
Results

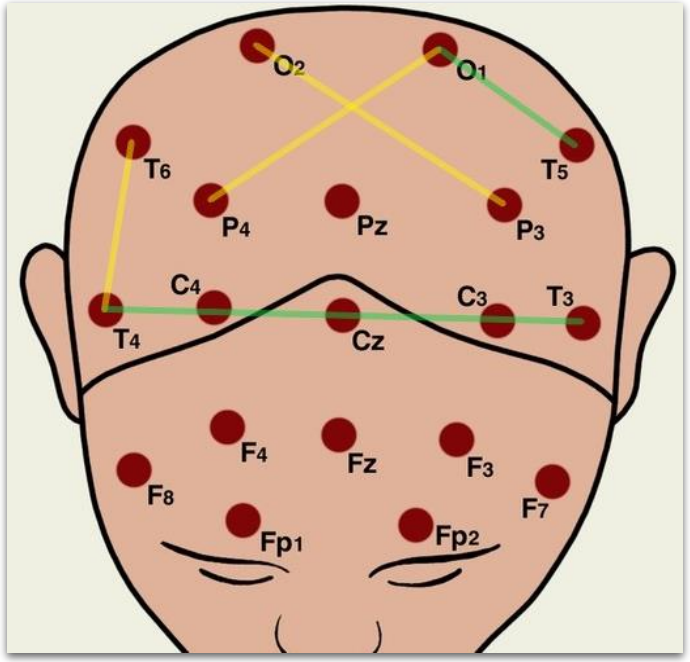
- The KNN model averaged 67% accuracy, LASSO averaged 71% accuracy

Feature Importance

- Regularized logistic regression used
- 10-fold cross validation with data partitioned in each fold using a stratified sub-sampling technique
- Cross validation used in each fold to tune regularized term choosing term that minimized the MSE
- Features with non-zero coefficients are recorded for significance

Feature Importance





Yellow = negative correlation with diagnosis, Green = positive correlation with diagnosis

Conclusions

Present Successes:

- We were able to confirm that correlation between certain areas of the brain is a significant factor in schizophrenia
- We were able to determine the specific areas of the brain where that correlation was significant
- We were able to rule out certain simple brain patterns as factors

Future Work:

- Feature Engineering: more in-depth waveform analysis
- Data: grouped by symptom, combined with other datapoints, combined with similar disorders, etc.
- Modeling: clustering to consider separating different disorders with similar symptoms

Q&A