Classification of Alcoholics from EEG Signals

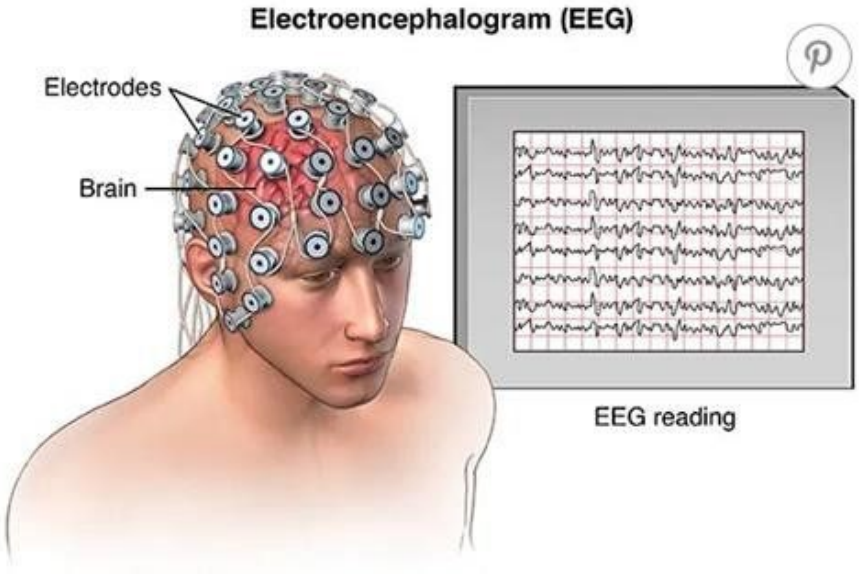
Azeezat Azeez and Miriam Sierra

June 2022 Erdös Institute Data Science Bootcamp

GitHub - ErdosInstitute-Predict-who-is-alcoholic/Predict-who-is-alcoholic

Contents

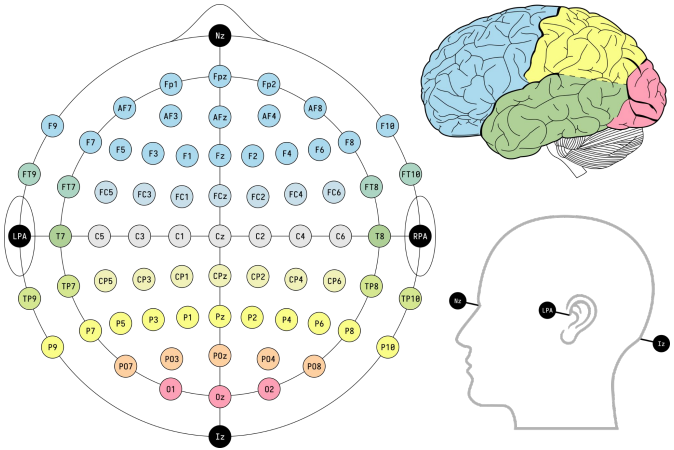
- Background and Motivation
- Methods
- Concluding Remarks



Background: Electroencephalography

The EEG(Electroencephalography) is a non-invasive neuroimaging technique used to record electrical activity from the scalp, representative of real time neural activity at high temporal resolution.

In Clinical setting EEG is primary used to diagnose Epilepsy



Project Description: Predict who is an Alcoholic?

Data

Open-source EEG data from the UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/EEG+Database

Description

This data arises from a large study to examine EEG correlates of genetic predisposition to alcoholism. It contains measurements from 64 electrodes placed on subject's scalps which were sampled at 256 Hz (3.9-msec epoch) for 1 second.

Subjects

Two groups:

- Alcoholic
- Control

Sets

Train set:

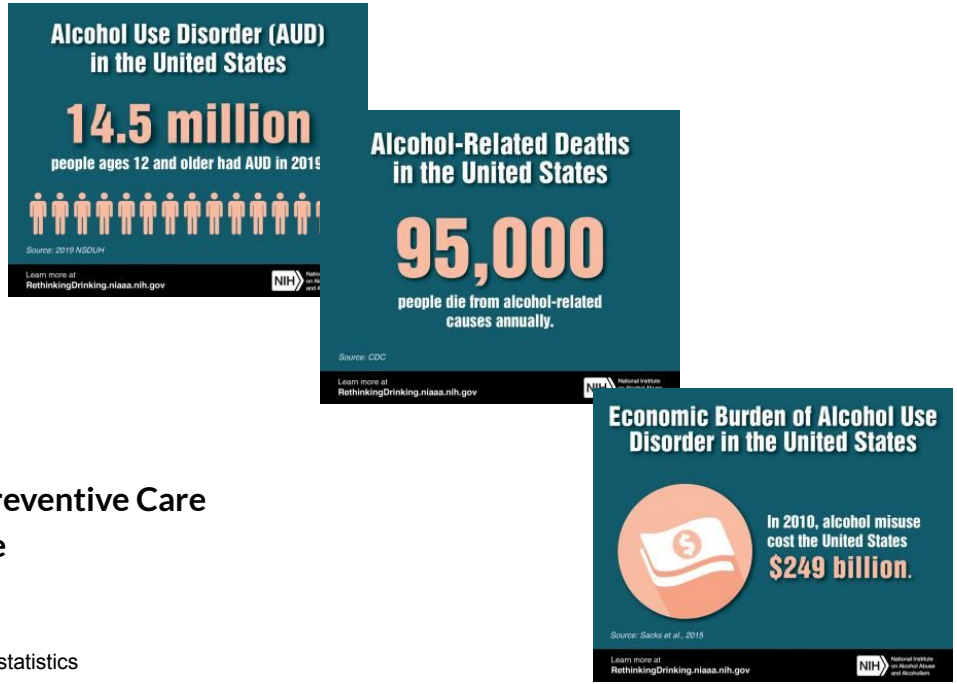
- 20 unique patients (10 control/10 alcoholic)
- 30 trials per patient

Motivation: Preventive Care & Improved Social Welfare

14.5 M people suffer from alcohol use disorder in the US

Current method of diagnosis is largely <u>qualitative</u> (self reports) which can introduce bias

The <u>quantitative</u> ability to diagnose discover high risk individuals based on biomarkers of EEG data has real world benefits:



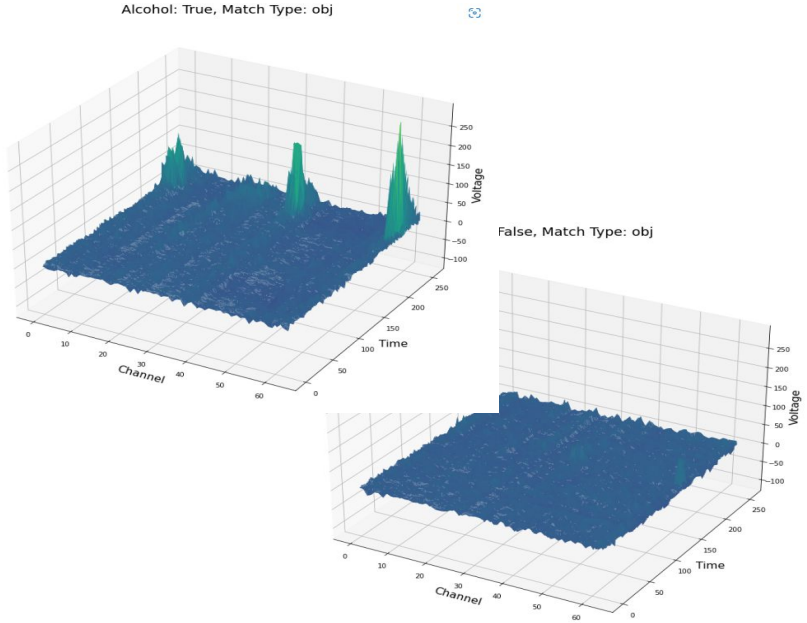
- Medical: Identify high risk and vulnerable populations→ **Preventive Care**
- Social: Reduce the social burden → **Improve Social Welfare**
- Financial: **Preventive Care + Improve Social Welfare**

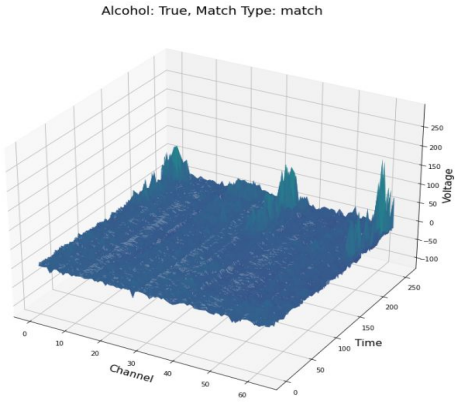
https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics

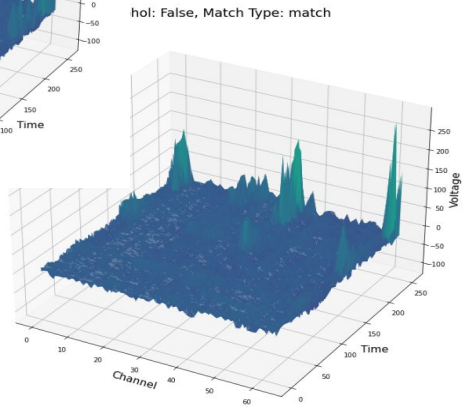
Approach: Classification Algorithms

1. Exploratory Analysis
2. Preprocessing: EEG data using Literature Standard
3. Feature Extraction and Dimension Reduction
4. Machine Learning Algorithms
    a. Supervised Learning (Hyperparameter tuning of best classifier)
        i. Logistic Regression
        ii. Random Forest
        iii. Support Vector Machine
    b. Neural Networks: Classification using deep 1D convolutional neural network

Exploratory Analysis







Preprocessing

The frequency of EEG signals range from 0.01 Hz to around 100 Hz, which can be divided into five frequency bands

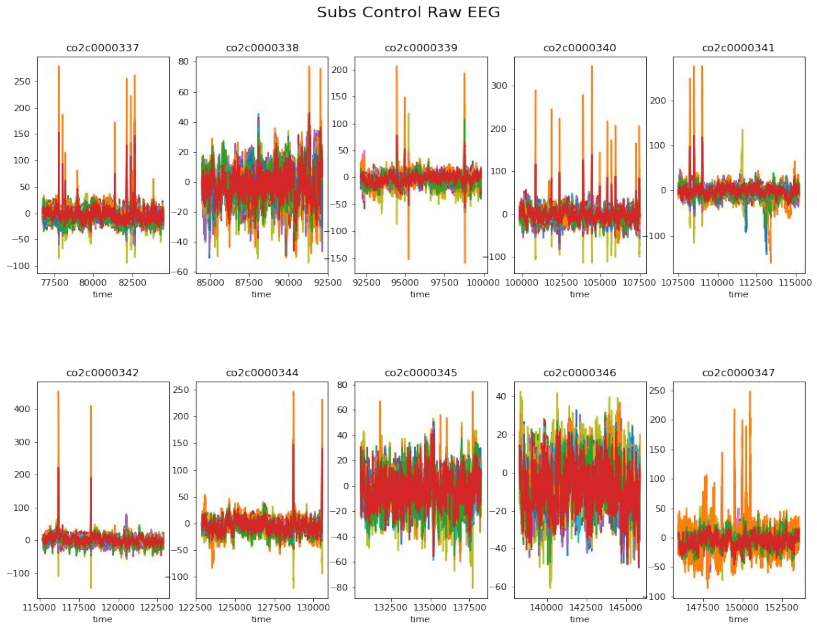
Band Name	Frequency (Hz)	Interpretation
Delta	<4	Deep sleep
Theta	4–8	Relaxed state and meditation
Alpha	8–13	Relaxed state of consciousness
Beta	13–30	active thinking

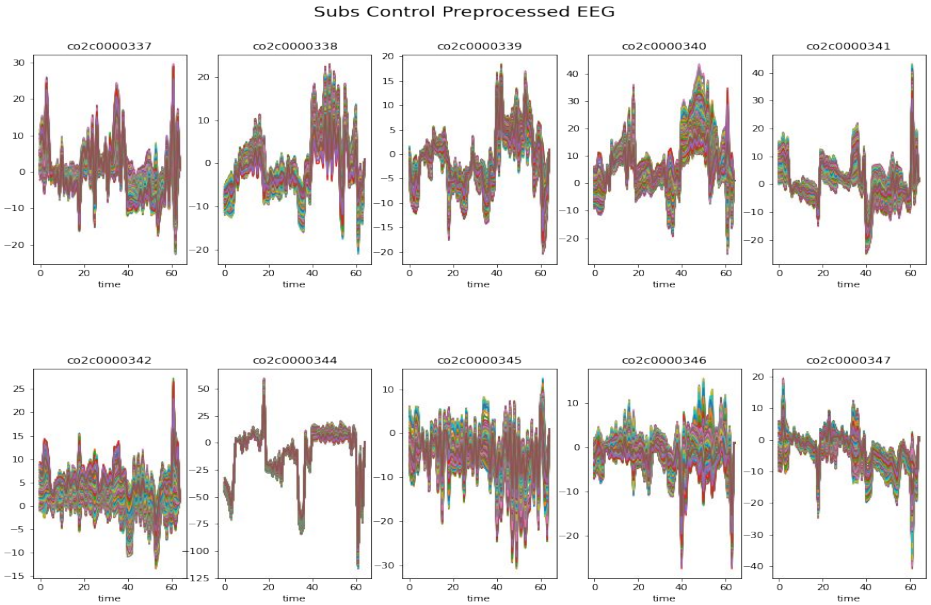
MNE Python Package: MEG and EEG data analysis toolbox

1. Bandpass filtering [0.01- 40Hz]
2. Epoch Data by trial
3. Remove bad Epochs

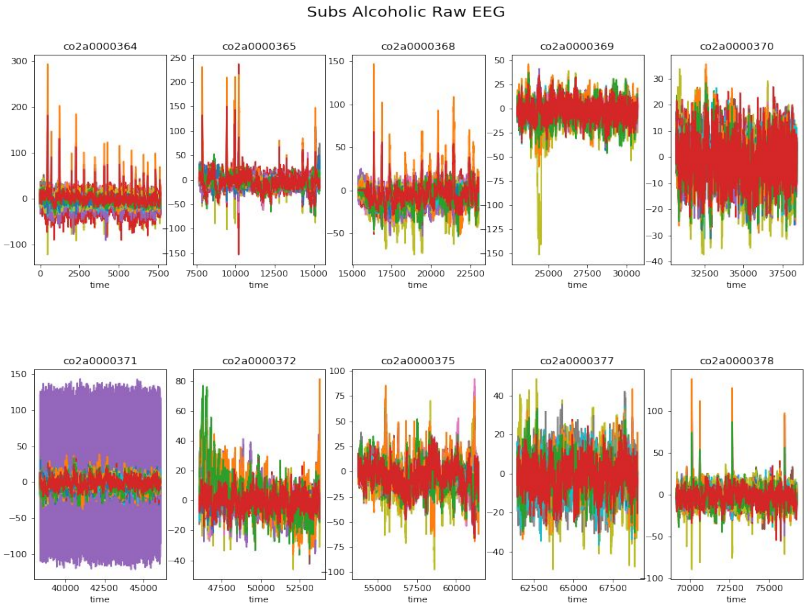
Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7(267):1–13, 2013. doi:10.3389/fnins.2013.00267.
Jiang, Xiao, Gui-Bin Bian, and Zean Tian. "Removal of artifacts from EEG signals: a review." Sensors 19.5 (2019): 987.

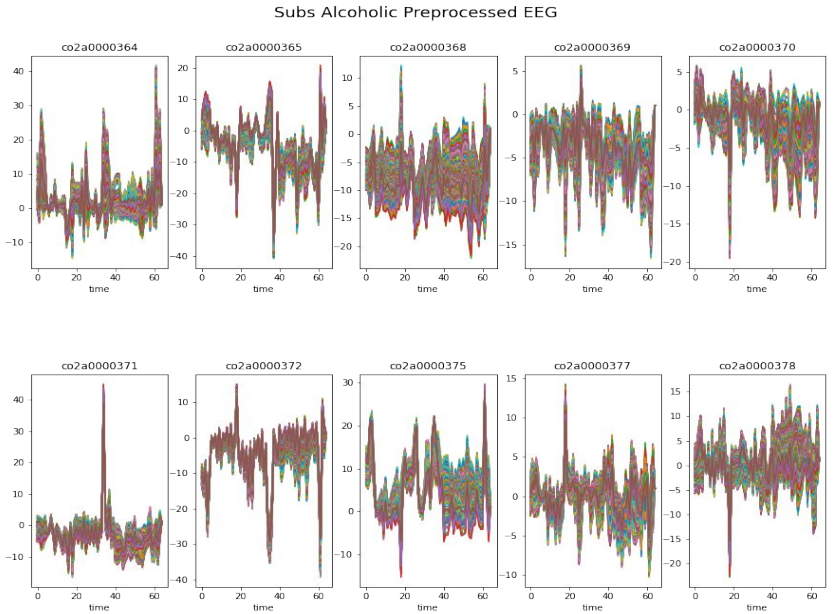
Raw Data Vs Preprocessed Data





Raw Data Vs Preprocessed Data





Feature Extraction & Dimension Reduction

For each subject we extract 12 common Statistical Metrics to reduce subject wise dimension(30 trails*65 channels):

Mean, Std, ptp, Variance, Minimum, Maximum, rms, abs_diff_signal, Skewness, kurtosis, argminim, argmaxim

Feature Importance

Random Forest

	Feature	Importance Score
43	P1	0.030397
28	FC2	0.023847
39	FZ	0.023237
12	CP2	0.022846
56	PZ	0.020928

Extra Trees

	Feature	Importance Score
43	P1	0.037095
28	FC2	0.036153
39	FZ	0.033858
12	CP2	0.031603
55	POZ	0.024527

Model Assessment and Selection

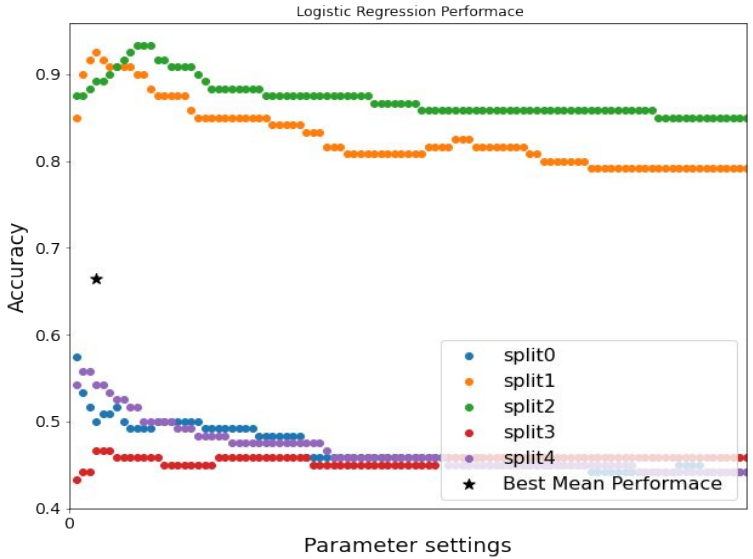
	Logistic Regression		SVC		Random Forest	
	64 Channels	5 Channels	64 Channels	5 Channels	64 Channels	5 Channels
Accuracy	65.2%	65.5%	67.2%	66.3%	70.7%	67.5%
Precision	66.4%	65.3%	68.3%	65.2%	72.0%	67.8%
Recall	69.0%	68.3%	69.0%	79.3%	68.0%	71.7%
f1	66.8%	66.6%	67.9%	68.3%	67.5%	68.8%
ROC AUC	72.2%	68.2%	72.6%	68.6%	74.3%	71.7%

Classification: Logistic Regression

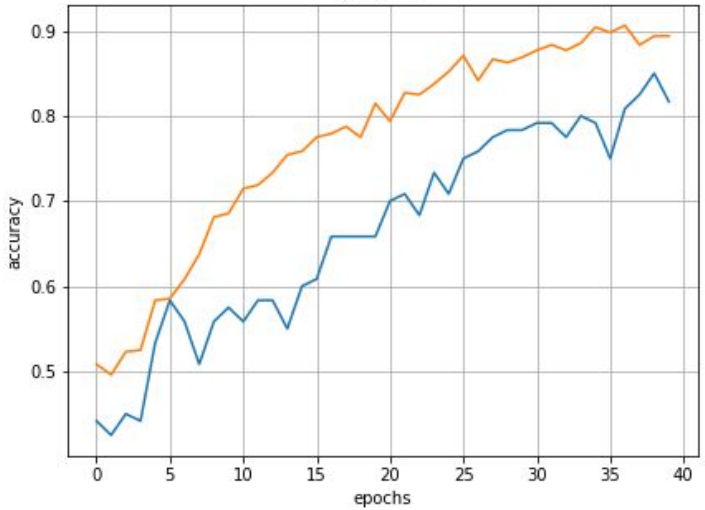
Param_grid:[0,1,0.0001]

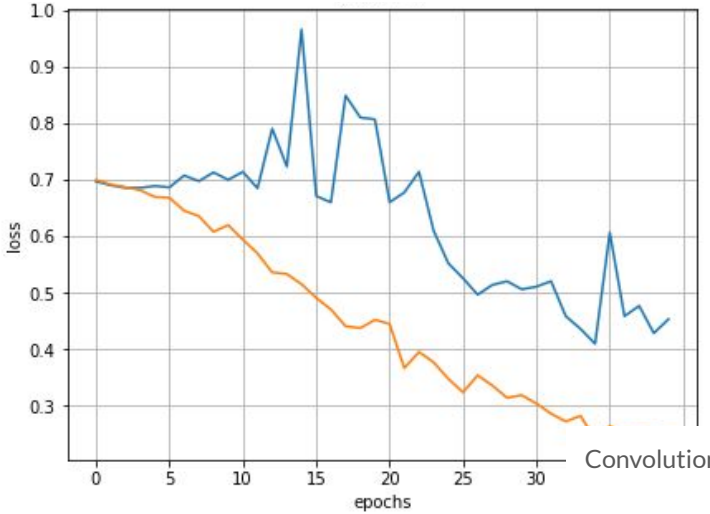
k-Fold Cross Validation: n=5

Average Best Performance: 0.665 at 0.0004



Classification using deep 1D convolutional neural network





Convolution and Pooling Sequential Model

- Total params: 1,326
- Trainable params: 1,316
- Non-trainable params: 10

Summary

The intersection of Data Science and Medical Imaging offer a unique opportunity to create clinical biomarkers for Patients

In this Case the potential to offer Preventive Care & Improved Social Welfare for those suffering from Alcohol Use Disorder

While our current model performance is promising, there is still room to improve the diagnostic utility of the Algorithms.

Thank You Erdös Institute!