Introduction

Post-traumatic stress disorder (PTSD) is a disorder that develops in some people who have experienced a shocking, scary, or dangerous event (https://www.nimh.nih.gov). Post-concussive syndrome (PCS) describes the constellation of symptoms that commonly occur after mild traumatic brain injury (TBI), and patients who suffer more than one brain injury are at increased risk (https://www.ncbi.nlm.nih.gov). It is highly common in individuals who have a traumatic brain injury, especially combat veterans. Such subjects tend to show overlapping symptoms of both PCS and PTSD.

In this project, we are going to use a dataset that contains signals between two brain regions that are obtained from brain images (scanned in Auburn University – MRI Center). 139 active-duty male US Army soldiers were recruited for this study from Fort Benning, GA, and Fort Rucker, AL, USA. All groups were matched for age, race, education, and deployment history.

Participants were scanned at the Auburn University MRI Research Center using a Siemens 3T MAGNETOM Verio Scanner with a 32-channel head coil. Resting-state data were collected using a T2*-weighted multiband echo-planar imaging (EPI) sequence with parameters: TR = 600 ms, TE = 30 ms, FA = 55°, multiband factor = 2, voxel size = 3 × 3 × 5 mm^3, and 1000 time points per run. Brain coverage included the cerebral cortex, subcortical structures, midbrain, and pons, but excluded the cerebellum. Participants kept their eyes open, fixating on a white cross against a dark background during scanning.

After extracting the time series, functional connectivity (FC) among the 200 regions was computed using Pearson's correlation coefficient for all region pairs, resulting in 19,900 FC values. These values were then utilized as features for the classification process. Since the whole-brain coverage was unavailable, time series were obtained from only125 regions for PTSD, leading to a correspondingly reduced number of FC connections for this dataset.

MRIs and other tools are common methods for collecting information on the brain, such as brain waves and neural activity, to understand such conditions better in a physical manner. Often traumatic conditions on the brain (or affecting it) are profound yet not easy to observe directly. The goal of the project is to use classification techniques to automate the detection of these mental health problems using brain signals.

For the other aspect of the project, since this is a high-dimensional dataset where there are more variables than observations, we also would like to account for the feature importance and would like to extract the brain regions/lobes associated with PTSD.

Data processing:

As the main goal of the project is the detection of PTSD via brain image data, we combine PTSD and PCS patients as one group. Therefore, the response variable becomes control group (0) and the affected group (1). The number of healthy individuals (control) is 45, whereas the number of patients is 94. In order to be able compare different models, we split our dataset into 2: training set and validation set by using the 80/20 proportions.

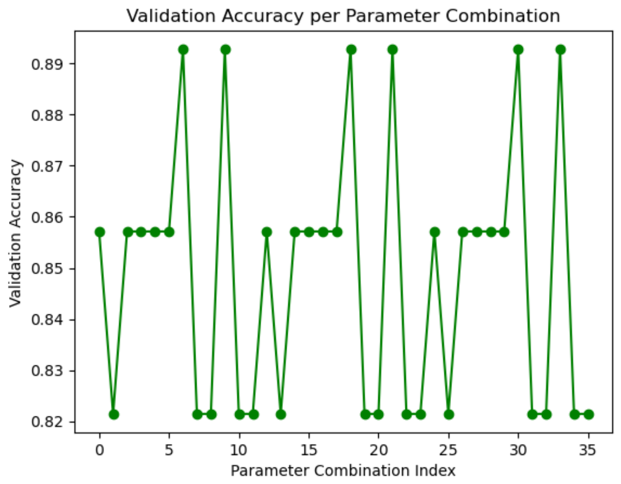
Model fitting:

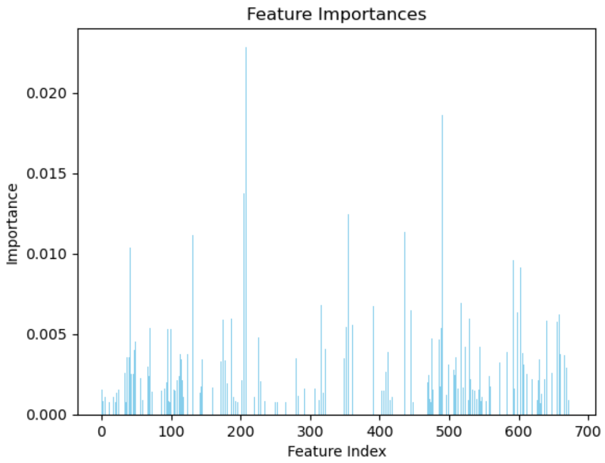
Random forest and lasso logistic regression are the most acceptable machine learning techniques in binary classification. In addition to classification, they also manage to provide feature importance by using different measures such as Gini index and penalization. That's why we employ these 2 techniques for our project.

For random forest, we tried different parameters as shown below.

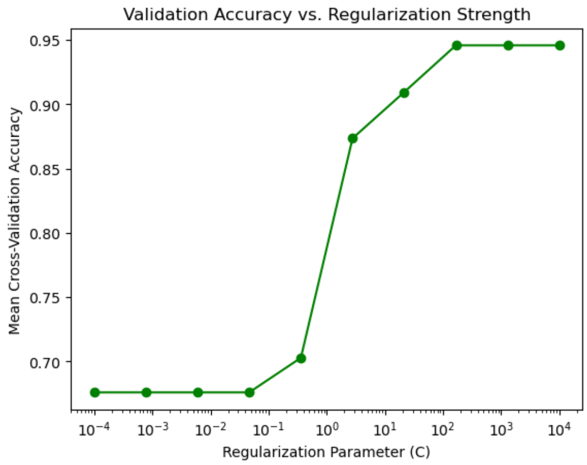
- The minimum number of samples required to be at a leaf node: [**1**,2]
- The minimum number of samples required to split an internal node: [2,**5**]
- The number of trees in the ensemble: [**50**,100,150]

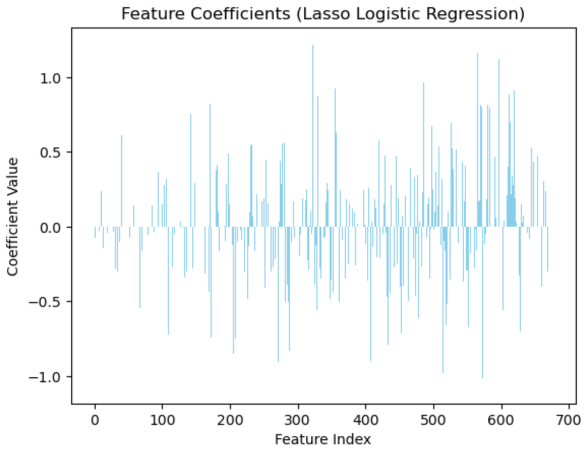
The highest accuracy is obtained by the parameters that are bolded above. The validation accuracy is found to be 0.857. The following plots show the different combinations of parameters and the feature importance of the final random forest model.





For the lasso logistic regression model, we use 5-fold cross-validation, and the validation accuracy is found to be 0.949. Plots corresponding to lasso logistic model can be found below. It is worth noting that the feature importance plots are not necessarily informative due the excessive amount of variables.

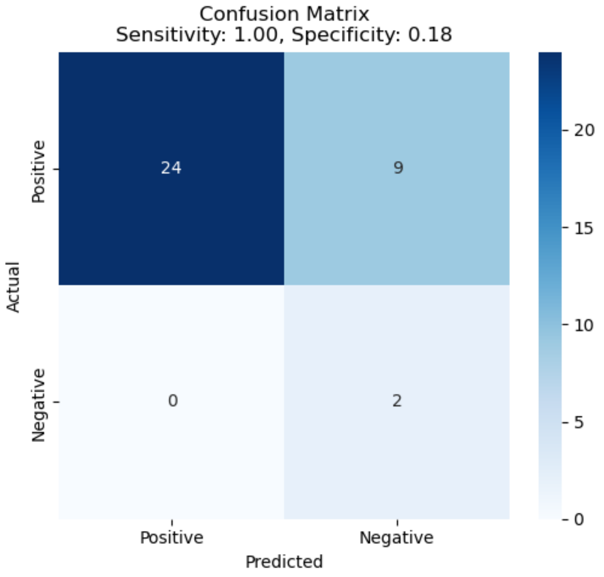




Before finalizing the model, we checked the variables (189) that are selected by both random forest and lasso logistic model: The following table shows an aggregated summary of the results. As each variable represents the signals between two brain regions, we prefer focusing on the brain lobes/regions that contain the most selected signals by both models. Also, one should remember that there are too many to mention each.

Brain region	Number of times of selection
Frontal	26
Temporal	17
Occipital	7
Cingulum	5
Lingual	5
Caudate	4
Insula	4
Postcentral	4
Precentral	4
SupraMarginal	4
Cuneus	3
Fusiform	3
Rolandic	3
Angular	2
Calcarine	2
Parietal	2
Thalamus	2
Vermis	2

Now we can pick our final model. Since validation accuracy is better in the lasso model, we decide to use that model as our final model in prediction. We have a test dataset that has 35 observations and is separated from the dataset that is used for training. The confusion matrix can be found in the figure below.



The test accuracy of the model is 0.74, which is not great for detection of PTSD. However, we can observe that the sensitivity of the model is 1, which is the highest possible value. This shows the power of the model in predicting the positive cases.

Conclusion:

We use high dimensional brain image data to automate the detection of PTSD and to extract the important brain regions. Since the aim is to detect, we aggregate the PCS and PTSD. For feature importance purposes, random forest and lasso logistic are employed. The temporal and the frontal lobes are found to be the most important features as well as occipital, cingulum and lingual which have been reported to have alterations in PTSD before (Yin et al. 2012; Zhang et al. 2016; Liu et al. 2015). This can give doctors an insight of the importance of the brain signals, specifically between frontal and temporal lobes. Lasso logistic model has been selected as the final model with a 74% test accuracy. The model has the highest sensitivity which shows perfect detection of the positive cases.

References:

1. Lanka, Pradyumna, et al. "Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets." *Brain imaging and behavior* 14 (2020): 2378-2416.

2. Yin, Y., Jin, C., Eyler, L., Jin, H., Hu, X., Duan, L., et al. (2012). Altered regional homogeneity in post-traumatic stress disorder: a restingstate functional magnetic resonance imaging study. Neuroscience Bulletin, 28(5), 541–549.

3. Zhang, Q., Wu, Q., Zhu, H., He, L., Huang, H., Zhang, J., & Zhang, W. (2016). Multimodal MRI-based classification of trauma survivors with and without post-traumatic stress disorder. Frontiers in Neuroscience, 10, 292.

4. Liu, F., Xie, B., Wang, Y., Guo, W., Fouche, J.-P., Long, Z., Wang, W., Chen, H., Li, M., Duan, X., Zhang, J., Qiu, M., & Chen, H. (2015). Characterization of post-traumatic stress disorder using resting-state fMRI with a multi-level parametric classification approach. Brain Topography, 28, 221–237.