Alzheimer's & Frontotemporal Dementia (FTD) Classification from EEG Data

Sandia National Laboratories
Brian Keith (bkeith9@gatech.edu)
2024-02-22

Georgia Tech

Project Background

What to Expect:

Information related to Sandia National Laboratories' (SNL) project scope, my twist on their prompt, as well as the motivation for the specific problem chosen.

Georgia Tech

Project Scope and Core Ideas

The Sandia National Laboratories' (SNL) project prompt revolved around using machine learning models to detect degradation in hardware components as they aged. The motivation for this is that physically testing components is time consuming and expensive.

This idea makes for a very interesting project and has numerous real-world use cases as well as approaches (*Shahraki et al., 2017*). While in SNL's introduction mostly focused hardware components, like machine bearings, the core concept of the project is to highlight how signals change over time and how we can use models to detect those changes.

Part of the reason I chose this project is that I have a background in exactly this concept. During my time with my previous employer, ANDRITZ, I worked on part of a team whose job was to do remote monitoring of KCF sensors installed on Pulp and Paper manufacturing equipment. This provided insight into the importance of models to detect issues before they occur, because not only does it reduce workloads for the on-site maintenance staff, but it can also allow for maintenance to be scheduled during planned downtime which translates cost-savings in a production environment.

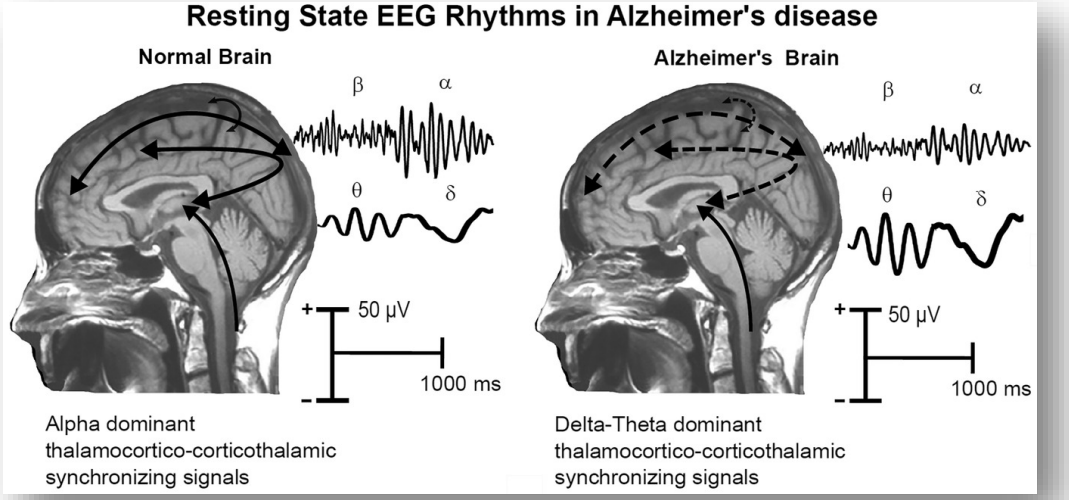
My Idea

While it would have been easy for me to take my existing domain knowledge and directly apply it to this project, I wanted to challenge myself and take the concept proposed by SNL to apply it in a way that perhaps SNL had not directly considered when creating the project.

As I researched and considered ways to scope this project, I had an idea, which may be intuitive to those with a medical background but felt like a revelation to myself. The human brain is really just one big, complex electrical circuit, so what if we could model patterns of electrical signals in the brain to detect neurodegenerative diseases?

As it turns out, this is a burgeoning area of research in the medical field and there are researchers hard at work trying to answer exactly this question.

After discussing and getting approval from the SNL project lead, Stephen Smith, I decided to move forward with this idea. While the idea does stray from the exact motivations outlined by SNL, it retains the core concept of using signal data to detect and utilize patterns that exist in the electrical signals for the purpose of detecting a fault or failure.



Tentative physiological model of the generation of resting state eyes-closed electroencephalographic (rsEEG) rhythms in the brain of age-matched old cognitively unimpaired (CU) persons and Alzheimer's disease (AD) patients (*Babiloni et al., 2021*).

Georgia Tech

Motivation

Before I provide specific details of the project, I believe it's important to discuss the motivation behind using models to diagnose neurodegenerative diseases. As, it differs in some ways from the direct monetary motivations laid out in SNL's project proposal.

There is a growing trend in Alzheimer's as well as other neurodegenerative diseases which are projected to continue increasing over the coming 30 years (*Steinmetz et. al., 2019*). I as well as many others reading this may have a loved one or friend who they have seen affected by these diseases.

When it comes to treatment of these diseases, while there are no current treatments that can reverse Alzheimer's disease, accurate and early diagnosis can help prepare families of the affected and extend the good quality of life years for individuals with the disease (*Rasmussen et. al., 2019*).

While any medical diagnosis should include a variety of factors such as consultations with medical professionals and clinical diagnostics, one proposed method to assist medical professionals in diagnosing neurodegenerative diseases is electroencephalography (EEG), which measures brain electrical activity. The signals from the EEG readings can then be used to automatically find patterns that might indicate the presence of neurodegenerative diseases using machine learning models (*Miltiadous et. al., 2023*).

https://www.jax.org/news-and-insights/2019/December/will-i-get-alzheimers

Georgia Tech

EEG Background

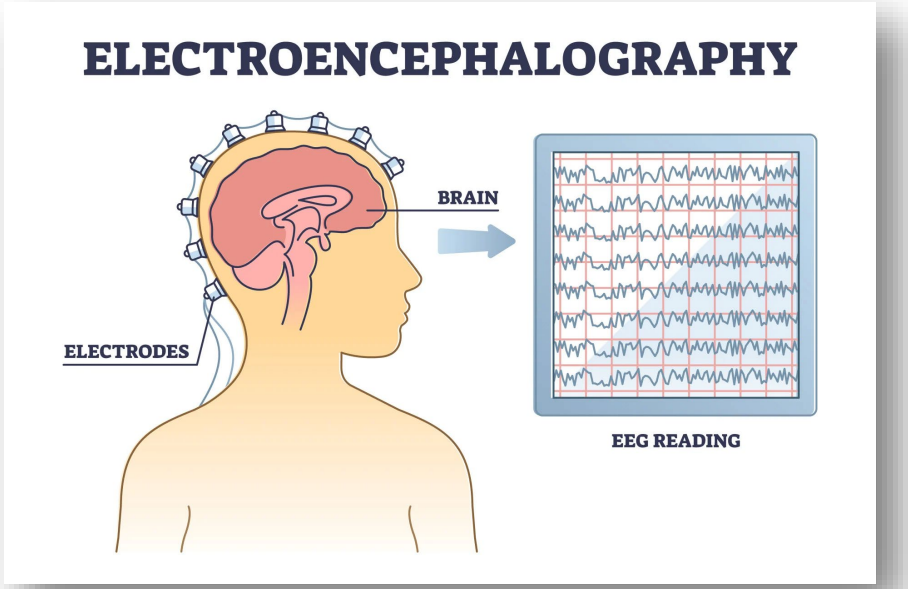
What to Expect:

A brief, high-level, overview of information on EEGs for those without any knowledge on what they are or how they're used to assist in understanding the subsequent sections.

Georgia Tech

What are EEGs?

- An electroencephalogram (EEG) is a medical test which utilizes electrodes attached to a patient's scalp to measure electrical activity in the brain (*Guy-Evans, 2023*). An EEG does not detect the activity of individual neurons, but instead measures the activity of small areas of the brain which can be used to indicate the level of activity in that region of the brain (*Id*).
- These tests are used to detect a variety of neurological phenomena such as epilepsy, strokes, dementia, and other brain disfunctions (*Id*).
- While the information related to the magnitude of the waves is important, one of the key features used to analyze EEG data is signal decomposition of the data into the 5 widely recognized brain waves (*Miltiadous et al., 2021*):



https://www.simplypsychology.org/what-is-an-eeg.html

Band	Frequency (Hz)
Delta (δ)	0.5 – 4
Theta (θ)	4 – 8
Alpha (α)	8 – 13
Beta (β)	13 – 25
Gamma (γ)	25 – 45

The signal data gathered via the EEG can be decomposed into these bands using techniques such as Fast Fourier Transforms to calculate the Power Spectral Density and Relative Band Power of each of the bands that makes up the overall signal.

Georgia Tech

Project Goals & Plan

What to Expect:

A high-level outline of the project's goals and plan. The progress through the project plan will be discussed in the subsequent section.

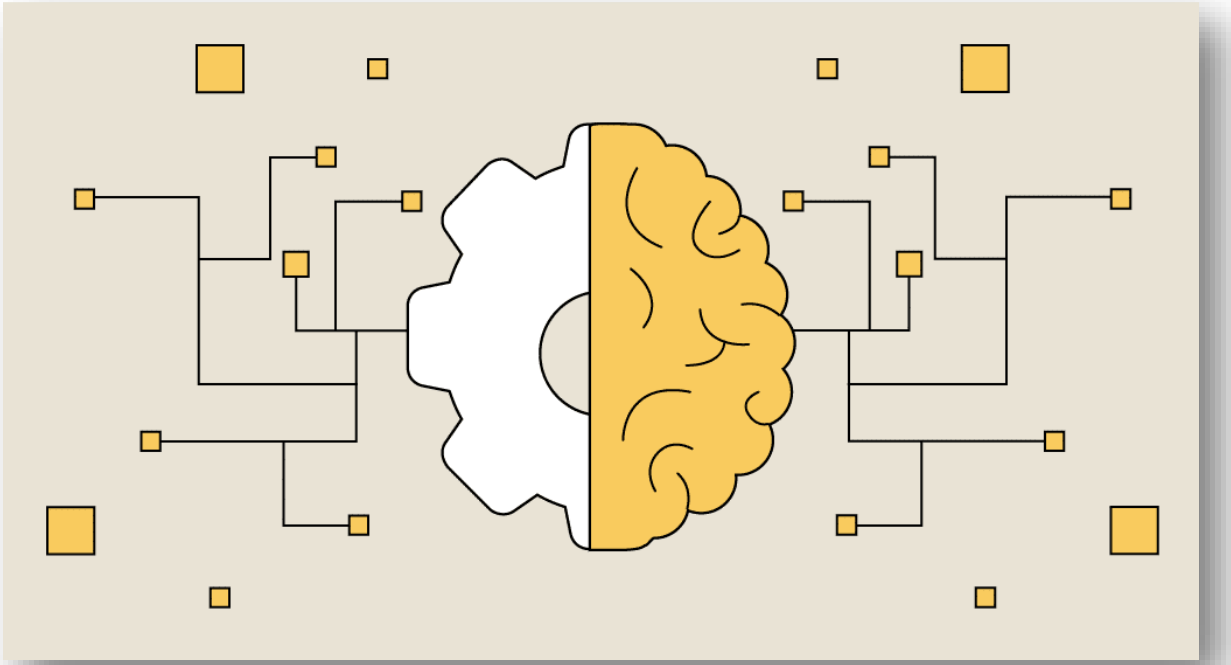
Georgia Tech

Project Goals

While I am not an expert when it comes to anything in the medical field, much less neuroscience (not exactly rocket science is it?), the goal of this project is to combine the basic of information that I can gather from research with advanced analytics and modeling techniques to develop a model which can accurately classify patients into the following categories:

- **Control – No Cognitive Impairment (C)**
- **Alzheimer's Disease (AD)**
- **Frontotemporal Dementia (FTD)**

My initial interest was in accurate classification of Alzheimer's Disease, however, the dataset (discussed later) also contained information related to FTD patients, so I pivoted my approach to include that disease as well.



Plan – Overview

Find Dataset

- Find a dataset with EEG readings. This is a classification problem, so the data will need labels to allow for supervised learning algorithms such as Random Forests.

Data Handling

- This step includes all the work necessary to take the dataset and prepare it for use in modeling including getting it into a format that can be processes efficiently by Python, dealing with any outliers, etc.

Feature Engineering

- Determine the input features for the models. This may involve techniques such as looking at how the features relate to the classifications for each patient, looking at the interactions of the features, and other techniques to determine the best set of parameters to use.
- This may also involve research related to EEGs and how others with more specialized domain knowledge have used the data to train similar models.

Modeling

- This step will be the bulk of the planned work and involves developing the models to classify patients into the different groups based on the EEG data.
- At its heart, the proposed problem is a classification problem, so I plan to explore multiple classification based models during the project such as Random Forests, Support Vector Machines, and Neural Networks.
- This step also includes the work of evaluating and comparing the model performances to determine which is the most appropriate for the task at hand.

Please note: This is only meant to generalize the overall process, the real process will likely be iterative and may involve additional steps. For example, different models may require different features and development of a particular model may have feature selection as part of the hyperparameterization process

Georgia Tech

Progress to Date

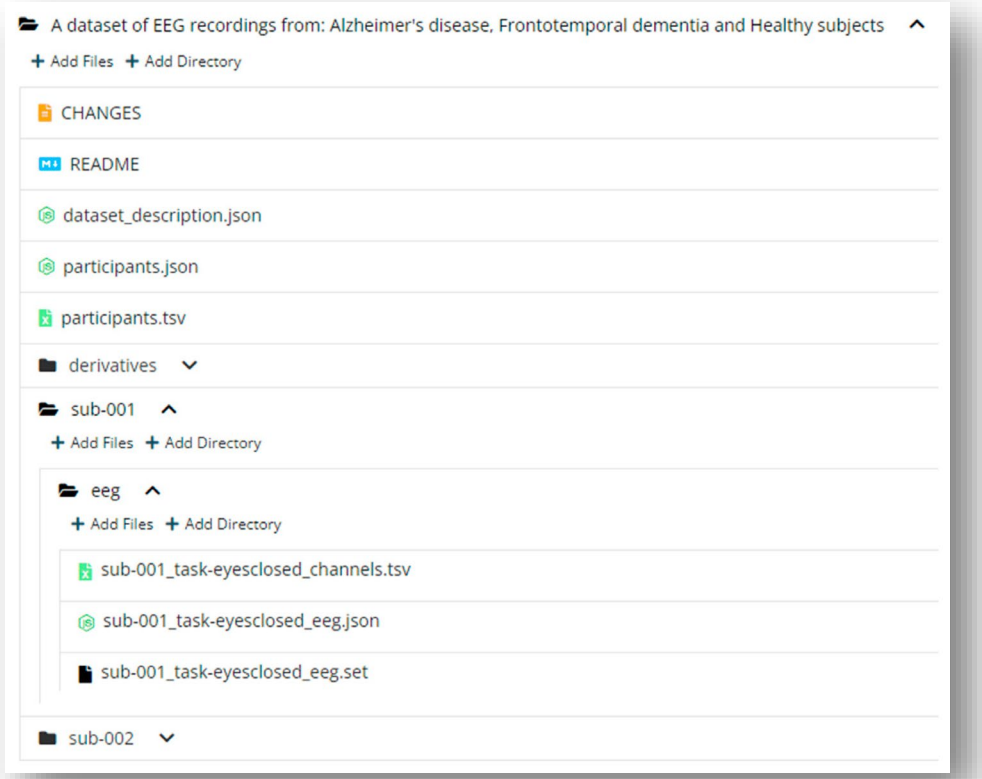
What to Expect:

An overview of the dataset that I'll be using for the project, the data handling and preprocessing performed to prepare the data, and the feature extraction and modeling that has been performed thus far.

Georgia Tech

The Dataset – Overview

- After researching potential datasets, I settled on a dataset from Greek researchers who published the dataset under a Creative Common CC BY license (*Miltiadous et al., 2023*).
    - The primary reason for my choice of this dataset was the researchers published the dataset with the goal of others using it for further research and therefore it was very well documented and some of the very domain specific preprocessing of the dataset had already been performed (more on that later).
- The data set includes EEG data from 88 patients, 36 with Alzheimer's Disease, 23 with Frontotemporal dementia, and 29 healthy control subjects.
- The data was provided as a ZIP file by the researchers with the folder structure shown in the image to the right. The structure, for the purposes of my analysis consisted of three different datasets:
    1. Information describing the patients, which was extracted from the *participants.tsv* file. This dataset provides a link between the dataset for a given participant and the classification of the patient, i.e., Alzheimer's Disease, Frontotemporal dementia, or healthy control as well as supplementary information such as their gender.
    2. The raw EEG data as *.set* files broken out by participant.
    3. The EEG data with some domain specific pre-processing to remove artifacts in the EEG readings such as eye movement, also provided as *.set* files broken out by participant.
- The two EEG datasets contained information from 19 different electrodes placed at various locations which were sampled at 500 Hz with a resolution of 10 μV/mm. Information related to the time in milliseconds for each sampled datapoint was also provided in the EEG data.



The structure of the dataset following the BIDS format (*Miltiadous et al., 2023*).

Georgia Tech

The Dataset – Loading & Consolidation

- For this analysis, I chose to use Python. Therefore, before doing anything, I needed to load and transform the data out of the *.set* files and into a more Python friendly format such as pandas.
- I will not go into the specifics of the methods used. However, essentially, I wrote some custom functions to scrape the directories for the files of interest, loaded the files with *scipy.io.loadmat()*, then looked for specific keys within the files to extract the data into pandas then concatenated all the different files into one large 35M row DataFrame.
  - During the process, I also had to design a custom recursive DataFrame explosion algorithm due to the structure of the input *.set* files.
- Once I had parsed the data into pandas, I performed some datatype correction and then exported the data to a *.pkl* file for later processing. An example of the *.info()* for the raw EEG dataset can be seen in the picture to the right.
  - Column #'s 3 - 21 are the labels associated with the 19 sensors mentioned previously.

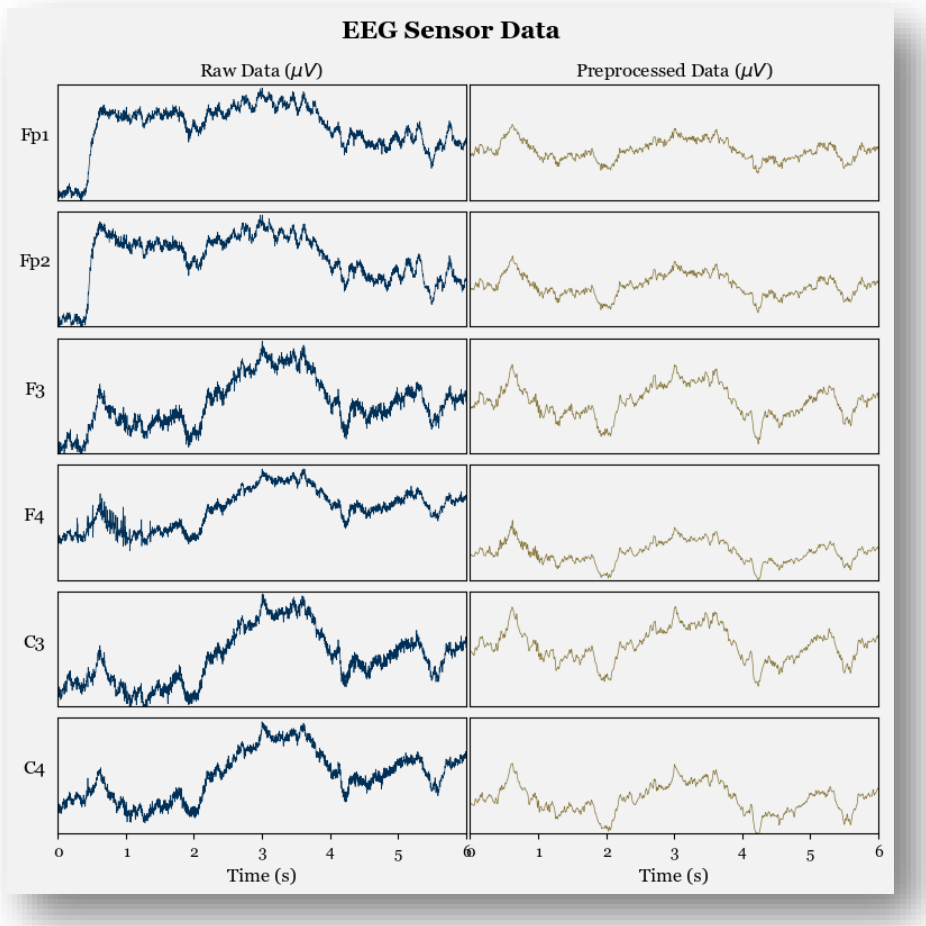
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35295150 entries, 0 to 35295149
Data columns (total 22 columns):
 #   Column          Dtype
---  ------          -----
 0   participant_id  category
 1   time_s          float64
 2   time_ms         int32
 3   Fp1             float32
 4   Fp2             float32
 5   F3              float32
 6   F4              float32
 7   C3              float32
 8   C4              float32
 9   P3              float32
 10  P4              float32
 11  O1              float32
 12  O2              float32
 13  F7              float32
 14  F8              float32
 15  T3              float32
 16  T4              float32
 17  T5              float32
 18  T6              float32
 19  Fz              float32
 20  Cz              float32
 21  Pz              float32
dtypes: category(1), float32(19), float64(1), int32(1)
memory usage: 2.9 GB
```

Example of the .info() returned for the Raw EEG data after data consolidation.

Georgia Tech

The Dataset – Raw or Preprocessed?

- One nuance of the data provided by the researches was that I needed to decide which of the two datasets to use, the raw EEG data or the data that was preprocessed to correct a variety of data issues such as:
  - Re-referencing the electrodes based on two reference electrodes
  - Frequency range correction
  - Performing artifact rejection routines related to eye and jaw artifacts
- While ideally, I would have liked to use the raw data from the EEG machine for my analysis, based on the research I have done related to the preprocessing performed by the researchers, I believe that some of the techniques needed to recreate the preprocessed data would outside of the scope of my capabilities due to domain specific technical knowledge required for some of the techniques.
- For this reason, I have decided to use the preprocessed data for my analysis. This does simplify some of the data wrangling work required, however, in the spirit of the project from SNL, there is still significant signal processing work to extract the features from the EEG signals which I will be performing myself.

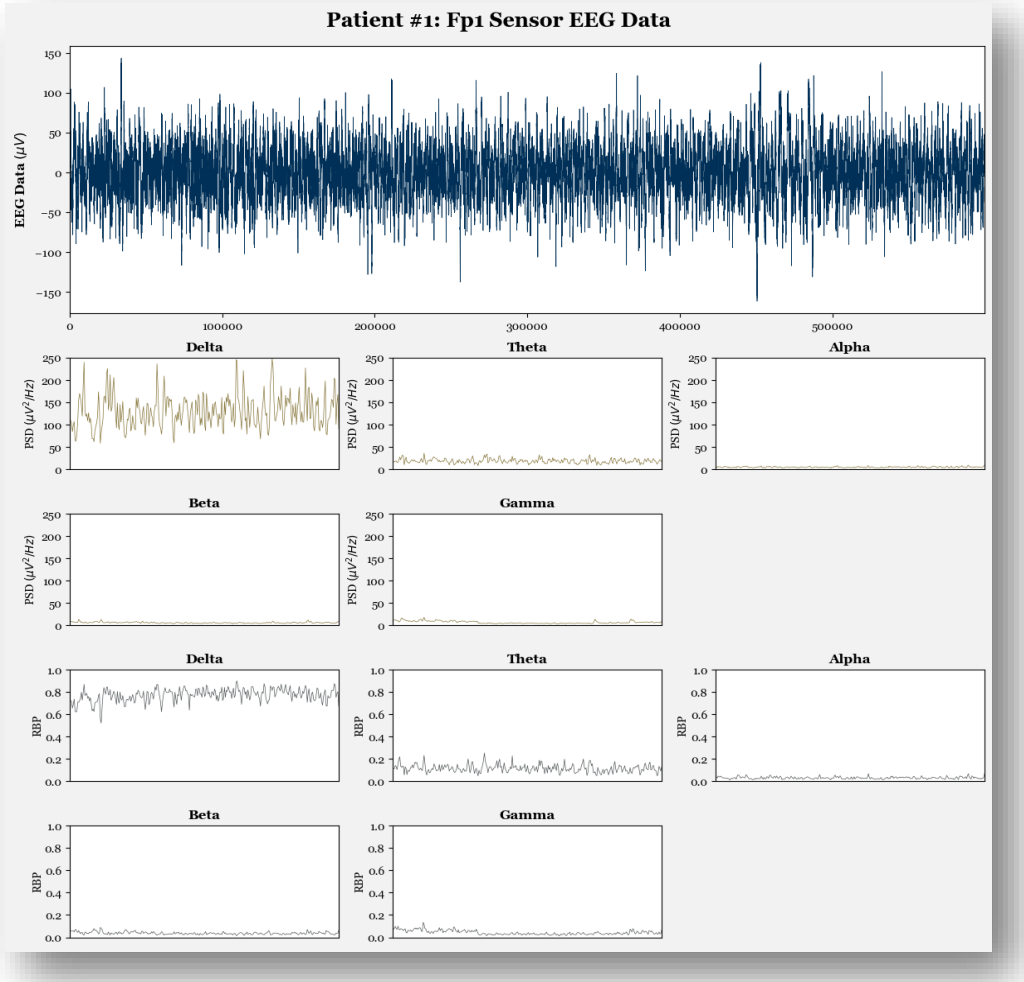


A visualization of the Raw vs. Preprocessed data for Participant #1 across a few sample sensors.

Georgia Tech

Feature Extraction

- Before I could begin any modeling of the data, I needed to preformed signal decomposition for each of the participants as well as for each of the 19 sensors.
- The idea of the signal decomposition is to take the signal data from the EEG and decompose the signal into the respective power spectral density (PSD) for the 5 different brain wave bands, Delta, Theta, Alpha, Beta, and Gamma. From the PSD, we can also calculate the Relative Band Power (RBP) of each band giving us a normalized metric between 0 and 1 for the contribution of a given PSD to the overall signal.
- While this is an area where I need to perform further research and iteration, for now, I am using Welsh's method (*Welsch, 1967*) to decompose the signal based on the frequencies shown on the "What are EEGs?" slide. One key point is that for the signal decomposition, I performed Welsh's method over 4 second segments with a 50% overlap in the input signal. This choice is currently relatively arbitrary and is based on the method used by the researchers. I will be experimenting with tuning this in the future.
- An example of this decomposition can be seen in the figure to the right for one participant and one sensor. This was applied to each of the sensors and participants giving me a total of 95 features (19 sensors x 5 bands) related to PSD and RBP, respectively (190 in total).

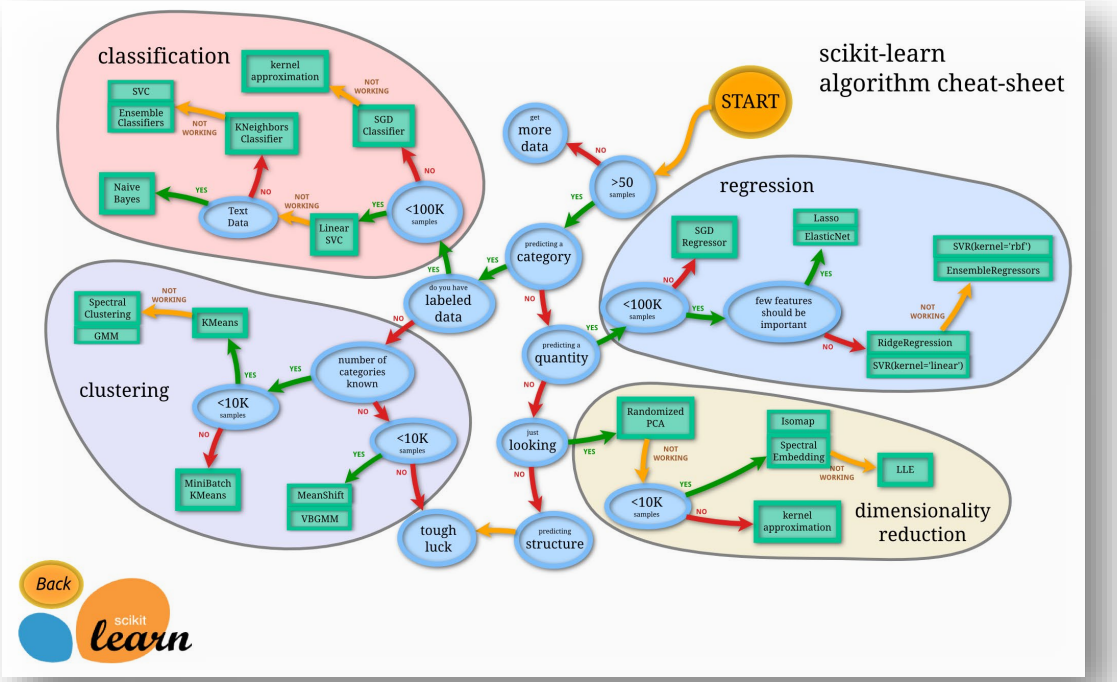


A visualization of the decomposed EEG data using Welsh's method.

Georgia Tech

Modeling

- I am just beginning to experiment with the modeling process, so I won't detail much information in this section. However, some of the initial tests I have done with non-hyperparameterized random forests have been promising with up to 90% overall accuracy.
- There are multiple considerations I need to think through, even when it comes to things such as how to split the data. My current thinking is that I will likely need to split the data based on the participants, and not based on rows since I want the model to wholistically consider the EEG data for a given participant. Since the number of participants in each group is relatively small, I'll likely need to employ cross validation techniques to ensure my model's performance characteristics are reliable.
- When it comes to modeling, I generally view the whole pipeline as part of the modeling process, so part of my plan is to not only test different classification models, but also to test different features, different methods of generating the features, different validation methods, etc. to arrive at the best possible model.



https://scikit-learn.org/stable/_static/ml_map.png

Remaining Work

What to Expect:

Summary of the work remaining for the rest of the semester.

Georgia Tech

Remaining Work Summary

Having settled on my dataset, if after delving deeper into the modeling I don't find a need for additional data, the bulk of the work remaining is in the modeling process. My plans for the remainder of the semester are to test a variety of different classification models, features, signal decomposition techniques, and cross validation methods.

More specifically, some of the current thoughts are:

- Features:
  - What is the best way to decompose the signal? Is the 4 second sampling interval with 50% overlap in the Welsch method the best, or did the researches choose it arbitrarily?
  - What are the best features to utilize? Does using all 190 features lead to a better model? Is only using RBP better since it's a normalized metric? Are there some sensors, which would be in different areas of the brain, more predictive than others?
- Models:
  - Which models should I try? There are obvious choices like Random Forests, SVM, KNN, and Logistic Regression, but could we get better results using neural networks, deep learning, or more advanced CART algorithms that utilize boosting/bagging?

It truly is a long road ahead, but I believe that the end-product will be a reliable model that is able to accurately distinguish between the different diagnostic groups.

https://upload.wikimedia.org/wikipedia/commons/thumb/7/74/The_Long_Road_Ahead.jpg/2560px-The_Long_Road_Ahead.jpg

Georgia Tech

Thank You!

Georgia Tech

References

Georgia Tech

References

1. Ameneh Forouzandeh Shahraki, Om Parkash Yadav, and Haitao Liao. A Review on Degradation Modelling and Its Engineering Applications [J]. Int J Performability Eng, 2017, 13(3): 299-314.
2. Babiloni et. al. (2021). Measures of resting state EEG rhythms for clinical trials in Alzheimer's disease: Recommendations of an expert panel. Alzheimer's & Dementia, 17(9), 1528–1553. https://doi.org/10.1002/alz.12311
3. Steinmetz et. al. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. The Lancet Public Health, 7(2), e105–e125. https://doi.org/10.1016/s2468-2667(21)00249-8
4. Rasmussen, J., & Langerman, H. (2019). Alzheimer's Disease – Why We Need Early Diagnosis Degenerative Neurological and Neuromuscular Disease, Volume 9, 123–130. https://doi.org/10.2147/dnnd.s228939
5. A. Miltiadous et al. , "Machine Learning Algorithms for Epilepsy Detection Based on Published EEG Databases: A Systematic Review," in IEEE Access, vol. 11, pp. 564-594, 2023, doi: 10.1109/ACCESS.2022.3232563
6. Olivia Guy-Evans (2023, September 19). EEG Test (Electroencephalogram): Purpose, Procedure, And Risks. Simply Psychology. https://www.simplypsychology.org/what-is-an-eeg.html
7. Abhang, P. A., Gawali, B. W., & Mehrotra, S. C. (2016). Technological basics of EEG recording and operation of apparatus. In Elsevier eBooks (pp. 19–50). https://doi.org/10.1016/b978-0-12-804490-2.00002-6
8. Miltiadous, A., Tzimourta, K. D., Γιαννακέας, N., Tsipouras, M. G., Afrantou, T., Ioannidis, P., & Tzallas, A. T. (2021). Alzheimer's Disease and frontotemporal Dementia: A robust classification method of EEG signals and a comparison of validation methods. Diagnostics, 11(8), 1437. https://doi.org/10.3390/diagnostics11081437
9. Andreas Miltiadous and Katerina D. Tzimourta and Theodora Afrantou and Panagiotis Ioannidis and Nikolaos Grigoriadis and Dimitrios G. Tsalikakis and Pantelis Angelidis and Markos G. Tsipouras and Evripidis Glavas and Nikolaos Giannakeas and Alexandros T. Tzallas (2023). A dataset of EEG recordings from: Alzheimer's disease, Frontotemporal dementia and Healthy subjects. OpenNeuro. [Dataset] doi: doi:10.18112/openneuro.ds004504.v1.0.6
10. P. Welch, "The use of the fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms", IEEE Trans. Audio Electroacoust. vol. 15, pp. 70-73, 1967.

Georgia Tech