# Interpretable Deep Learning for Diagnosis of Human Brain Disorders using Neuroimaging

by

Janzaib Masood

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
December 12, 2020

Keywords: Neuroimaging, Deep Neural Networks, Explainability

Approved by

Gopikrishna Deshpande, Professor of Electrical and Computer Engineering
Thomas S. Denney, Professor of Electrical and Computer Engineering
Jeffery S. Katz, Professor of Psychological Sciences

# Abstract

Deep neural networks are increasingly being used in neuroimaging research for the diagnosis of brain disorders and understanding of human brain. They are complex data driven systems that work in a black-box fashion once they are trained. Despite their impressive performance, their applicability in medical applications will be limited unless there is more transparency on how these algorithms arrive at their decisions. Interpretability algorithms help in bringing transparency in the decision-making of deep neural networks.

In this work, we combined the understanding generated by multiple interpretable deep learning algorithms, to explain our resting state functional connectivity classifiers. Convolutional neural network classifiers were trained for discriminating between patients and healthy subjects; we worked with post-traumatic stress disorder (PTSD), autism spectrum disorder (ASD) and Alzheimer,Äôs disease. We used the variants of gradient and relevance-based interpretability algorithms. Permutation testing and cluster mass thresholding was used to identify the significant discriminating functional connectivity paths between patients and control subjects.

For PTSD, the classifier provided >90% accuracy and we found that different interpretability algorithms gave slightly different results, most likely because they assume different things about the model and data. By taking a consensus across methods, the interpretability became more robust and was found to be in general agreement with prior literature on connectivity alterations underlying PTSD. For ASD, classification performance, and hence interpretability, varied widely across data acquisition sites (56% - 94%). Harmonization of data across sites provided incremental improvement in accuracy, but not enough to make interpretability largely consistent. We found that interpretability makes sense only for some sites that provide high enough accuracy. Our results demonstrate that robust interpretability across methods and models requires substantially higher accuracy than is currently possible in many neuroimaging datasets. This should be a cautionary tale for researchers wanting to use interpretability of artificial neural network classifiers in neuroimaging.

# Acknowledgments

I am immensely thankful to my advisor Gopikrishna Deshpande, for being extremely supportive during the course of this work and helped me improve my abilities in statistical analysis. I would also like to express my true gratitude to Professor Thomas Denney and Professor Jeffery Katz for evaluating my thesis. I am very thankful to all of my teachers at Auburn and in Balochistan. I cannot forget the huge support from Pradyumna Lanka and Rangaprakash Deshpande, who shared their preprocessed MRI data that I used in this work. I would also like to share my sincere gratitude to Alex Ing, for sharing and helping me understand his work in cluster focused permutation testing.

I also appreciate the effort of Julie Rodiek who is always available to help with everything at the MRI Research Center.

This work would not have been wrapped up successfully without the support of my parents, brothers and friends throughout the process.

## Table of Contents

Abstract . . . . . . . . . . ii

Acknowledgments . . . . . . . . . . iv

List of Abbreviations . . . . . . . . . . xii

1 Introduction . . . . . . . . . . 1

1.1 Motivation . . . . . . . . . . 1

1.2 Introduction . . . . . . . . . . 2

1.3 Thesis Organization . . . . . . . . . . 4

2 Literature Review . . . . . . . . . . 5

2.1 Resting State Functional Connectivity . . . . . . . . . . 5

2.2 Brain Disorder Diagnosis and Functional Connectivity . . . . . . . . . . 6

2.2.1 Autism Spectrum Disorder . . . . . . . . . . 6

2.2.2 Alzheimer's disease and Mild Cognitive Impairment . . . . . . . . . . 6

2.2.3 Post-traumatic stress disorder . . . . . . . . . . 7

2.3 Deep Learning for Brain Disorder Classification . . . . . . . . . . 7

2.4 Interpretability of Deep Learning . . . . . . . . . . 8

3 Dataset and Preprocessing . . . . . . . . . . 11

3.1 Dataset Description . . . . . . . . . . 11

3.2 Preprocessing . . . . . . . . . . 12

4 Deep Neural Network Classifiers . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
4.1 Selection of the Neural Network Classifier and Associated Parameters . . . . . 14
4.1.1 Convolutional Neural Networks . . . . . . . . . . . . . . . . . . . . . . . 16
4.1.2 Training a Deep Neural Network Model . . . . . . . . . . . . . . . . . . 17
4.2 Structuring and Training Brain Disorder Classifiers . . . . . . . . . . . . . . . . 18
4.2.1 Post-traumatic Stress Disorder and Alzheimer's Disease Models . . . . 21
4.2.2 Autism vs Controls Model . . . . . . . . . . . . . . . . . . . . . . . . . 22
4.3 Saving the models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
5 Generating Explanations of Classifier Decisions . . . . . . . . . . . . . . . . . . . . 25
5.1 Interpertable Machine Learning . . . . . . . . . . . . . . . . . . . . . . . . . . 25
5.1.1 Gradient Heatmaps . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
5.1.2 Smoothgrad . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
5.1.3 Input ◦ Gradient . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
5.1.4 Integrated Gradient . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
5.1.5 Guided Backgpropagation (GBP) . . . . . . . . . . . . . . . . . . . . . . 27
5.1.6 Deconvnet Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
5.1.7 Layer-wise Relevance Propagation-z (LRP-z) . . . . . . . . . . . . . . . 28
5.1.8 Layer-wise Relevance Propagation-Epsilon . . . . . . . . . . . . . . . . 30
5.1.9 Layer-wise Relevance Propagation ($\alpha$, $\beta$) . . . . . . . . . . . . . . . . 31
5.2 Individual Heatmaps versus General Heatmaps . . . . . . . . . . . . . . . . . 31
6 Statistical analysis and Consensus Heatmaps . . . . . . . . . . . . . . . . . . . . . . 36
6.1 Permutation Test with Difference of Means Statistic . . . . . . . . . . . . . . . 37
6.1.1 Correcting for Multiple Comparisons . . . . . . . . . . . . . . . . . . . 37
6.1.2 Analyzing PTSD CNN Model using Permutation Test . . . . . . . . . . 39
6.2 Permutation Test with Cluster Mass Statistic Threshold . . . . . . . . . . . . . 40

6.2.1 Analyzing the PTSD Heatmap using CMS Permutation Method . . . . 43

6.3 Consensus Heatmaps . . . . . . . . . . . . . . . . . . . . . . . . . . . 43

7 Multiple Sites, Feature Size and Interpretability . . . . . . . . . . . . . . . . . . . . . 49

8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

Bibliography . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57

List of Figures

Figure 4.1 Comparing FFNN and CNN Models for trained for classification of PTSD and Control Subjects. . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
Figure 4.2 Example of a Convolutional neural network being used in a functional connectivity data classification task for PTSD and Control subjects. . . . 17
Figure 4.3 Comparing models test accuracies with different activation functions for the classification of PTSD/Controls subjects. . . . . . . . . . . . . . . . 19
Figure 4.4 Comparing the efficacy of optimizers used for CNN model. This data is based on our PTSD vs Control subjects classification. . . . . . . . . . . . 21
Figure 4.5 Loss curves indicating the epoch where the training process is halted. This example is based on a CNN models classifying our PTSD and Control subjects. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
Figure 4.6 Dropout Regularization introduced for the ABIDE model to reduce overfitting. Different dropout probabilities were tried in ABIDE for improvement in performance.1% dropout gave a better training and testing accuracy. 23

Figure 5.1 Deconvolution process chart here shows how in the deconvolution method information from layer2 is reconstructed with the use of location based switching information in the layers. To perform the opposite operation of each layer, the parameter matrices used in the transformation are transposed. 28
Figure 5.2 Left: A neural network-shaped classifier is shown on the left during forward pass. Right: the backward flow of relevances is shown. $w_{ij}$ are the connecting weights and $a_i$ is the activation of neuron $i$. $R_i^l$ is the relevance of neuron $i$ that is to be computed. . . . . . . . . . . . . . . . . . . . . 29
Figure 5.3 Gradient Heatmaps of 4 individuals PTSD subjects, classified as PTSD by the model. These heatmaps were generated by the 96% accurate PTSD classification CNN model (trained by 10-fold cross validation). The thickness of a path represents the contribution made that path to support PTSD decision. In this figure we are only showing the top 10 supporting paths of each heatmap. There numbers in the heatmaps are not subject to any upper or lower limits. . . . . . . . . . . . . . . . . . . . . . . . . . . 32
Figure 5.4 This brain visualization is showing which paths were most common among gradient heatmaps of 4 PTSD subjects. Each connection in this visualization was found in 3 of the subjects. The heatmap was generated by the 96% accurate PTSD classification CNN model (trained by 10-fold cross validation). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
Figure 5.5 Explanations generated for a PTSD subject, classified as PTSD by the 96% accurate CNN model, using 4 different interpretability techniques. The thickness of a path represents the contribution made that path to support PTSD decision. In this figure we are only showing the top 10 supporting paths of each heatmap. There numbers in the heatmaps are not subject to any upper or lower limits. . . . . . . . . . . . . . . . . . . . . 34

Figure 5.6 Left: This is a figure is showing which paths were common among all heatmaps generated for one PTSD subject. : Right: This figure shows which were not common among any of the four heatmaps of the PTSD subject. These algorithms include gradient, integrated gradients, LRP-z and LRP-epsilon method. The heatmaps were generated by the 96% accurate PTSD classification CNN model (trained by 10-fold cross validation). 35

Figure 6.1 Matrices contain the raw heatmaps generated using interpretability methods, and each matrix represents one model class (PTSD / Control). $H_{path1}$ contains information for path1 from all subjects. A large number of permutations of $H_{path1}$ are generated, and for each permutation the difference of means ($x$) is calculated. A p-value is calculated by seeing where $x_{actual}$ lies in the distribution of test statistic ($x$). The process is repeated to calculate a p-value for each brain path. An FDR corrected p-value threshold is applied to pass only significant brain paths to visualize and analyze. . . 38

Figure 6.2 Gradient Heatmap of significant paths that make PTSD/Control subjects different one each other. The test statistic in the permutation test was the difference of means. The color bar shows p-values of 0.05 as 0.0 mapped to 0.95 and 1 respectively. The darkest blue path identified least significant path and red most shows the most significant brain path. The raw PTSD and Control groups of heatmaps were generated for a CNN classifier (with 96% cross-validation accuracy). . . . . . . . . . . . . . . 40

Figure 6.3 LRP-z of significant paths that make PTSD/Control subjects different one each other. The test statistic in the permutation test was the difference of means. The color bar shows p-values of 0.05 as 0.0 mapped to 0.95 and 1 respectively. The darkest blue path identified least significant path and red most shows the most significant brain path. The raw PTSD and Control groups of heatmaps were generated for a CNN classifier (with 96% cross-validation accuracy). . . . . . . . . . . . . . . . . . . . . . . . . . . 41

Figure 6.4 Raw heatmap information is filled into actual correlation matrices to form groups A and B. A large number of permutations of the matrices are created and a t-test is performed between groups for each permutation. Applying the t-test on each permutation results in a binary mask with ones representing significant content of heatmap and zeros representing insignificant paths. A vector of cluster mass statistics is calculated for the actual data mask and all other permutations. A CMS value for an ROI is measured by counting its connections with other ROIs. From each CMS vector the maximum value is enlisted, and a distribution of max CMS is created. FWER with corrected p-value threshold is applied to calculate a cutoff CMS value. This CMS value threshold is applied on CMS vector of actual data. The thresholded CMS vector is transformed back into a binary correlation mask for visualization and analysis. . . . . . . . . . . 42

Figure 6.5 Visualization of significant paths generated by gradient heatmap algorithm, that is calculated using the cluster mass thresholding in the permutation test. These significant paths make PTSD/Control subjects different one each other. The test statistics in this permutation test were the difference of means and cluster mass statistic. The raw PTSD and Control groups of heatmaps were generated for a CNN classifier (with 96% cross-validation accuracy) for doing this analysis. . . . . . . . . . . . . . 43

Figure 6.6 PTSD: Consensus heatmap with score 8 or greater, showing the significant paths calculated using the permutation testing with difference of mean statistic. The dark blue color in figure represents paths were common among none of heatmaps and by 1 (red) represent paths that were common among all the heatmap techniques. . . . . . . . . . . . . . . . . . 44

Figure 6.7 PTSD: Consensus heatmap with score 8 or greater, showing the significant paths calculated using the permutation testing with CMS statistic-based thresholding. The dark blue color in figure represents paths were common among none of heatmaps and by 1 (red) represent paths that were common among all the heatmap techniques. . . . . . . . . . . . . . 45

Figure 7.1 Contribution of each site to the ABIDE-I multiple sites dataset. The number in the pie chart indicates the number of subjects in a site. . . . . . . . 50

Figure 7.2 A demonstration of the accuracy values for each ABIDE-I site,Äôs model, in comparison with the number of subjects in the site. . . . . . . . . . . . 52

Figure 7.3 A consensus heatmap generated by calculating a consensus between 10 interpretability methods for data acquired at UM site from ABIDE-I dataset. A score of 1 means a path was identified my all interpretabilty algorithms. This is consensus is thresholded at 0.9, meaning agreement between 9 or more of the 10 interpretability methods. . . . . . . . . . . . . . . . . . . 53

Figure 7.4 A consensus heatmap generated by calculating a consensus between 10 interpretability methods for data acquired at NYU site from ABIDE-I dataset. A score of 1 means a path was identified my all interpretabilty algorithms. This is consensus is thresholded at 0.9, meaning agreement between 9 or more of the 10 interpretability methods. . . . . . . . . . . . 54

# List of Tables

Table 2.1 Brain Disorder Classification Studies Using Machine Learning . . . . . . 9
Table 2.2 List of neuroimaging studies that used Interpretability techniques . . . . . 10

Table 3.1 Shows a summary of the datasets which are later used to train classifiers. Reduced features refer to the number of features we receive after we apply t-test on the content of the connectivity matrices. . . . . . . . . . . . . . 12

Table 4.1 A summary of the CNN model used for PTSD/Controls classification task. 22
Table 4.2 A summary of the CNN model used for Alzheimer's disease/Controls classification task. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
Table 4.3 A summary of the CNN model used for ASD/Controls classification task. 24
Table 4.4 Performance of brain disorder neural network classifiers saved at end of model training. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

Table 6.1 A summary of the regions of interest associated with discriminating connections identified by permutation testing based consensus explanation. The classification model under consideration is PTSD/Controls Convolutional Neural Network (with 96% cross validation accuracy) . . . . . . . 46
Table 6.2 A summary of the discriminating regions of interest identified by permutation testing (with Cluster mass thresholding) based consensus explanation. The classification model under consideration is PTSD/Controls Convolutional Neural Network (with 96% cross validation accuracy) . . . 48

Table 7.1 Comparing the classification and interpretability performance of the ABIDE-I CNN models (with 80% training set and 20% testing set).CV denotes cross-validation and Acc denotes model accuracy. . . . . . . . . . . . . . 51

# List of Abbreviations

| | |
|---|---|
| ABIDE | Autism Brain Imaging Data Exchange |
| ACG | Anterior Cingulate Gyrus |
| AD | Alzheimer’s Disease |
| ADNI | Alzheimer’s Disease Neuroimaging Initiative |
| Amyg | Amygdala |
| ASD | Autism Spectrum Disorder |
| CAL | Calcarine Fissure |
| CAU | Caudate Nucleus |
| CMS | Cluster Mass Statistic |
| CNN | Convolutional Neural Network |
| CUN | Cuneus |
| CV | Cross-validation |
| DCG | Median Cingulate and Paracingulate Gyri |
| DMN | Default Mode Network |
| DTI | Diffusion Tensor Imaging |
| EEG | Electroencephalography |

| | |
|---|---|
| FC | Functional Connectivity |
| FDR | False Discovery Rate |
| FFG | Fusiform Gyrus |
| FFNN | Feed-forward Neural Network |
| fMRI | Functional Magnetic Resonance Imaging |
| FWER | Family-wise Error Rate |
| GBP | Guided Backpropagation |
| IPG | Inferior Parietal Gyrus |
| ITG | Inferior Temporal Gyrus |
| Ling | Lingual Gyrus |
| LRP | Layer-wise Relevance Propagation |
| LSTM | Long Short-term Memory Network |
| MFG | Middle Frontal Gyrus |
| MOG | Middle Occipital Gyrus |
| MRI | Magnetic Resonance Imaging |
| MTG | Middle Temporal Gyrus |
| NYU | New York University |
| OLF | Olfactory Cortex |
| PAL | Pallidum |
| PCG | Posterior Cingulate Gyrus |
| PCUN | Precuneus |

| | |
|---|---|
| PET | Positron Emission Tomography |
| PoCG | Postcentral Gyrus |
| PreCG | Precentral Gyrus |
| PTSD | Post-traumatic Stress Disorder |
| REC.G | Gyrus Rectus |
| RNN | Recurrent Neural Networks |
| ROI | Region of Interest |
| rsfMRI | Resting State Functional Magnetic Resonance Imaging |
| SFG | Superior Frontal Gyrus |
| SMG | Supramarginal Gyrus |
| sMRI | Structural Magnetic Resonance Imaging |
| SOG | Superior Occipital Gyrus |
| SVM | Support Vector Machine |
| Tha | Thalamus |
| UM | University of Michigan |

# Chapter 1

# Introduction

## 1.1 Motivation

Human brain is known to be the most complex entity in the nature, and understanding parts of the human brain with certainty is challenge in neuroscience. The brain disorders are difficult to diagnose because it is difficult to identify underlying complex mechanism with high fidelity. The diagnosis in most of the clinical setups are done by the knowledge of the medical professionals using MRI images (among other diagnostic tools). And also, every individual is different; this adds more to the problem of understanding human brain and its disorders. So, we need mathematical processing of brain data, that is able to draw conclusions by looking at data that is available.

Deep neural networks are becoming increasingly common for the classification of neuroimaging data, to diagnose and learn about human brain disorders. With the help of enough training data, these models give promising classification performance to differentiate between healthy and patients. The deep learning classification algorithm gives a flag at the output, saying a 'yes' or 'no' about a patient. So, this level of detail is not enough, and it challenges the deployment of such systems in a clinical or research setup where human lives are involved. We have a mechanistic understanding of the deep neural network models but there is no clear understanding of how/why they make their decisions after they are trained to perform a task. Neuroimaging needs more elaboration about the classification decisions because human brain data contains immense number of features and the problems could be anywhere. Without a strong explanation, a decision is only a data driven opinion, which could be biased or have any other problem inherited from dataset. For instance: if all the data for brain disorder study is collected from a specific age group, gender, geography or scanning equipment; the decision-making will generalize less regarding the disorder and specialize more regarding the dominant metadata (age, gender, geography, scanner specifications). One such example can be seen in [1], showing how the classifier focused on copyright text in an image and less on the actual content. Data imbalance between the classes in the population, also is a challenge in medical imaging and can produce misleading performance metrics (the accuracy for a data based decision could be 95%, this could be occurring because one class comprises the larger part the sample size and not due to the model's learning of data).

The interpretable deep learning algorithms give an explanation of neural network decisions in the input space in the form of a heatmap. The heatscores in a heatmap guide the users of a data driven system, about input features that play an important role. There are several such algorithms that give explanations of the classifier diagnoses decisions. Every explanation algorithm has a separate working principle; many of them fail under certain conditions of hyperparameters and even basic perturbations in the input [2]. None of such methods is known to give a perfect solution. The combined power of all these algorithms is used here to generate disorder specific heatmaps in the human brain space. Our results inform us about the significant connectivity features that make healthy and diseased subjects different. Collection of large datasets at one imaging site is costly and challenging to manage by one organization or team, so it is very important to combine and use data from different sites. Mathematically it is more understandable to work with more data, however, there are several unseen biases that reduce the quality of such combined data. We show an example of such datasets and identify why a neural network performs ineffectively on it. An interpretability analysis of multisite data models could give more general understanding of disorders independent of site or scanner specific effects.

## 1.2 Introduction

Magnetic resonance imaging (MRI) is a non-invasive technique of studying structure and functioning of the brain by measuring the changes in the properties of the molecules in the subjects' brain [3]. The imaging information collected from MRI machines are used for diagnosis of mental disorders and research for understanding of human brain. From the magnetic resonance imaging scanner there are two main types of images generated, the structural MRI (sMRI) images and functional MRI (fMRI) images. The structural images contain good spatial resolution and give anatomical understanding of the brain, while the functional MRI gives better temporal resolution and basically the data is obtained over the course of brain activity. In fMRI the goal is to observe the human brain activations while it is functioning in real-time as closely as possible. Usually in fMRI a subject is supposed to perform certain tasks, the data is collected while the user performs the tasks, for example: eye blinking, finger taping and watching a movie etc.

A more effective and convenient form of fMRI is the resting state fMRI (rsfMRI). In rsfMRI the patient/subject is put in the MRI machine and is supposed to be not involved in any physical activity or specific thinking or goal-directed behavior [4]. Resting state functional connectivity matrices from rsfMRI data are calculated by the calculating the correlation of the time series of a region of interest (ROI) with every other ROI over time. A connectivity matrix contains the correlations among all regions of interest, of a brain. The elements of connectivity matrices show how activation of ROIs affect each other in default or resting state.

In this study, we use rsfMRI connectivity data for creating an interpretable tool for the diagnosis of brain disorders. The disorders considered in the experiments here are post-traumatic stress disorder (PTSD), Alzheimer's disease and Autism. The connectivity matrices are reduced in size with the help of a feature selection algorithm, this helps getting rid of the features that are less likely to be discriminatory. The reduced connectivity matrices are used to train a deep neural network model for the classification of people between healthy and those individuals with disorders.

The trained deep neural networks are used in further analysis using interpretable deep learning techniques. These methods read a model, an input and the models' decision to generate an explanation for an individual subject or image. In order to understand a disorder, we use explanations generated for our entire training data (large groups of patient and control heatmaps) in a permutation testing mechanism. The permutation test helps generate one single heatmap or explanation that reflects the significance of discriminating connectivity paths between groups; we correct for multiple comparison in the testing process. We used permutation testing with two different strategies, difference of means statistic and a cluster mass statistic thresholding method.

We take our study a step further, by seeing whether all the explanation algorithms give similar resulting paths of significance for the disorders. To do this, we build a composite pipeline that combines the power of multiple interpretability algorithms to give explanations of diagnosis decisions. The explanations generated by this interpretability pipelines show visually the magnitude of consensus for a significant path. This way, we are able to narrow down to brain paths that are more important or less important. Out of the three datasets in question, we get convincing explainable results for Post-traumatic stress disorder and Alzheimer's disease. For Autism, our data comes from multiple sites, so initially we did not achieve any considerable performance of the classification models. It means, the models were not able to capture trends in this dataset, so we used site harmonization to show that the problem is occurring because data came from multiple sources. We processed each sites' data separately to show this problem.

## 1.3 Thesis Organization

The chapter 2 gives a basic view of the relevant literature and content The chapter 2 gives a basic view of the relevant literature and content specific detail is available in the following chapters. Chapter 3 gives a description of the datasets, preprocessing and how the features extraction was performed. Chapter 4 focuses on convolutional neural networks and training classifers for the disorders. Chapter 5 explains the variants of gradient and relevance-based algorithms for generating heatmaps. Chapter 6 shows to how to calculate general heatmaps of a disorder and make an integrated pipeline for the explanations, using all the interpretability algorithms together. This concept is illustrated using PTSD data. Chapter 7 considers ABIDE multisite dataset and demonstrates that robustness and agreement of explanations generated by interpretability algorithms with previous literature is dependent upon the fidelity of the classification model in terms of the prediction accuracy that it can provide. Chapter 8 gives a conclusions and future work directions regarding the thesis.

# Chapter 2

# Literature Review

## 2.1 Resting State Functional Connectivity

In the beginning of fMRI, researchers would usually explore the brain functioning in the presence of a stimulus (such as sensory, motor or some type of cognitive task) to see how brain activations occur. Resting state fMRI was first defined by Biswal in 1995 [5] when he tried having his subjects do literally nothing in the MRI scanner. Instead of random spontaneous neuronal activity at rest, he saw structure, organization and correlations among parts of the brain. The rsfMRI concept became more famous when Marcus Raichle in 2001 [6] identified a default mode network in the brain at rest. This default network would show activation at rest and decreased activity when subjects were engaged in a cognitive task. Further exploring the default mode network (DMN), Michael Greigius showed that at rest, it shows correlated oscillations [7]. A large number of researchers have suggested that networks of brain regions that activate and deactivate together during tasks maintain signatures of their connectivity, that can be studied even at rest [4]. A number of other networks have also been observed at rest including sensorimotor, vision, hearing and memory [8]. The slow, synchronized oscillations within each network (that are independent of one another) are robust, and steady during sleep and under anesthesia [9, 10]. In rsfMRI the data can be obtained with a dedicated scan, in which individuals are instructed to simply rest, or by inferring resting-state data from segments of rest embedded within a series of tasks [11]. The lack of task requirements in rsfMRI makes it more attractive, especially for patients who have difficulty with understanding or executing instructions. The resting-state functional connectivity (FC), usually is measured by calculating a correlation between the time-series of the brain regions of interest under analysis. Resting-state FC can show the intrinsic organization of human brain, and the spontaneous activations at rest are also predictive of task and behavior performance [12].

## 2.2 Brain Disorder Diagnosis and Functional Connectivity

Resting-state functional connectivity also helps diagnose the disorders in human brain. In resting-state functional connectivity we assume that the default networks of the brain do not function normally. For example: the disruption of default mode network has been seen in Alzheimer's disease, autism, schizophrenia and depression [13]. For autism, several studies have shown abnormal patterns of connectivity in various networks related to memory, language, emotion processing and social cognition [14]. Similarly, there are studies in which resting state functional connectivity information has been used to look in Post-traumatic stress disorder (PTSD). During the period from 1990 to 2017, more than 200 papers used functional connectivity features alone or multi-modality features including functional connectivity to classify or predict brain disorders [15]. Given below are three brain disorders which we will explore in more detail using our experiments in later chapters.

### 2.2.1 Autism Spectrum Disorder

It is a complex neurodevelopmental disorder, consisting of a long range of symptoms. With different levels of disability, it affects a person's interactions, communications and learning [16]. This disorder begins in childhood and prevails through the lifetime; in United States, it is estimated that ASD has a prevalence of 1:68 [17]. The treatment costs of an ASD patient in lifetime exceed one million dollars [18]. There is still no clear understanding of autism, it might be due to genetics, brain structure and functioning, developmental and environmental factors.

### 2.2.2 Alzheimer's disease and Mild Cognitive Impairment

Mild cognitive Impairment (MCI) is a syndrome which causes memory loss greater than expected by aging [19]. About 3-19 percent of people, older than 65 years suffer from MCI. The symptoms of MCI are not as dangerous as Alzheimer's disease (AD) and people with MCI can carry out their normal activities [20]. However, in AD the cognitive functioning is also disturbed in addition to memory loss [16]. AD is the most common type of dementia, causes problems with memory, thinking and behavior [21]. AD occurs increasingly in the people older than 60 years, its significance as a public health problem was became evident [22]. Between 2000 and 2013, the death results from AD increased remarkably 71 percent, making AD the sixth leading cause of death in the United States [23].

### 2.2.3 Post-traumatic stress disorder

Post-traumatic stress disorder (PTSD) is a brain disorder common in people who go through some traumatic and tense situations. It is one of the most prevalent disorders in war veterans [24]. An investigation by US military for mental health problems identified 9.8 percent of veterans returning from Iraq, 4.7 percent from Afghanistan, and 2.1 percent from other locations were at risk of PTSD [25]. The biological basis for PTSD from several studies is reviewed by [26] and neuroimaging studies, for example: [24, 27].

## 2.3 Deep Learning for Brain Disorder Classification

In the last decade, we have seen the increasing use of traditional machine learning and deep neural networks in neuroimaging. These complex algorithms have performed very well on classification tasks. Most common of these algorithms are support vector machines (SVM), random forests, feed-forward neural network (FFNN), convolutional neural networks (CNN), recurrent neural networks (RNN) and auto-encoders. Working of these algorithms is explained in the next chapter. Deep learning has been used in neuroimaging with data of all modalities. But here, we will focus more on the studies that used functional connectivity or particularly resting state data to classify brain disorders. Table 2.1 below summarizes some example studies from the recent years. In this research, we do not necessarily claim to have created the best classification models. Our focus is mainly about how to make these classifiers more transparent, trustworthy and interpretable for use in clinical setup.

## 2.4 Interpretability of Deep Learning

Interpretability refers to the degree of clear understanding of a system. Tang et al [28] say: ‚ÄúInterpretability is the degree to which a human can understand the cause of a decision‚Äù. Greater interpretability in a machine learning or deep learning system means that it is easier for humans to comprehend why it takes certain decisions. With the increasing number of applications and the improving performance of deep learning in neuroimaging, the question of interpretability becomes important. Because the ultimate goal of neuroimaging is to understand the human brain, diagnose brain disorders and help in curing the disorders. The research and decision-making systems are extremely challenging to deploy in clinical setup for actual use. Neuroimaging deep learning systems must be made interpretable because of the huge importance of human lives. As researchers we have a mechanistic understanding of how parts of a deep neural network work. For example: the relationship of nonlinearity with the depth of the layers, types of activation functions, convergence of the learning process etc are all understood very well in theory. However, it is a challenge to generate a clear post-decision explanation of the neural network output. For example: a good explanation would exactly identify the reasons why an MRI scan of a given patient is attributed with dementia. There are a few studies that have explored the question of interpretability of neuroimaging classification models, as shown in Table 2.2. These research studies are improving over time, but as of now, there is no study that shows the explanation of deep models convincingly. This research is the one of the first in generating explanations for neural network classification decisions for resting-state functional connectivity data. The interpretable machine learning algorithms themselves will be described later in the text.

Table 2.1: Brain Disorder Classification Studies Using Machine Learning

| Authors | Modality | Model | Dataset | Acc | Sen | SPE |
|---|---|---|---|---|---|---|
| Shi et al [29] | MRI, PET | Stacked Deep Polynomial Network and Linear SVM | ADNI, 202 Subjects | 97.13 | 95.93 | 98.53 |
| Suk et al [30] | MRI, PET | Deep Boltzmann Machine with an SVM | ADNI, 398 Subjects | 95.35 | 94.65 | 95.22 |
| Korolev et al [31] | MRI | 3D Convolutional Neural Network based on ResNet and VGGNet | ADNI, 231 Subjects | 88 | - | - |
| Thomas et al [32] | fMRI | Long Short Term Memory Networks | 130 Subjects | - | - | - |
| Choi et al [33] | FDG-PET, AV45-PET | Multimodal 3D Convolutional Neural Network | ADNI, 492 Subjects | 96 | 93.5 | 97.8 |
| Shao et al [34] | fMRI(FC with ALFF) | Deep Forest: a combination of multiple classifiers | ADHD-200 | 82.73 | 77.67 | 87.23 |
| Riaz et al [35] | fMRI(FC) | SVM | ADHD-200 | 86.7 | 77.27 | 90.16 |
| Mao et al [36] | rs-fMRI | 4D CNN with LSTM Convolution | ADHD-200 | 71.3 | 73.2 | 68.7 |
| Wang et al [37] | sMRI | 3D CNN | ADHD-200 | 69.01 | - | - |
| Deshpande et al [38] | fMRI | Fully Connected Cascade Neural Network | ADHD-200 and Multisite Data of 1177 Subjects | 95 | - | - |
| Kong et al [39] | fMRI | Deep Neural Network | 364 Subjects | 90.39 | 84.37 | 97.38 |
| Eslami et al [40] | rs-fMRI | Autoencoder and Single Layer Perceptron | ABIDE-1 with 1112 Subjects | 70.3 | 68.3 | 72.2 |
| Heinsfeld et al [41] | rs-fMRI | Autoencoder and Deep Neural Network | ABIDE | 70 | 74 | 63 |

**Abbreviations**: MRI (Magnetic Resonance Imaging), PET (Positron Emission Tomography), SVM (Support Vector Machine), CNN (Convolutional Neural Network), LSTM (Long Short-term Memory Network), ADNI (Alzheimer's Disease Neuroimaging Initiative), ABIDE (Autism Brain Imaging Data Exchange), ADHD (Attention Deficit Hyperactivity Disorder), ResNet (Residual Network), VGGNet (Visual Geometry Group Network), fMRI (Functional Magnetic Resonance Imaging), rsfMRI (Resting state functional MRI), EEG (Electroencephalography), DTI (Diffusion Tensor Imaging), sMRI (Structural Magnetic Resonance Imaging).

Table 2.2: List of neuroimaging studies that used Interpretability techniques

| Author | Modality | Model | Interpretability Method |
|---|---|---|---|
| Strum et al [42] | EEG | Multilayer Pooling Network | Layer-wise Relevance Propagation |
| Thomas et al [32] | fMRI | CNN and LSTM | Layer-wise Relevance Propagation |
| Oh et al [43] | fMRI | Convolutional Autoencoder and CNN | Gradient Method |
| Rieke et al [44] | sMRI | CNN | Occlusion, Gradient and Guided Backropagation |
| Tang et al [28] | Whole Slide Images | CNN | GradCam |
| Kawahara et al [45] | DTI | CNN | Gradient |

**Abbreviations**: CNN (Convolutional Neural Network), LSTM (Long Short-term Memory Network), fMRI (Functional Magnetic Resonance Imaging), EEG (Electroencephalography), DTI (Diffusion Tensor Imaging), sMRI (Structural Magnetic Resonance Imaging).

# Chapter 3

# Dataset and Preprocessing

The resting state fMRI data used in this study originates from the opensource neuroimaging databases. These include ADNI (Alzheimer's Disease Neuroimaging Initiative), ABIDE (Autism Brain Imaging Data Exchange) and data for PTSD was acquired in house at AU MRI Research Center. Preprocessing and feature extraction of this data has already been formed by Pradyumna Lanka [46, 47] and is kept opensource. We are using his dataset with minor additions.

## 3.1 Dataset Description

The resting state fMRI data for ABIDE contains 566 healthy subjects, 339 with autism and 93 with Asperger's syndrome. In ADNI we have 132 subjects in total, with 34 EMCI (early MCI) subjects, 29 LMCI subjects, 29 with Alzheimer's disease and 35 healthy controls. For PTSD the data had been collected in house, this resting state fMRI data was acquired from 87 active duty male US Army soldiers who served in Iraq or/and Afghanistan, with 28 controls, 17 diagnosed with only PTSD and 42 diagnosed with both PTSD and post-concussion syndrome (PCS). Our in-house data collection was approved by Institutional Review Board at Auburn University and Headquarters U.S. Army Medical Research and Material Command (HQ USAMRMC IRB). A 3T Siemens MAGNETOM Verio Scanner was used to scan participants, with a 32-channel head coil. Resting state data was collected from the subjects in two runs using a T2* multiband echoplanar imaging (EPI) sequence. The data acquisition parameters were, TR = 600ms, TE = 30ms, FA = 55 degrees, multiband factor = 2, voxel size = 3x3x5(millimeter cube) and

1000 points in time. The brain coverage for the volumes was restricted to only cortical and subcortical areas, and cerebellum was not included. We use binary classification to classify between controls and patients in these 3 datasets, so we can focus particularly on the diseases in question and the explainability of models. Table 3 shows the information about how our data was organized and its properties of concern.

Table 3.1: Shows a summary of the datasets which are later used to train classifiers. Reduced features refer to the number of features we receive after we apply t-test on the content of the connectivity matrices.

| **Disorders** | **Dataset** | **Classes** | **Number of Subjects** | **Reduced Features** |
|---|---|---|---|---|
| Post-traumatic Stress Disorder | AU MRI Research Center | 2 | 174 | 677 |
| Alzheimer's Disease | ADNI | 2 | 132 | 665 |
| Autism | ABIDE | 2 | 988 | 1357 |

## 3.2 Preprocessing

For preprocessing the resting state fMRI data, a Matlab package called DPARSF (Data Processing Assistant for Resting-State fMRI Toolbox) was used by [48]. The preprocessing steps performed on data include 3D volume realignment, slice-timing correction, co-registering of T1-weighted structural images to mean of the functional image, nuisance variable regression, mean global signal, white matter and signals of cerebrospinal fluid and 6 motion parameters. The dataset was normalized with to MNI (Montreal Neurological Institute) template. The BOLD (blood oxygen level dependent) timeseries of voxels was deconvolved by estimating the hemodynamic response function (HRF) using blind deconvolution process. These timeseries signals were filtered temporally using a bandpass filter of 0.01Hz-0.1Hz. Using CC200 template [49], mean timeseries of 200 functionally homogeneous brain regions was extracted. Pairwise functional connectivity values were calculated using the 200 timeseries signals, by measuring the Pearson‚Äôs correlation coefficients between signals. This process results in

19800 FC features. Whole brain coverage for PTSD and ADHD were not available, so time-series were extracted only using 125 and 190 regions. 19800 features for each example subject are a lot for a machine learning model to take care of, given the small sample sizes. So, the number of features was reduced to around 1000, using a t-test and an FDR correction of $p \leq 0.05$ was used to find these significant FC values.

# Chapter 4

# Deep Neural Network Classifiers

## 4.1 Selection of the Neural Network Classifier and Associated Parameters

We have seen previously in Table 2.1 that machine learning algorithms and deep neural networks are being used extensively in research, to perform classification and other types of decisions. The temporal nature of the data has already been taken care of, by calculating Pearson correlation coefficients. And all of our datasets have number of features between 600 and 1400, using a t-test. To build a deep learning classifier for this data, the obvious option is a fully connected network. In fully connected networks the feature vector is transformed through layers linearly and a non-linear operation is performed at the output of each layer [50]. The size of the layers is reduced step by step to the number of classes to generate, ,Äùyes,Äù or ,Äùno,Äù decisions about a class. This type of neural network has a greater number of parameters, because each neuron node in a layer is connected with all other nodes in the layers before and after itself. So, in this kind of network we will have to optimize millions of parameters. It is impractical because we are working with a smaller number of subjects or a smaller number of examples.

There are a number of factors that guide the type of classifier used, and we will discuss them briefly below. First, feature length in one subjects' data is between 600 and 1400, which is comparable to the feature length (724) of MNIST handwritten digits dataset [51]. MNIST digits classification using fully connected network hardly performs with an accuracy of 90 percent for 60000 sample size. Even, the 60000 images were not enough to generate an extraordinary result. However, the convolutional neural network accuracy for MNIST went to 97 percent without adding any complexity [52].

Second, our reduced feature vectors (features that pass the t-test for group differences in the training data) were resized as images and fed into a convolutional neural network. In the FC matrix or correlation space, we can think of each number as a pixel in an image. We can do this, because the convolution operation in a CNN is equivalent to a linear transformation (of fully connected networks) [53]. So, we reshape the feature vectors into squares and work with them similar to how images are dealt with in machine learning. If a feature vector does not evenly become a square, then we append constant numbers to fill the empty spots of the closest square. This process is equivalent to the padding operation in a convolution layer, to maintain the desired dimensions. Adding a constant feature to every single input instance, does not create a different because it is same for all classes and the gradient updates for the corresponding parameters are all zeros.

Third, we are dealing with our dataset examples as images in a convolutional neural network, however our sample sizes (PTSD 174, ADNI 132 and ABIDE 988) are not large enough to applications of convolutional neural network that have been successful in other fields. When we reshape our feature vectors into 2D images, the landmark/location for each feature stays the same in the entire dataset. In other words, the pose of content in our images does not change and it makes the dataset easier to learn, even with a smaller sample size. The negative effects of pose variation on image datasets and convolution-based classification was done in [54]. They saw that the change in the pose of an image challenges and fools a trained classifier. For our application, in training and even on deployment in a real situation, the pose would never change. Due to the above reasons, we think it is safe to use a convolutional neural network and it results in better classification accuracy using less data.

Figure 4.1 shows a performance comparison between a convolutional and feed-forward network on our PTSD dataset. The models were trained 10 times, from the start with a random initialization of weights and biases.

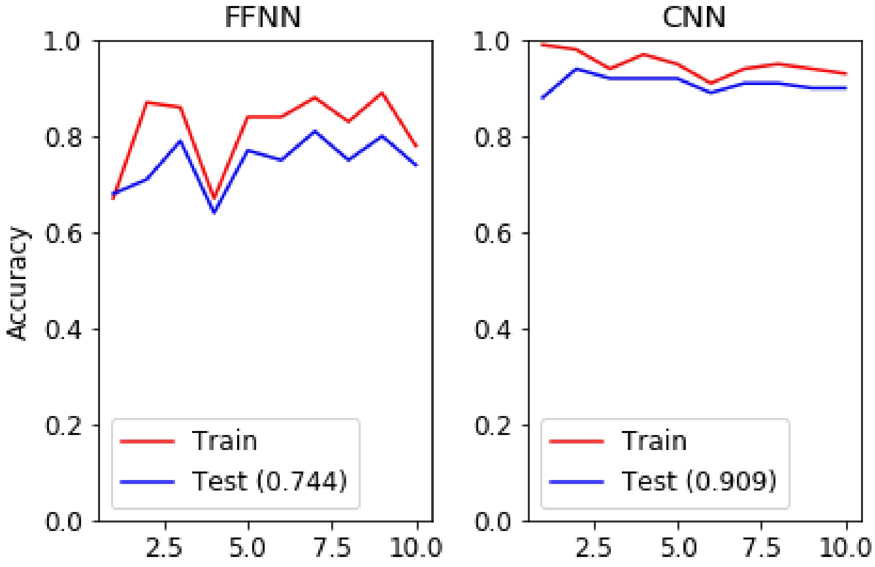
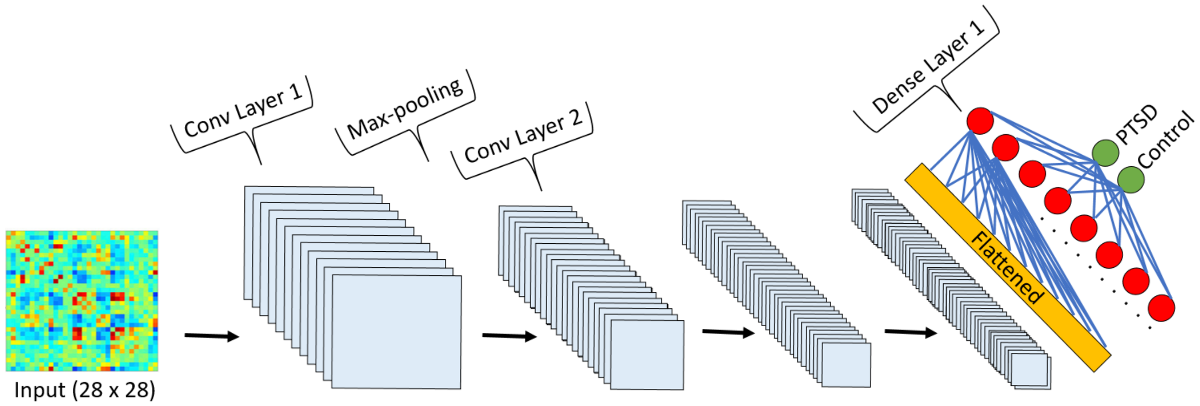


Figure 4.1: Comparing FFNN and CNN Models for trained for classification of PTSD and Control Subjects.

### 4.1.1 Convolutional Neural Networks

CNNs are a type of deep neural networks that have proven to be highly successful in medical image processing. CNN architecture exploits the spatial relationships of the content of the image; this feature is unique to this type of neural networks which makes them good at processing image information. They are basically related to both digital image processing theory and human visual cortex. CNNs have a hierarchical layered structure in which every layer process,Äôs, certain features of the image which are similar in human vision system [55]. The initial layers detect and process low level features (dots, lines) whereas the higher layers make sense of the complex features (for e.g. curves, corners, textures and their combinations). Most common ingredients of a CNN are convolutional layers, activation functions, pooling layers and fully connected layers. Figure 4.2 shows an image of our neural network architecture which is similar to Alex-net [56] and Lenet [52] . The convolutional layers apply a 2D convolutional operation on their input and kernel consists of the weights of the neural networks and adds a bias parameter to every output pixel. The weights and biases can be thought of as knobs, upon which the working of the machine depends. Every convolution layer output is passed through an activation function and then followed by a pooling layer. The activation function is simply a non-linear function. The purpose of the pooling layer is to reduce the size of the output images by subsampling; most of the famous architectures use average pooling or maximum pooling. So, the architecture consists of cascaded Conv-Pool layers and after passing the image to a few layers the size becomes very small.

Then the 2D image is flattened in the form of a vector and further reduced in size using fully connected layers. The fully connected layers also simply do a non-linear transformation are like layered graphs in which every input node has a connection to all output nodes. The connection is fully connected layers are weights/biases upon which the transformation depends. The output of the transformation is passed through a non-linear activation function to produce non-linearity. Finally, the output layer of a CNN classifier is made to have nodes equal to the number of classes.



]

Figure 4.2: Example of a Convolutional neural network being used in a functional connectivity data classification task for PTSD and Control subjects.

### 4.1.2 Training a Deep Neural Network Model

By training a deep neural network we refer to the process of adjusting its parameter, so that later when an input example of data is passed through it, it gives the desired output. This process can be broken down into four parts.

First, given a number of example images we pass them forward through our classifier models to generate decision. Now, some of these decisions are correct and some are incorrect. We have both the predictions of the model and actual labels of data.

Second, a cost function is required to evaluate the success of a model by measuring the magnitude of incorrect classifications. Usually an effective error is selected as the cost function, in our work we use sum of squared error. The smaller the cost function, the better the model performance.

Third, adjusting the parameters (weights and biases) of the model guided by the cost function value. Most commonly these adjustments are done using gradient descent algorithm. The partial derivatives of the cost function are calculated with respect of every parameter of the neural network. The neural network parameters are updated by subtracting the corresponding partial derivative values from them.

Finally, this process of doing predictions, calculating the new cost value, the partial derivatives and updating parameters is repeated over and over until the model starts performing significantly correct classifications.

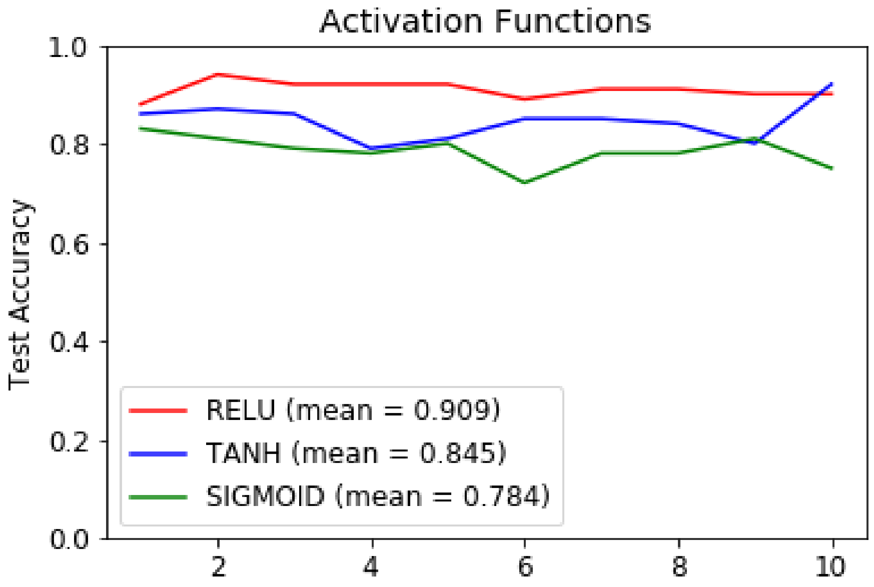
## 4.2 Structuring and Training Brain Disorder Classifiers

The selection of hyperparameters of the CNN comes through a mathematical understanding of different parts in a neural network and observing the learning process. We put our sample size, feature length and data quality in consideration as we tune the structure of CNNs. The hyperparameters include all the things that a developer sets, for better learning. Some of these parameters for CNNs are number of convolution layers, size of convolution kernel in each layer, number of kernels in convolution layers, selection of pooling mechanism, choice of activation function, gradient update strategy and learning rate etc.

There are many structure-related hyperparameters to start with, but since our samples,Äô shape matches well with the shape of MNIST examples, so we started off with LENET base architecture [52] and applied changes to it according to our data and application. LENET was the first convolutional network tried for MNIST and also one of the first few in the deep learning history.

We used max-pooling for the subsampling step in the classification networks. There 2 most common subsampling methods for CNNs (average pooling and max-pooling). Max-pooling Max-pooling [50] suits our application better because it passes only the most important brain paths over to the next layer (this process seems like we lose information; but it is better than a wrong assumption). The other reason we avoid another type of pooling is because in our 2D FC input, 2 adjacent numbers do not necessarily have to identify adjacent paths in brain space.

We replaced the tangent sigmoid activation function of convolution layers with rectified linear unit (Relu function). Relu was introduced in 2010 by Hinton [57] and has proven to be the best non-linear function to use in the hidden layers. For the last layer of the classifier we used softmax function, so the sum of the activation units is 1, there are sharp differences between unit activations. Figure 4.3 shows our comparison of the classification accuracies for Relu, tangent hyperbolic (Tanh) and Sigmoid function, after training the model 10 times with a random initialization of weights and biases. We found that Relu function produced the best mean test accuracy.



]

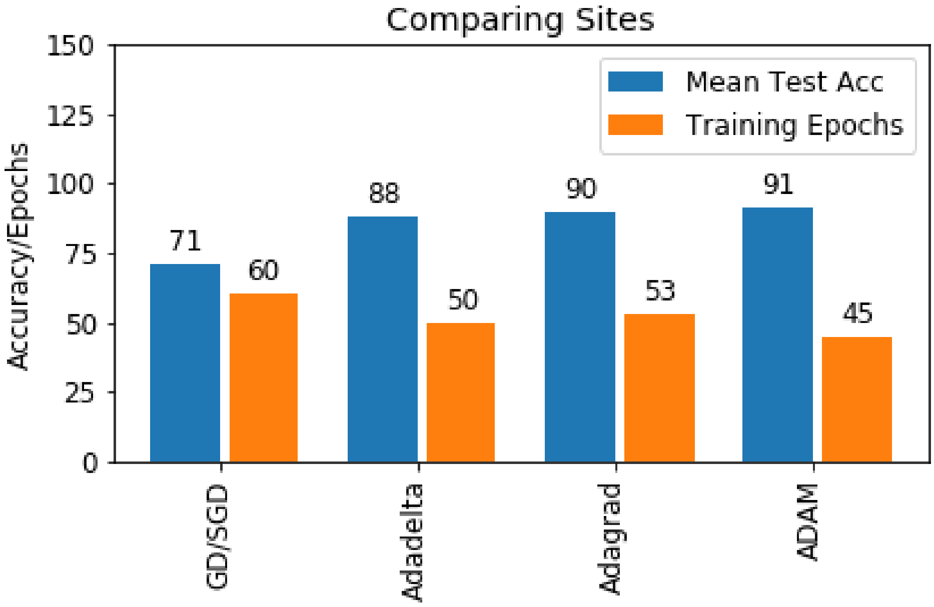
Figure 4.3: Comparing models test accuracies with different activation functions for the classification of PTSD/Controls subjects.

In general, for image classification, the kernel size is chosen as 3x3 or 5x5 or above (commonly most of the literature and applications use 3x3). A comparison of bigger and smaller kernel windows is discussed in [58]. We chose the smallest possible kernel (3x3) size, because in our 2D inputs, we do not assume any flat continuity in the input space and the dimensions are small. We try to put the focus on every edge in the image and it is possible only with smaller kernels. In all three models we worked with, the stride in the convolution process was set to 1, so that we do not skip any information.

In all the datasets we worked with, we had to use 3 convolutional layers instead of 2 layers of LENET. This is mainly because with 3x3 kernel size, the feature length does not become small enough in 2 layers. And when it is not small enough, the flat layers in the end add an unacceptable number of parameters. So, it is best to use another convolution layer instead of having long dense layers. We increased the number filters in the convolution layers from 16 to 32 and then 64, this was kept same for the networks. Normally, it is good to use as many filters as possible to capture all kinds of possible trends, but we could not afford to further increase the number of filters because we did not have enough data to optimize them.

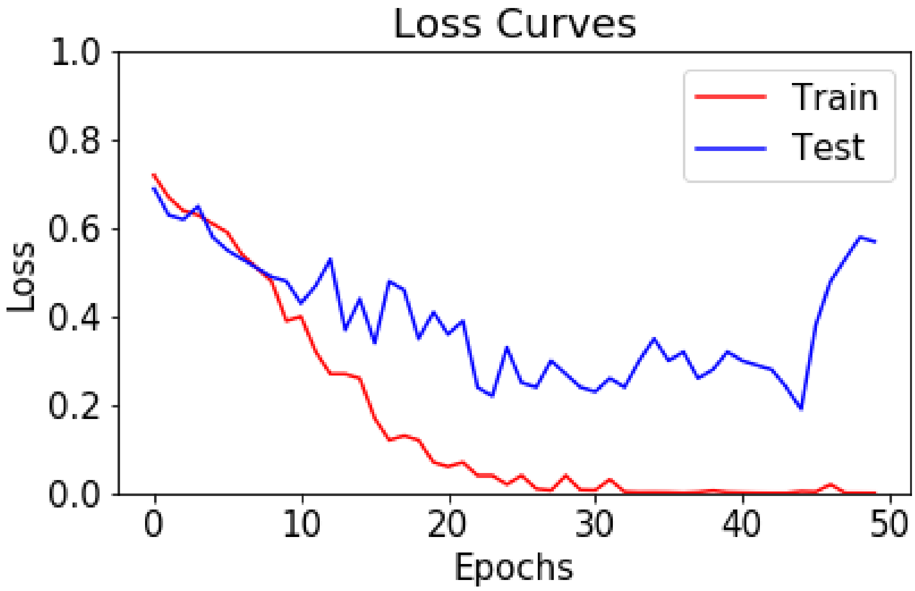
In the training process, that dataset was split into 80 percent training-set and 20 percent testing-set. Initially we used gradient descent algorithm [50] to optimize the weights and biases of the CNN. However, we found that the performance was not always the same for CNN training and stopped at a level. Sometimes the models learned enough and sometimes the learning just stopped at particular accuracy (indicting getting stuck in local minima). So, we employed ADAM optimizer [59] for a faster and consistent learning. With this ADAM as our parameter optimization algorithm, the problem of halts in learning was reduced and the networks learned according to their capacity in smaller number of epochs. Figure 4.4 shows the accuracy and training epochs comparison for the different optimizers available in Keras library.

An epoch indicates the number of passes of the complete training set in the learning process. When we mention the number of epochs for training, we are reporting the dataset iteration at which the gap between training and testing accuracies starts to increase consistently. Another way of deciding when to stop the training is, to look at training and testing loss values. Figure 4.5 shows the loss curves under training of PTSD CNN model. We can see that there is a point in the process where further training increases the testing loss (45 in case of PTSD model). Testing loss looks more edgy because it is based to totally unseen data.



]

Figure 4.4: Comparing the efficacy of optimizers used for CNN model. This data is based on our PTSD vs Control subjects classification.



]

Figure 4.5: Loss curves indicating the epoch where the training process is halted. This example is based on a CNN models classifying our PTSD and Control subjects.

### 4.2.1 Post-traumatic Stress Disorder and Alzheimer's Disease Models

The feature size for the PTSD and Alzheimers' subjects differs for only 12 paths (PTSD 677, ADNI 665). We used the same structure of CNN models described previously, for both of these disorders because the number of subjects as well is comparable. The only difference between the two is the number of epochs it took to train the models to achieve its highest accuracy. Tables 4.1 and 4.2 show the exact dimensions of the input, it passes through the network and corresponding number of parameters in each layer. It took 45 epochs to train the PTSD/Controls model to an accuracy of testing accuracy of 91 percent. And, for Alzheimers' disease it took 50 epochs to train the model to a testing accuracy of 77 percent.

Table 4.1: A summary of the CNN model used for PTSD/Controls classification task.

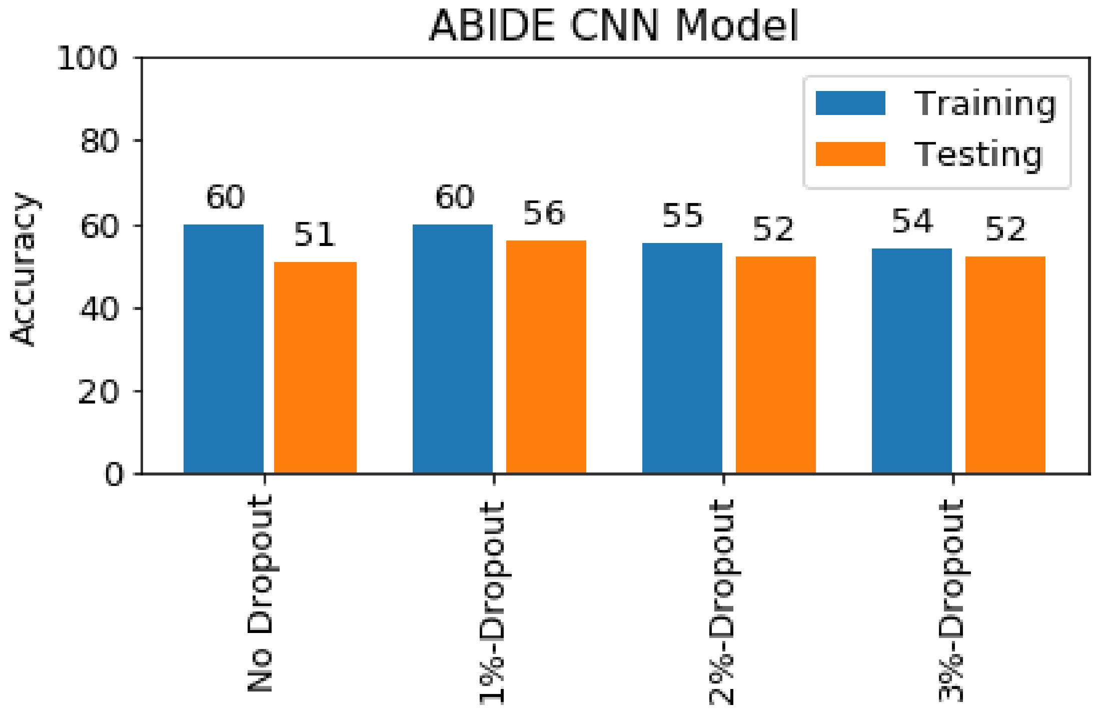
| **Layer Number** | **Layer** | **Output Shape** | **Number of Parameters** |
|---|---|---|---|
| 1 | 2D Convolution | 25 x 25 x 16 | 160 |
| 2 | 2D Convolution | 23 x 23 x 32 | 4640 |
| 3 | Max Pooling | 11 x 11 x 32 | 0 |
| 4 | 2D Convolution | 9 x 9 x 64 | 18496 |
| 5 | Flatten | 5184 | 0 |
| 6 | Dense | 8 | 41480 |
| 7 | Dense | 2 | 18 |
| Total Parameters = 69794 | | | |

Table 4.2: A summary of the CNN model used for Alzheimer's disease/Controls classification task.

| **Layer Number** | **Layer** | **Output Shape** | **Number of Parameters** |
|---|---|---|---|
| 1 | 2D Convolution | 24 x 24 x 16 | 160 |
| 2 | 2D Convolution | 23 x 23 x 32 | 4640 |
| 3 | Max Pooling | 11 x 11 x 32 | 0 |
| 4 | 2D Convolution | 9 x 9 x 64 | 18496 |
| 5 | Flatten | 5184 | 0 |
| 6 | Dense | 8 | 41480 |
| 7 | Dense | 2 | 18 |
| Total Parameters = 69794 | | | |

### 4.2.2 Autism vs Controls Model

For autism we had 1179 features and 988 subjects, the size of this dataset is good enough to train a classifier for a binary task. However, the data was collected at different sites (we work on site variation factor in the dataset, in a later chapter). Table 4.3 shows the structure and parameters of the model. With model structures tested for previous disorders, we repetitively found a large gap between the training and testing accuracy. We trained the model enough to get to a training accuracy of 60 percent, the testing was only 51 percent; this indicated no actual learning and more overfitting. An overfit model performs well on the training data, but on testing data it fails badly. So, we added 2 regularization layers to avoid this problem. Particularly, we used dropout regularization [60] mechanism and varied its setting. In this method we set a probability by which the neuron units in a layer are turned off, in the training stage. This makes the output of a layer less dependent on the layer before it, due to the randomness involved, resulting in a more robust model. The addition of dropout layers with 1 percent probability, reduced the gap between training and testing (60 / 56 percent at 67 epochs), the improvement was only marginal. Given below is a summary of the dimensions in the CNN and its trainable parameters. Figure 4.6 shows the effects of different dropout probabilities on model learning and overfitting control.



]

Figure 4.6: Dropout Regularization introduced for the ABIDE model to reduce overfitting. Different dropout probabilities were tried in ABIDE for improvement in performance.1% dropout gave a better training and testing accuracy.

Table 4.3: A summary of the CNN model used for ASD/Controls classification task.

| Layer Number | Layer | Output Shape | Number of Parameters |
|---|---|---|---|
| 1 | 2D Convolution | 35 x 35 x 16 | 160 |
| 2 | 2D Convolution | 33 x 33 x 32 | 4640 |
| 3 | Max Pooling | 16 x 16 x 32 | 0 |
| 4 | 2D Convolution | 14 x 14 x 64 | 18496 |
| 5 | Dropout | 14 x 14 x 64 | 0 |
| 6 | Flatten | 12544 | 0 |
| 7 | Dense | 8 | 100360 |
| 8 | Dropout | 8 | 0 |
| 9 | Dense | 2 | 18 |
| Total Parameters = 123,674 | | | |

## 4.3 Saving the models

After training, we saved these models and also saved a 10-fold cross-validated version of these models. Given below is a summary of the models we saved after training.

Table 4.4: Performance of brain disorder neural network classifiers saved at end of model training.

| Disorders | Epochs | Train/Test Acc | Cross-validation Acc |
|---|---|---|---|
| Post-traumatic Stress Disorder | 45 | 94% / 91% | 96% |
| Alzheimer's Disease | 50 | 85% / 76% | 85% |
| Autism | 67 | 60% / 56% | 58% |

# Chapter 5

# Generating Explanations of Classifier Decisions

Once we have classification models that are able to classify functional connectivity data generated from resting state fMRI of a subject; we become more interested in understanding how has the model under consideration, captured the trends in data (FC data in our case). This helps answer questions like, what are the most important features that are significant in a disease.

## 5.1 Interpertable Machine Learning

An interpretable machine learning algorithm is aimed at making the models transparent. It reads 3 blocks as input and to generate an explanation. These include an input example, a trained classification model, and the model's decision for the given example. The explanation has exactly the same dimensions as the input example. Let's say that we have a model *F* that reads an input *x* of *d* dimensions and generates a classification decision *y* of *c* dimensions in equation 5.1. An interpretability algorithm is denoted by *E* that reads *F*, *x* and *y* to generate a heatmap or explanation in 5.2. A brief description of the 10 methods of interpretability we used in this study is given below.

$$y = F(x) \tag{5.1}$$

$$H = E(F, x, y) \tag{5.2}$$

### 5.1.1 Gradient Heatmaps

The gradient method [61] is the simplest way of generating an explanation. In this method we simply calculate partial derivatives of the output with respect to every input feature. This is implemented by running a backward pass from output to the input layer (like it is done in normal training procedures), and one last derivative is calculated with respect to the actual input. In light of equations 5.1 and 5.2, it can be denoted as 5.3; where $E_i(F, x, y)$ is the heatmap value for feature at index $i$ of an input.

$$E(F, x, y)_i = \frac{\partial y}{\partial x_i} \tag{5.3}$$

### 5.1.2 Smoothgrad

Smoothgrad [62] is an improved version of Gradient method, it reduces noise and visual diffusion in explanations, by averaging heatmaps over noisy replicas of an input. For a given example *x*, the smoothgrad explanation $E_{sg}$ can be calculated as shown in equation 5.4. Where $g_i$ are the noise vectors that belong to $N(0, \sigma^2)$ and drawn from i.i.d from a normal distribution.

$$E_{sg}(F, x, y) = \frac{1}{N} \sum_{i=1}^{N} E(x + g_i) \tag{5.4}$$

### 5.1.3 Input ∘ Gradient

This method is referred to as input times gradient method. The explanation here is an element-wise product of input example and gradient matrix. This setup reduces the problem of gradient saturation and reduces diffusion [63]. Equation 5.5 shows how $E_{i.g}$ is calculated.

$$E_{i.g} = E(F, x, y) \circ x \tag{5.5}$$

### 5.1.4 Integrated Gradient

Integrated Gradient method [64] is another version of the gradient method that helps make the visualization more appealing and understandable. In this method we integrate explanation calculated by scaled versions of the input. This method is not very random compared smoothgrad.

$$E_{ig} = (x - \bar{x}) \int_{0}^{1} \frac{\partial F(\bar{x} + \alpha(x - \bar{x}))}{\partial x} d\alpha \tag{5.6}$$

### 5.1.5 Guided Backgpropagation (GBP)

Guided Backpropagation works on the same principle as gradient method, except for the additional constraint in the calculation of gradients. In GBP, when gradients are backpropagated through the neural network, the negative gradients are set to zero or passed through Relu functions in every layer.

### 5.1.6 Deconvnet Method

This technique [65, 55] is primarily good for models that work using convolutions. Deconvolution method simply gives an explanation of what a convolution layers has learned. If we talk of a complete CNN model, then Deconvnet method creates a new unsupervised network that is an opposite image of actual model. It performs the opposite operation of all the layers including pooling.

The filters in the new network are the inverse of filters used in classification of an input. The deconvolution operation helps reconstruct the input of a convolution layer using its output layer. The relu operation is performed as it is in the actual network. If we think of the new network as backward pass, the relu functions here clip the negative gradient and hence stop noise from propagating backwards.

The opposite approximate of the max pooling operation usually is performed by employing a location matrix of switches (1 is placed for highest activation in the pooling window and rest are set to zero). The using this location based switching information, we transfer information from output of a pooling layer to previous layer. It is important to understand the an explanation generated by the Deconvnet, is a forward propagation unlike gradient or relevance methods. Given below is an image obtained from [55], where deconvolution was tried first time.

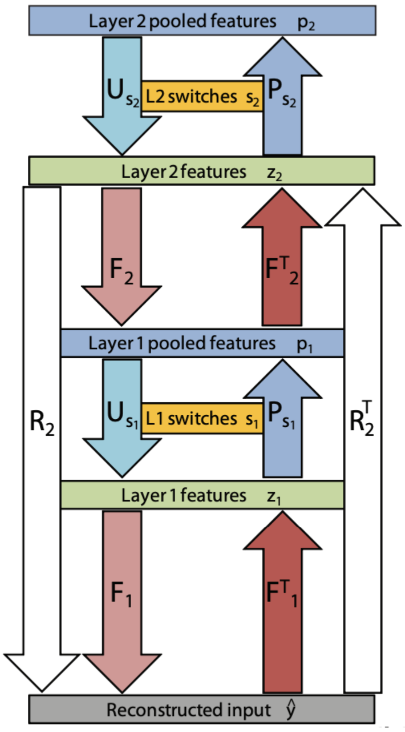


Figure 5.1: Deconvolution process chart here shows how in the deconvolution method information from layer2 is reconstructed with the use of location based switching information in the layers. To perform the opposite operation of each layer, the parameter matrices used in the transformation are transposed.

### 5.1.7 Layer-wise Relevance Propagation-z (LRP-z)

Layer-wise Relevance Propagation is a heatmap generating method for classifiers that can show the relevance of every input feature to the class decision [1]. The working principle of LRP is the breaking of sums into in their constituent parts. For example: in Figure 5.1 node 7 score is broken into three parts and fed back to the nodes 4, 5 and 6 according to their weights. This process continues from the output to the input over all the layers. In the last step we get relevance or contribution of every input feature in decision (either helping or opposing). It is a general framework under which a solution has to satisfy a major constraint called Layer-wise Relevance Conversation principle. This principle states that the sum of relevance scores across layers remains conserved i.e. there is no concept of sink or source. It could be represented mathematically as follows - and this equation summarizes the entire notion of LRP.

$$... = \sum_{d \in l+1} R_d^{l+1} = \sum_{d \in l} R_d^{l} = ...$$

where $R_d$ is the relevance/importance score for the $d^{th}$ dimension and $l$ represents the $l^{th}$ layer. Here we show a simple neural network-shaped architecture (Figure 2, from [1]) to derive a solution to the LRP. When a relevance is positive, it means the corresponding feature has helped in doing a particular decision; and if it is negative, it shows an opposition of the feature for the decision.

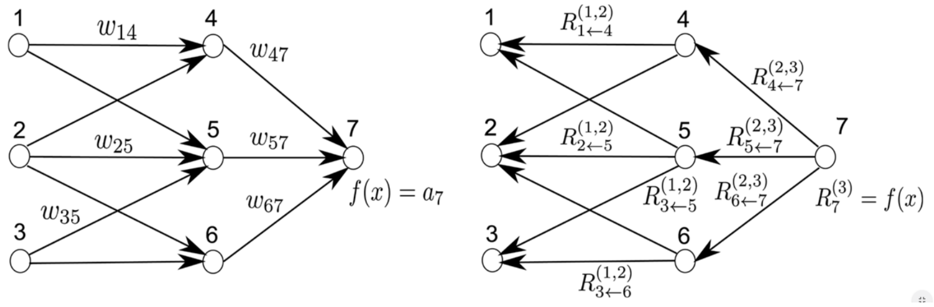

1
$w_{14}$
4
$w_{47}$
2
$w_{25}$
5
$w_{57}$
7
$w_{35}$
$w_{67}$
$f(x) = a_7$
3
6
1
$R^{(1,2)}_{1\leftarrow 4}$
4
$R^{(2,3)}_{4\leftarrow 7}$
2
$R^{(1,2)}_{2\leftarrow 5}$
5
$R^{(2,3)}_{5\leftarrow 7}$
7
$R^{(1,2)}_{3\leftarrow 5}$
$R^{(2,3)}_{6\leftarrow 7}$
$R^{(3)}_{7} = f(x)$
3
6
$R^{(1,2)}_{3\leftarrow 6}$

Figure 5.2: Left: A neural network-shaped classifier is shown on the left during forward pass. Right: the backward flow of relevances is shown. $w_{ij}$ are the connecting weights and $a_i$ is the activation of neuron $i$. $R_i^l$ is the relevance of neuron $i$ that is to be computed.

The forward pass through the network can be represented as follows -

$$a_4 = a_1 w_{14} + a_2 w_{24}$$

$$a_5 = a_1 w_{15} + a_2 w_{25} + a_3 w_{35}$$

$$a_6 = a_2 w_{26} + a_3 w_{36}$$

$$f(x) = a_7 = a_4 w_{47} + a_5 w_{57} + a_6 w_{67}$$

where $a_i$ is the activation for the $i^{th}$ neuron and $w_{ij}$ represents the connection weights from node $i$ to node $j$. Classifier score is represented by $f(x)$

Relevance computation part is similar to the backward propagation rule as can be seen from the following equations -

$$R_7^3 = f(x)$$

$$R_4^2 = R_{4\leftarrow 7}^{2,3} = R_7^3(\frac{a_4 w_{47}}{a_7}) = R_7^3(\frac{a_4 w_{47}}{a_4 w_{47} + a_5 w_{57} + a_6 w_{67}})$$
$$R_5^2 = R_{5\leftarrow 7}^{2,3} = R_7^3(\frac{a_5 w_{57}}{a_7}) = R_7^3(\frac{a_5 w_{57}}{a_4 w_{47} + a_5 w_{57} + a_6 w_{67}})$$
$$R_6^2 = R_{6\leftarrow 7}^{2,3} = R_7^3(\frac{a_6 w_{67}}{a_7}) = R_7^3(\frac{a_6 w_{67}}{a_4 w_{47} + a_5 w_{57} + a_6 w_{67}})$$

$$R_1^1 = R_{1\leftarrow 4}^{1,2} + R_{1\leftarrow 5}^{1,2} = R_4^2(\frac{a_1 w_{14}}{a_4}) + R_5^2(\frac{a_1 w_{15}}{a_5})$$
$$R_2^1 = R_{2\leftarrow 4}^{1,2} + R_{2\leftarrow 5}^{1,2} + R_{2\leftarrow 6}^{1,2} = R_4^2(\frac{a_2 w_{24}}{a_4}) + R_5^2(\frac{a_2 w_{25}}{a_5}) + R_6^2(\frac{a_2 w_{26}}{a_6})$$
$$R_3^1 = R_{3\leftarrow 5}^{1,2} + R_{3\leftarrow 6}^{1,2} = R_5^2(\frac{a_3 w_{35}}{a_5}) + R_6^2(\frac{a_3 w_{36}}{a_6})$$

The equation for LRP can be generalized as follows - where $z_{ij}$ is equal to $x_i w_{ij}$ and $z_i$ is equal to $\sum_{i'} x_{i'} w_{i'j}$

$$R_i^l = \sum_j \frac{x_i w_{ij}}{\sum_{i'} x_{i'} w_{i'j}} R_j^{l+1}$$

$$R_i^l = \sum_j \frac{z_{ij}}{z_j} R_j^{l+1} \tag{5.7}$$

### 5.1.8 Layer-wise Relevance Propagation-Epsilon

LRP-$\epsilon$ is an improved version of LRP, it consists of adding or subtracting a small positive number ($\epsilon$) from the denominator.

$$R_i^l = \begin{cases} \sum_j \frac{z_{ij}}{z_j - \epsilon} R_j^{l+1} & z \leq 0 \\ & (5.8) \\ \sum_j \frac{z_{ij}}{z_j + \epsilon} R_j^{l+1} & z > 0 \end{cases}$$

The purpose of the $\epsilon$ is to help limit of the relevance magnitude when $z_j$ is very close to zero (in other words if the sum of activations if later layer is very weak). The $\epsilon$ also can be used to have a control over the explanation; it will give sparse and less noisy explanations for bigger values.

### 5.1.9 Layer-wise Relevance Propagation ($\alpha$, $\beta$)

One of the techniques to redistribute relevances from higher to lower layers is called LRP-($\alpha$, $\beta$). In this we have a complete control over the flow intensity of positive intensities (using $\alpha$) and negative relevances (using $\beta$). Setting beta to zero allows to analyze only at features that support a decision in a positive way and vice versa. In this work, we employ two LRP-($\alpha, \beta$) explanations with parameters (2, 1) and (1, 0).

$$R_i^l = \sum_j (\alpha \frac{z_{ij}^+}{z_j^+} + \beta \frac{z_{ij}^-}{z_j^-}) R_j^{l+1} \tag{5.9}$$

## 5.2 Individual Heatmaps versus General Heatmaps

The interpretable machine learning algorithms described above, were able to generate an explanation for every single input (every subject). In figure 5.3 are a few gradient based explanations of subjects, classified as PTSD by our CNN model (trained by 10-fold cross-validation). Normally a raw heatmap identifies a very large number of paths with a high variance. We are showing here only the top 10 paths that supported the PTSD decision of the model. We saw that, even using the same model, and same way of generating the explanations for different PTSD subjects returns different explanations. Only 6 out of top-10 connections were found common among the 3 of the subjects, shown in figure 5.4.

In figure 5.5 we have showed heatmaps like before, generated using 4 different interpretability algorithms for 1 PTSD subject (using same model). These methods are gradient, integrated gradients, LRP-z and LRP-$\epsilon$. We are again seeing significant visual difference between the paths identified. This convinces us of the fact that there are similarities and differences between these methods as well. And there is literature [63, 1] that supports the fact that none of these algorithms work perfectly (there are interpretabilty algorithms that worked really well and fail badly in some situations). One simple example of failing interpretability methods is the presence of zeros in the input space. Input times gradient method, would definitely fail in such situation because we multiply the input space zeros to the gradients and heatmap skips

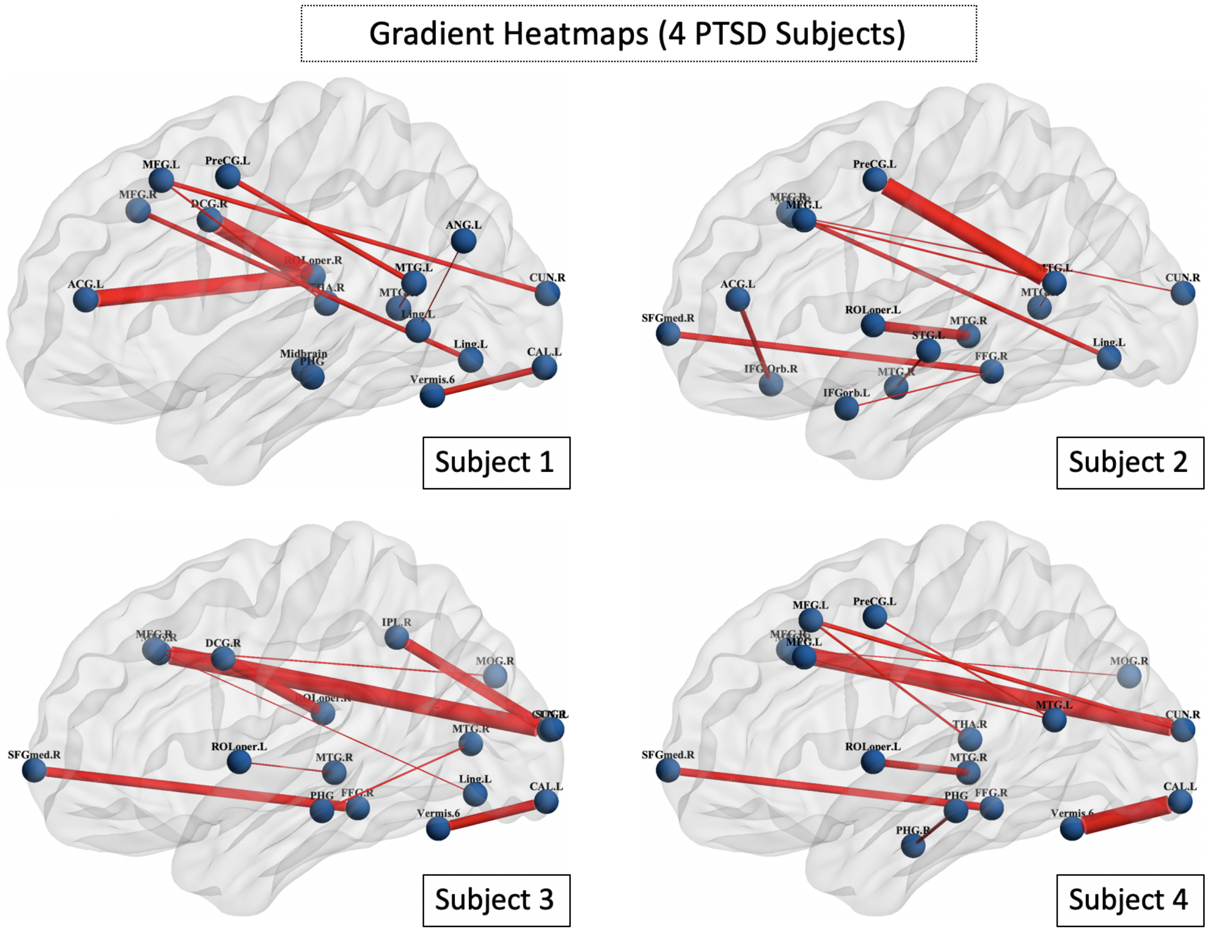


Figure 5.3: Gradient Heatmaps of 4 individuals PTSD subjects, classified as PTSD by the model. These heatmaps were generated by the 96% accurate PTSD classification CNN model (trained by 10-fold cross validation). The thickness of a path represents the contribution made that path to support PTSD decision. In this figure we are only showing the top 10 supporting paths of each heatmap. There numbers in the heatmaps are not subject to any upper or lower limits.

looking at some features which may be important. Figure 5.6 shows the brain paths were most common among the heatmap algorithms and paths that were most uncommon.

Given that every subject is different, it becomes more viable to generate explanations of the working of the model (generate group level heatmaps) in the first place. The general explanations also have to be in the input space and should identify that functional connectivity connections that are important for discerning between healthy subjects and patients. For this purpose, we pick a subset of data (where correct classifications occurred) and apply a permutation test to identify the generally significant paths for a trained classification model. The

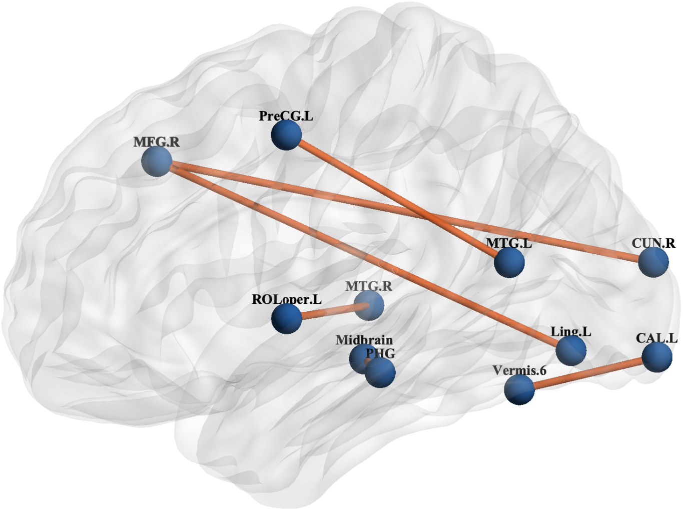


Figure 5.4: This brain visualization is showing which paths were most common among gradient heatmaps of 4 PTSD subjects. Each connection in this visualization was found in 3 of the subjects. The heatmap was generated by the 96% accurate PTSD classification CNN model (trained by 10-fold cross validation).

details to test are given in next chapter. And the question of individual subject level explanations becomes important when our model itself and the interpretability procedure is close to perfection.

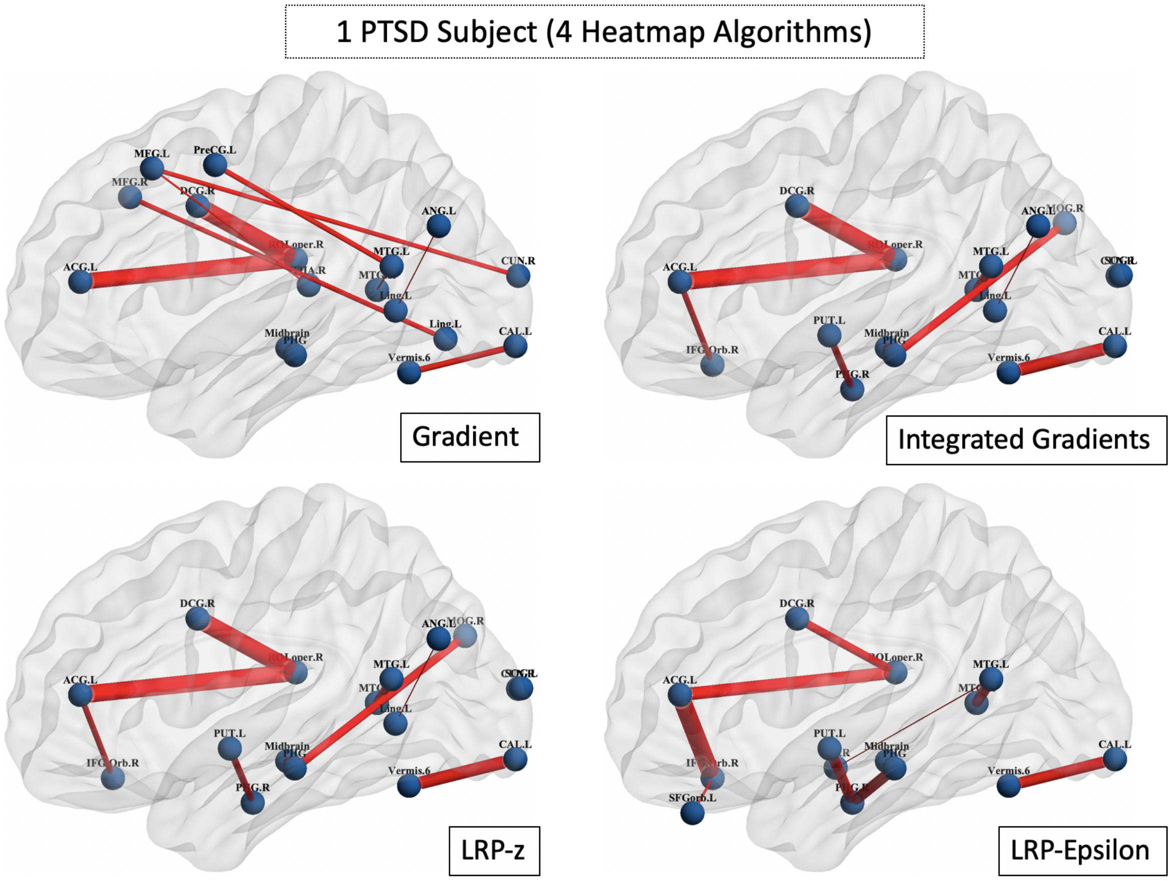


Figure 5.5: Explanations generated for a PTSD subject, classified as PTSD by the 96% accurate CNN model, using 4 different interpretability techniques. The thickness of a path represents the contribution made that path to support PTSD decision. In this figure we are only showing the top 10 supporting paths of each heatmap. There numbers in the heatmaps are not subject to any upper or lower limits.

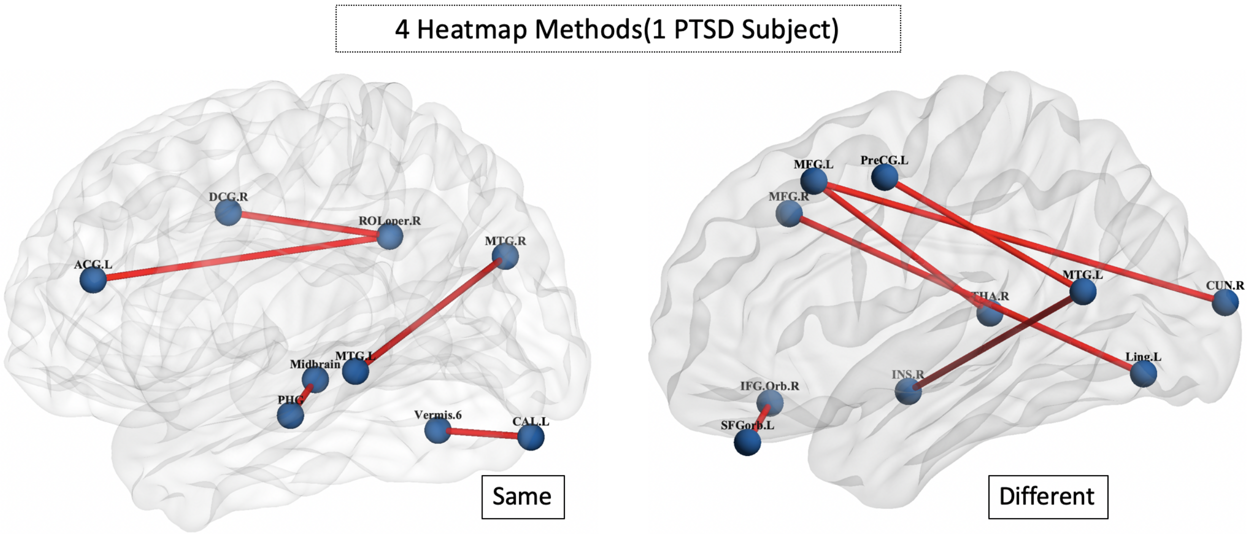


Figure 5.6: Left: This is a figure is showing which paths were common among all heatmaps generated for one PTSD subject. : Right: This figure shows which were not common among any of the four heatmaps of the PTSD subject. These algorithms include gradient, integrated gradients, LRP-z and LRP-epsilon method. The heatmaps were generated by the 96% accurate PTSD classification CNN model (trained by 10-fold cross validation).

# Chapter 6

## Statistical analysis and Consensus Heatmaps

The raw values in the heatmaps generated from the interpretability algorithms do not have the same distribution and their min and max (range) heatscores are different across interpretability methods. Each subject is different from one another. Therefore, even when the classification decisions are same, the explanations generated may look very different. We needed a systematic way of identifying the significant paths that make the subjects of two classes different from each other. This type of analysis would make our conclusions regarding the interpretability algorithms and the disorders more usable. Identification of significant brain paths (using heatmaps) gives us an understanding of the trained models as well as resting state brain connectivity signatures of mental disorders. For the identification of significant paths, we consider each path (or the corresponding heatscore in the heatmaps) as an independent entity. A hypothesis testing procedure is performed to see that, whether for a given path, the heatscore values in the PTSD group are different from those in the control group. The chose the permutation method for hypothesis testing because we do not have a closed form analytical null distribution for heatscore values. We tried two different strategies to generate heatmaps corresponding to a given disorder; they both involve a permutation test. The first method uses a permutation test and employs difference of means as the test statistic. The second method uses a t-test for the difference of means followed by a permutation test for correcting for multiple comparisons; the permutation test in this method uses cluster mass statistic.

## 6.1 Permutation Test with Difference of Means Statistic

A permutation test [66, 67] is a statistical significance testing method, also called exact test or a randomization test. In this test, we generate a very large number of permutations of our data in order to determine an empirical null distribution of a given test statistic. In this distribution we measure the area under the distribution for values greater than or less than (if it is a two-sided test, else just greater than) the corresponding test statistic (for example, mean difference between two groups) value obtained from real data. The area under the distribution represents the probability or p-value and if it is less than 0.05 we can conclude that there is less than 5% probability of the value from real data belonging to the null distribution. In case the test statistic is the mean difference between the groups, we can say that the groups are significantly different from one another.

Figure 6.1 shows our process of generating a statistically significant heatmap using permutation testing. Starting with path1, we enlist the path1 heatscores for our all subjects from both classes and calculate the test statistic, which is the difference of means of each group. The actual list of heatmap values for a path1 is permuted a large number of times (for example: 1000 permutations for PTSD experiment). For each permutation we measure and enlist the difference of means statistic. From the distribution of these test statistics we calculate a p-value for path1. To calculate the p-value we locate the real test statistic value in the distribution. The proportion of same of better statistic values in the total permutations gives us this probability value. This process is repeated for each FC path, to fill a matrix with p-values.

### 6.1.1 Correcting for Multiple Comparisons

In general, when multiple hypothesis tests happen in parallel (like in our case), the likelihood of getting at least one Type-I error increases with the number of tests performed. Type-I error or false positive is rejection of null hypothesis, while the null hypothesis is true. The probability of occurrence of a false positive in a single hypothesis test is $\alpha$, also called the significance level. The probability of at least one false positive for multiple tests can be given by the equation below. Here, $\alpha*$ is the significance level of one test and $\alpha$ is the significance level of $n$ tests.

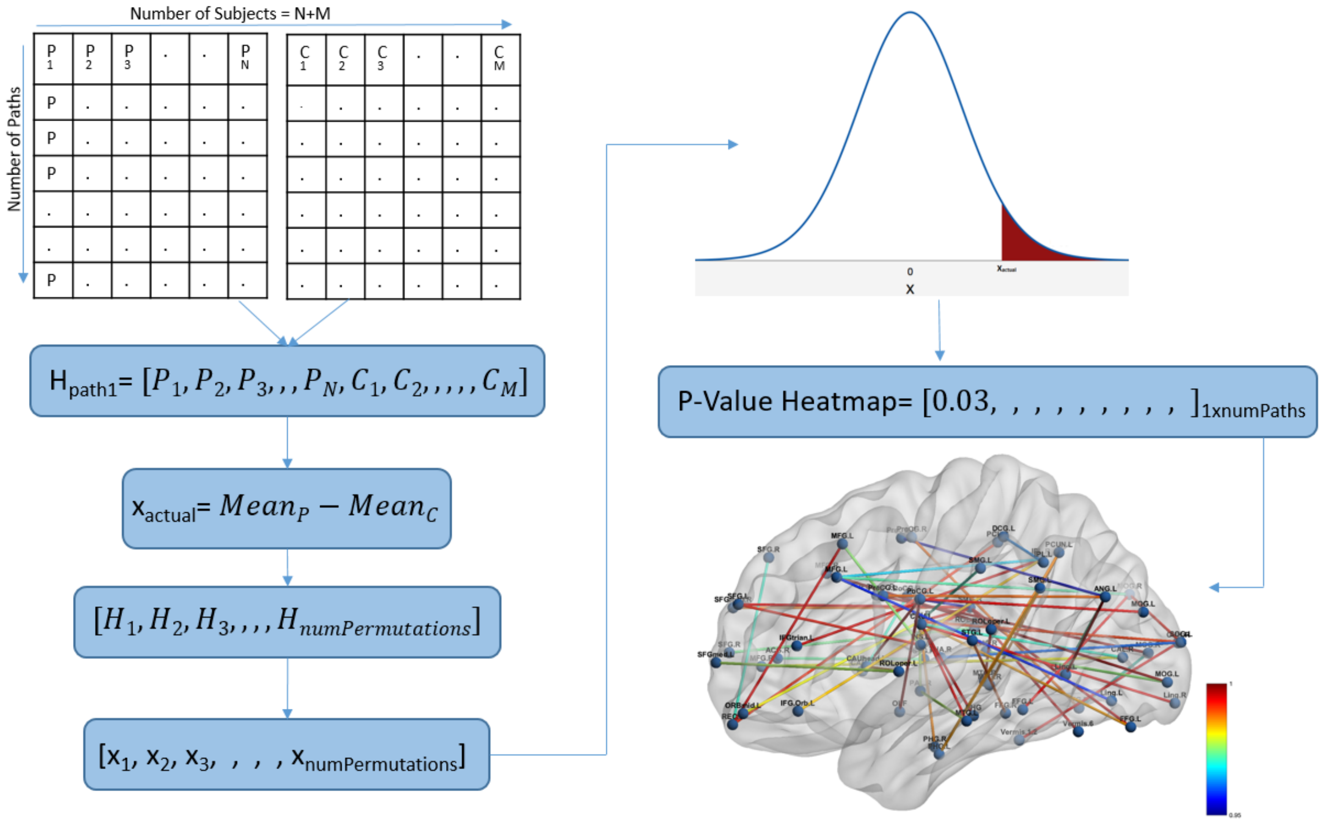

Number of Subjects = N+M
Number of Paths
P 1
P 2
P 3
P N
C 1
C 2
C 3
C M
$H_{path1}= [P_1, P_2, P_3,,, P_N, C_1, C_2,,,,, C_M]$
$x_{actual}= Mean_P - Mean_C$
$[H_1, H_2, H_3,,, H_{numPermutations}]$
$[x_1, x_2, x_3, , , , x_{numPermutations}]$
0
X
$x_{actual}$
P-Value Heatmap= [0.03, , , , , , , , , ]$_{1xnumPaths}$

Figure 6.1: Matrices contain the raw heatmaps generated using interpretability methods, and each matrix represents one model class (PTSD / Control). $H_{path1}$ contains information for path1 from all subjects. A large number of permutations of $H_{path1}$ are generated, and for each permutation the difference of means ($x$) is calculated. A p-value is calculated by seeing where $x_{actual}$ lies in the distribution of test statistic ($x$). The process is repeated to calculate a p-value for each brain path. An FDR corrected p-value threshold is applied to pass only significant brain paths to visualize and analyze.

This reduction of the significance level due to multiple tests is called multiple comparisons problem, because the $\alpha$ accumulates with more comparisons. It is necessary to compensate for the multiple comparisons issue, because $\alpha*$ will result in incorrect hypothesis testing decisions.

$$\alpha = 1 - (1 - \alpha*)^n \tag{6.1}$$

There are several techniques that help compensate for the problem of multiple comparisons such as Bonferroni correction [68], FWER (family-wise error rate) correction and FDR (False Discovery rate) correction [69] which uses the Benjamini Hochberg procedure [69] etc. In this permutation test we used FDR method, it minimizes the chances of Type-I error. We did not used FWER and Bonferroni correction because those two methods set the significance level for letting at least one false positive occurrence, and these methods did not allow for identification of any brain paths. For FDR correction at $\alpha$ we apply the BH (Benjamini Hochberg process) in three simple steps. First we sort our all p-values ($P_i$) obtained in total number of tests $n$. From the sorted p-values we find the test with the highest rank $i$ for which the p-value ($P_i$) satisfies the equation 6.2. In this way we declare the tests ranking 1,2,,,i as significant.

$$P_i \leq \alpha(\frac{i}{n}) \tag{6.2}$$

### 6.1.2 Analyzing PTSD CNN Model using Permutation Test

We repeated the above described permutation process on the PTSD/Control groups of heatmaps obtained using each of the interpretability algorithms, to get a heatmap of p-values. Given below, the figures 6.2 and 6.3, show the heatmaps (Gradient and LRP-$z$ respectively) with significant paths calculated through the permutation process. Since we are working with a large number of algorithms so the other 8 such heatmaps can be seen in the appendices section. In this figure, the colorbar differentiates between the relative significance of paths, where blue means the least significant and red means the most significant of all.

Each one of these algorithms identifies a number of paths (with count between 20 and 40) as significant. In figures 6.2, 6.3 and 6.4, we can see that there are considerable number of paths common between the heatmaps generated by these different algorithms. But, at the same time we can also see a lot of paths that are different. The similarity and differences occur due to the same working principle and change of mathematical formulation respectively, to calculate an explanation.

Currently, in theory there is no such interpretabilty method that explains the working of a model perfectly. At the same time, there is no clear understanding, goal or strategy being used in the artificial intelligence community to quantify the relative effectiveness of interpretability heatmaps in practical settings, such as neuroimaging-based diagnostics. In problems of neuroimaging, interpretability is a concern. Having explanations of decisions can be useful in neuroimaging diagnostics for discovery of ground truths. Unavailability of explanations reduces trust in the performance of a study, where machine learning does the decision-making.

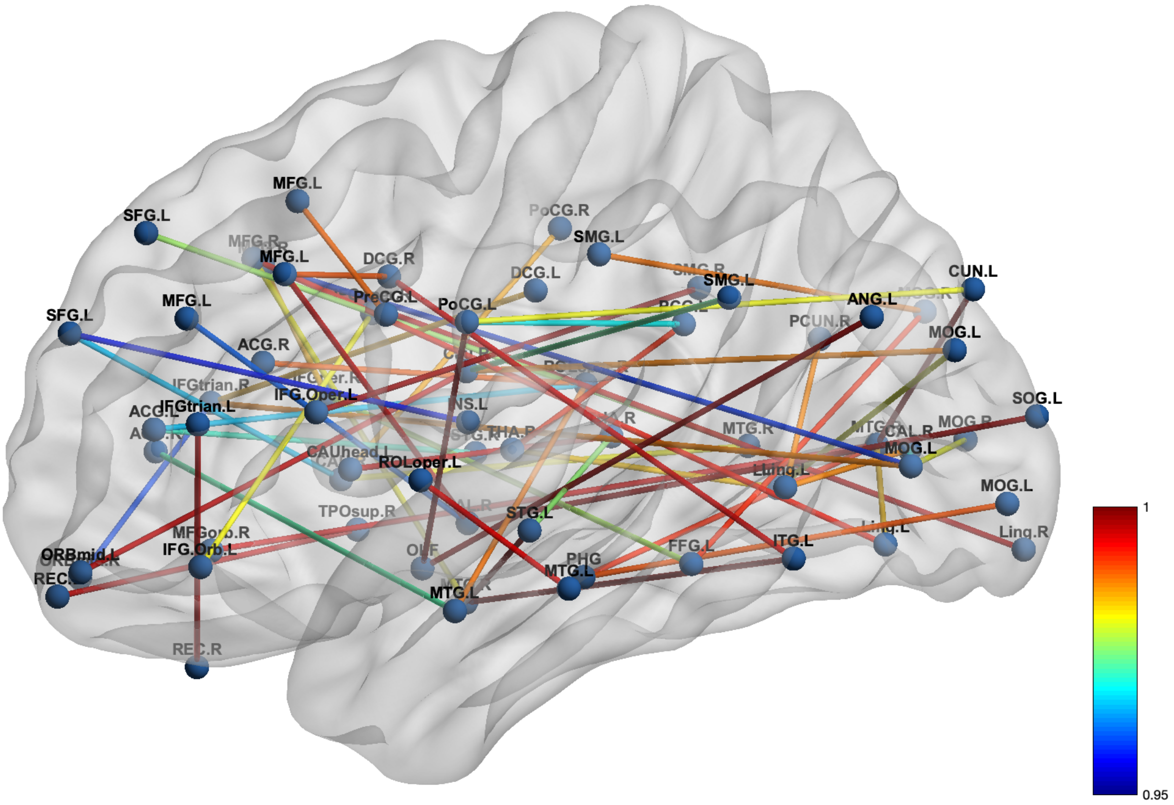


Figure 6.2: Gradient Heatmap of significant paths that make PTSD/Control subjects different one each other. The test statistic in the permutation test was the difference of means. The color bar shows p-values of 0.05 as 0.0 mapped to 0.95 and 1 respectively. The darkest blue path identified least significant path and red most shows the most significant brain path. The raw PTSD and Control groups of heatmaps were generated for a CNN classifier (with 96% cross-validation accuracy).

To tackle these issues described above, it becomes more viable to rely on an explanation generated by the combined effort of all existing interpretability methods. Because, in a consensus the study may incorporate and satisfy assumptions made by different explanation techniques. We use a voting procedure described in next section, where an agreement of all heatmap methods is calculated.

## 6.2 Permutation Test with Cluster Mass Statistic Threshold

In this permutation test we employ a cluster statistic referred to as cluster mass statistic (CMS) first introduced in [70]. This will help us identify significant paths of brain in groups of heatmaps. In this method, the first task is to calculate a matrix of t-statistics applied on the

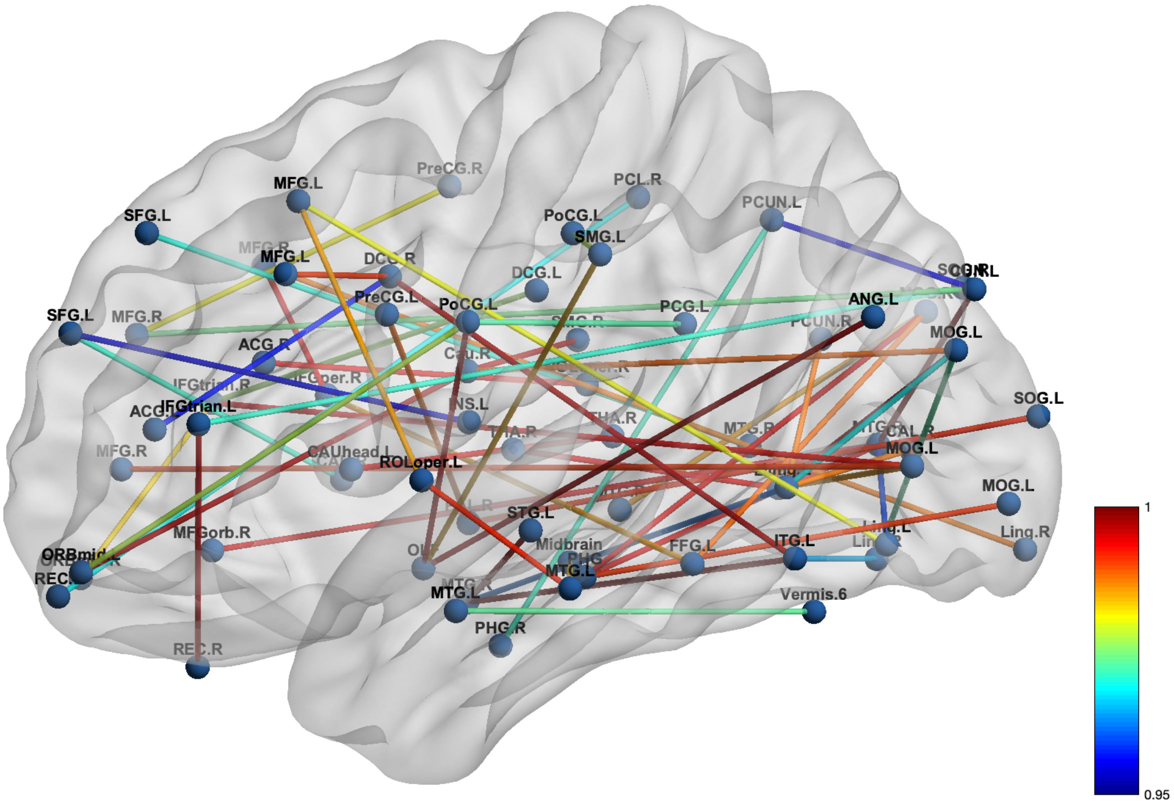
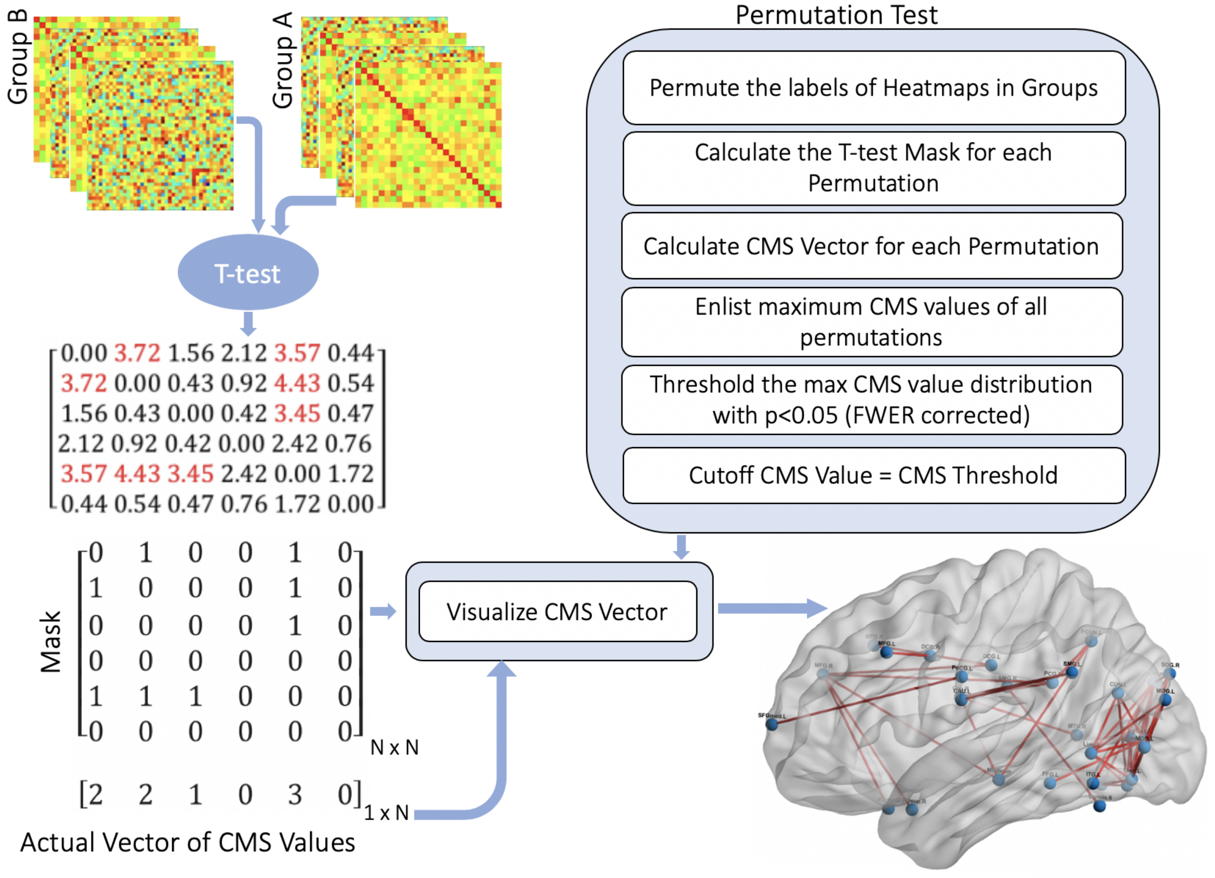


Figure 6.3: LRP-z of significant paths that make PTSD/Control subjects different one each other. The test statistic in the permutation test was the difference of means. The color bar shows p-values of 0.05 as 0.0 mapped to 0.95 and 1 respectively. The darkest blue path identified least significant path and red most shows the most significant brain path. The raw PTSD and Control groups of heatmaps were generated for a CNN classifier (with 96% cross-validation accuracy).

groups of connectivity matrix heatmaps coming from different disorder states (for e.g: Controls/PTSD). A threshold of $p \leq 0.05$ is applied to binarize the matrix of t-values generated by the t-test (in this mask, zeros in this matrix denote insignificant and ones denote significant brain paths). The purpose of this mask is to help with the process of calculating the cluster mass statistic. So, after using a students‚Äô t-distribution for the binary mask, we calculate a cluster mass statistic threshold using a permutation mechanism for finding paths that are significant, with reduced type I error. Reducing type-I error removes the paths that do not make a cluster at their nodes. The workflow of the process is given below and can be seen in figure 6.4.

Groups A and B refer to the two different classes of a disorder. A t-test mask is calculated between the heatmaps in both groups. The values in the binary mask are summed over rows to get a vector of CMS values. Each CMS number in the vector represents the number of links of


Group B
Group A
T-test
Permutation Test
Permute the labels of Heatmaps in Groups
Calculate the T-test Mask for each Permutation
Calculate CMS Vector for each Permutation
Enlist maximum CMS values of all permutations
Threshold the max CMS value distribution with p<0.05 (FWER corrected)
Cutoff CMS Value = CMS Threshold
Mask
N x N
1 x N
Visualize CMS Vector
Actual Vector of CMS Values

$$\begin{bmatrix} 0.00 & 3.72 & 1.56 & 2.12 & 3.57 & 0.44 \\ 3.72 & 0.00 & 0.43 & 0.92 & 4.43 & 0.54 \\ 1.56 & 0.43 & 0.00 & 0.42 & 3.45 & 0.47 \\ 2.12 & 0.92 & 0.42 & 0.00 & 2.42 & 0.76 \\ 3.57 & 4.43 & 3.45 & 2.42 & 0.00 & 1.72 \\ 0.44 & 0.54 & 0.47 & 0.76 & 1.72 & 0.00 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{N \times N}$$

$$[2 \quad 2 \quad 1 \quad 0 \quad 3 \quad 0]_{1 \times N}$$

Figure 6.4: Raw heatmap information is filled into actual correlation matrices to form groups A and B. A large number of permutations of the matrices are created and a t-test is performed between groups for each permutation. Applying the t-test on each permutation results in a binary mask with ones representing significant content of heatmap and zeros representing insignificant paths. A vector of cluster mass statistics is calculated for the actual data mask and all other permutations. A CMS value for an ROI is measured by counting its connections with other ROIs. From each CMS vector the maximum value is enlisted, and a distribution of max CMS is created. FWER with corrected p-value threshold is applied to calculate a cutoff CMS value. This CMS value threshold is applied on CMS vector of actual data. The thresholded CMS vector is transformed back into a binary correlation mask for visualization and analysis.

the corresponding ROI (node). The labels of heatmap groups are permuted for a large of number times and for each permutation we calculate the t-test mask. A CMS vector is calculated for each permutation to draw a distribution of max CMS values and a threshold is applied on it. This threshold is $p \leq 0.05$ and is corrected for family-wise error rate (FWER); it gives us a cutoff value of max-CMS. The cutoff CMS threshold is applied on actual CMS vector. The thresholded CMS vector and t-test mask matrix is used to visualize the significant connectivity paths.

### 6.2.1 Analyzing the PTSD Heatmap using CMS Permutation Method

Figure 6.5 shows gradient based PTSD heatmap, that identifies the functional connectivity paths that are significantly different between PTSD and Control subjects. A list of these CMS processed heatmaps is available in the supplementary section. We are showing the identified paths equally important in a brain sketch, because the visualization process is more mathematically correct if we avoid the actually passed CMS scores. The nature of these heatmaps seems very different from what we obtained in the simple permutation testing mechanism. In these heatmaps, there is more focus on the clusters instead of single paths.

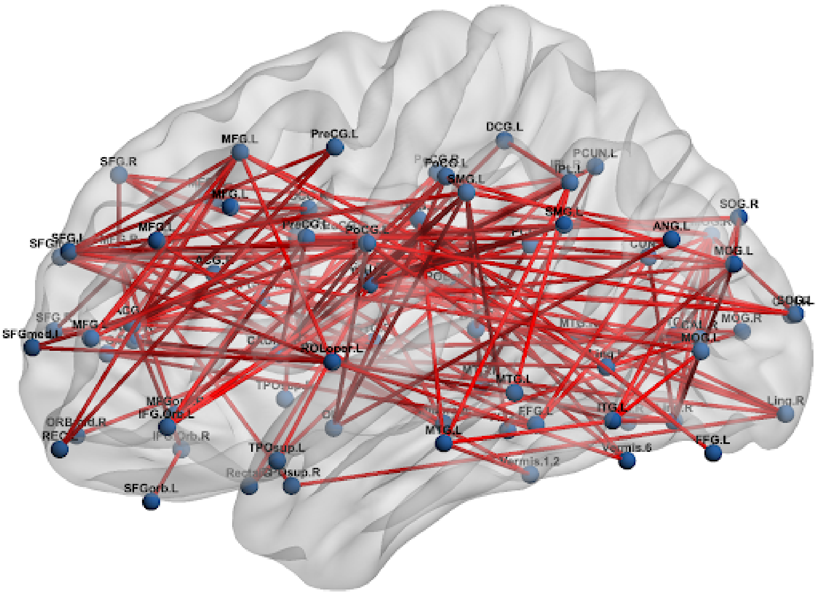


Figure 6.5: Visualization of significant paths generated by gradient heatmap algorithm, that is calculated using the cluster mass thresholding in the permutation test. These significant paths make PTSD/Control subjects different one each other. The test statistics in this permutation test were the difference of means and cluster mass statistic. The raw PTSD and Control groups of heatmaps were generated for a CNN classifier (with 96% cross-validation accuracy) for doing this analysis.

## 6.3 Consensus Heatmaps

Visually, we can easily identify considerable variation in the paths identified by each heatmap algorithm, even after applying the permutations methods. We can see in the figures 6.2, 6.3 and 6.4 that the paths identified (by heatmap methods) after applying the permutation process are not necessarily same. To generate a combined opinion using the information identified by all the heatmap algorithms, we make use of a voting process. In our analysis we are using 10 heatmap algorithms, and explanations from all these algorithms vote for every path to calculate a consensus. In figures 6.6 we have shown the consensus heatmap calculated, employing the permutation test with the difference of means as the test statistic. And, figure 6.7 shows the consensus heatmap calculated using the cluster mass statistic thresholding in the permutation test. In the consensus heatmap, the paths that are identified significant by all the interpretability algorithms are denoted by 1 and paths that are denoted by none of the methods, have a value of zero. For more confidence in the identified regions of interest and clarity of the study, we do not show brain paths that are identified by less than 8 of the interpretability algorithms.

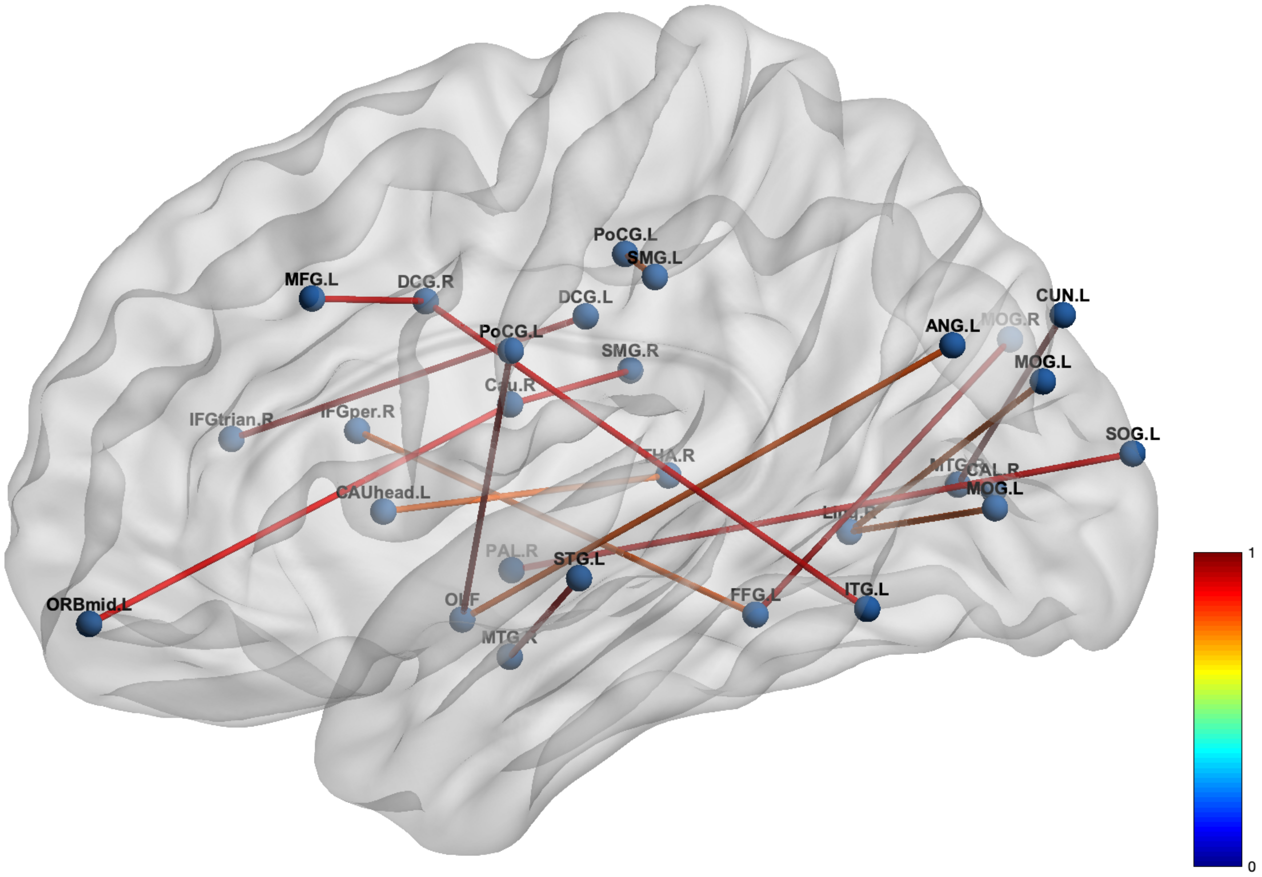


Figure 6.6: PTSD: Consensus heatmap with score 8 or greater, showing the significant paths calculated using the permutation testing with difference of mean statistic. The dark blue color in figure represents paths were common among none of heatmaps and by 1 (red) represent paths that were common among all the heatmap techniques.

PTSD normally occurs as an aftermath of a traumatic experience. Symptoms of PTSD are mainly but not limited to hyper-arousal, tendency to avoid traumatic stimulus, mood changes

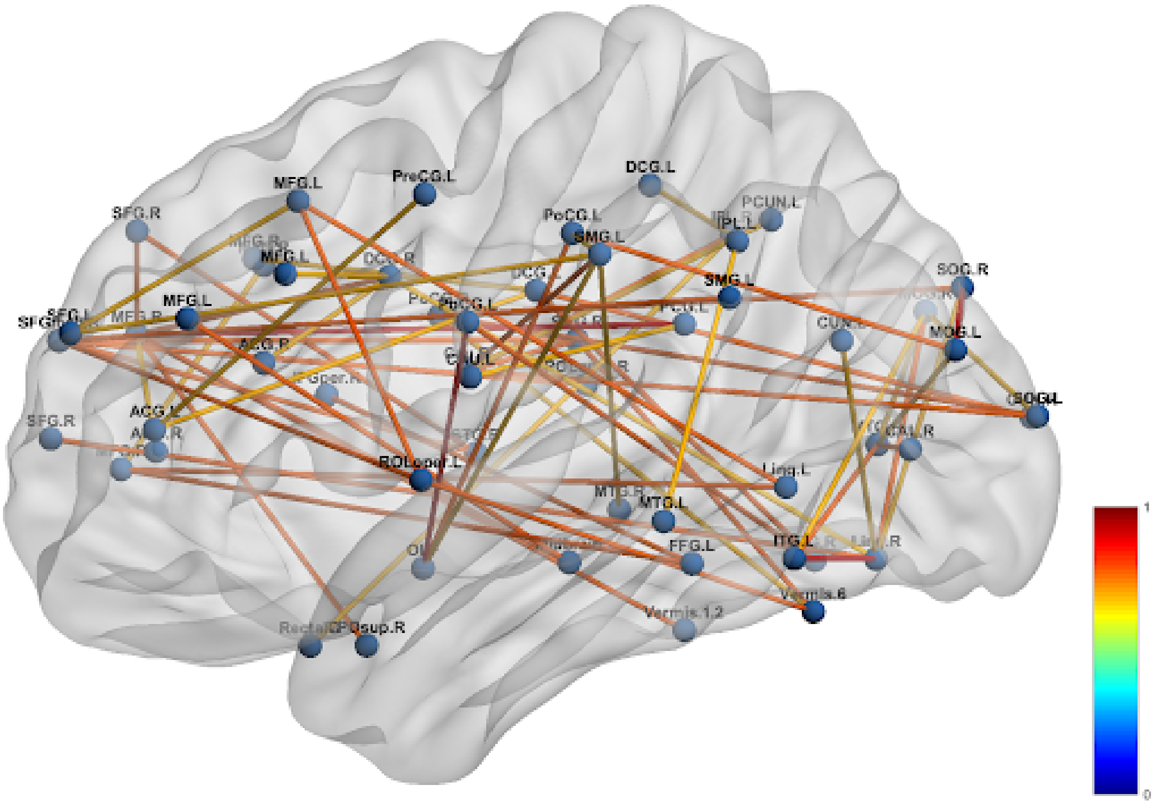


Figure 6.7: PTSD: Consensus heatmap with score 8 or greater, showing the significant paths calculated using the permutation testing with CMS statistic-based thresholding. The dark blue color in figure represents paths were common among none of heatmaps and by 1 (red) represent paths that were common among all the heatmap techniques.

and re-experiencing of traumatic situations. There has been detailed research about PTSD in neuroimaging and we compare our interpretability analysis with their findings. We reviewed 20 articles about the discriminating biomarkers of PTSD and controls, majority of which are resting state studies.

In the consensus brain map of significant paths discriminating PTSD subjects from Controls using simple permutation test method, we found that most of identified paths are existing in the literature. The key areas in the brain map that are supported by other literature are middle frontal gyrus, inferior frontal gyrus, middle temporal gyrus, inferior temporal gyrus and lingual gyrus. A reduction of volume in prefrontal regions and reduced activation in middle frontal gyrus has been seen in Shin 2001 [71, 72], Pitman 2012 [26]. The map also shows some subcortical areas such as thalamus, caudate nucleus and basal ganglia. The occipital discriminating regions have been seen in a few research publications including zhang 2016, lanka 2019, clancy

2020 and Lanius 2005 ([73, 46, 74, 75]). However, most of these works are not able identify why occipital regions‚Äô connectivity differs between PTSD and controls. Clancy in 2020 has observed an attenuation in the power of visual signals seen for PTSD subjects. The false positive nodes that we did not find significant in the publications are postcentral gyrus, angular gyrus and basal ganglia. Table 6.1 gives a list of ROIs identified as important discriminating features in brain map. Cingulate gyrus is an area that regulates the fear response, is affected in PTSD subjects; it was not identified by our simple permutation test explanation [26].

Table 6.1: A summary of the regions of interest associated with discriminating connections identified by permutation testing based consensus explanation. The classification model under consideration is PTSD/Controls Convolutional Neural Network (with 96% cross validation accuracy)

| **ROI Name** | **Abbr** | **Literature** |
|---|---|---|
| Middle Frontal Gyrus | MFG.L | [76, 77, 73, 24, 46] |
| Inferior Frontal Gyrus (triangular part) | IFG.traing.R | [75] |
| Inferior Frontal Gyrus (opercular part) | IFG.oper.R | [75, 78] |
| Middle Frontal Gyrus (orbital part) | ORBmid.L | [76, 79, 73] |
| Median Cingulate and Paracingulate Gyri | DCG.R | [80] |
| Median Cingulate and Paracingulate Gyri | DCG.L | [80] |
| Postcenteral Gyrus | PoCG.L | [] |
| Right Supramarginal Gyrus | SMG.R | [78] |
| Left Supramarginal Gyrus | SMG.L | [78] |
| Superior Temporal Gyrus | STG.L | [75, 81, 73] |
| Middle Temporal Gyrus | MTG.R | [82, 83, 84, 79, 81, 73, 80, 46] |
| Inferior Temporal Gyrus | ITG.L | [82, 76, 79, 83, 84, 85] |
| Fusiform Gyrus | FFG.L | [76] |
| Lingual Gyrus | Ling.L | [78, 85, 46] |
| Middle Occipital Gyrus | MOG.R | [76, 86, 73, 74] |
| Middle Occipital Gyrus | MOG.L | [76, 86, 73, 74] |
| Superior Occipital Gyrus | SOG.L | [76, 75, 74, 46] |
| Calcarine Fissure | CAL.L | [78, 13] |
| Angular Gyrus | ANG.L | [] |
| Caudate Nucleus | Cau.R | [13] |
| Thalamus | THA.R | [75, 87] |
| Basal Ganglia | CAUhead.L | [] |
| Olfactory Cortex | OLF | [78] |
| Pallidum | PAL.R | [81] |

In the consensus explanation, based on cluster mass thresholding in the permutation test, we found discriminating areas between PTSD and Controls that align even better with the existing literature. We found connections associated with anterior cingulate gyrus in this consensus which was not identified by simple permutation testing. This part plays a role in fear response regulation and does not perform its operation normally in PTSD subjects [26]. Connections of superior, middle and inferior opercular frontal gyri with other regions were identified in the consensus. Mood alteration is an important aspect of PTSD that may be mediated in part by orbitofrontal and medial cortical networks [88]. In PTSD subjects, an abnormal coactivation of frontal regions with amygdala may be due to the inability of subjects to perform executive actions normally in an emotional context [88]. The consensus also emphasizes on connections of middle temporal, inferior temporal and lingual gyri; this is consistent with the existing literature as well. Inferior parietal gyrus, cuneus, calcarine fissure, olfactory cortex, vermis, caudate nucleus, fusiform gyrus are areas identified in the consensus explanation and can also be seen in literature summarized in Table 6.2.

Table 6.2: A summary of the discriminating regions of interest identified by permutation testing (with Cluster mass thresholding) based consensus explanation. The classification model under consideration is PTSD/Controls Convolutional Neural Network (with 96% cross validation accuracy)

| ROI Name | Abbr | Literature |
|---|---|---|
| Anterior Cingulate Gyrus | ACG.L | [76, 86, 75, 89, 90, 71, 91] |
| Anterior Cingulate Gyrus | ACG.R | [76, 86, 75, 89, 90, 71, 91] |
| Superior Frontal Gyrus | SFG.R | [76, 75] |
| Superior Frontal Gyrus | SFG.L | [76, 75, 77] |
| Middle Frontal Gyrus | MFG.R | [76, 77, 73, 24, 46] |
| Middle Frontal Gyrus | MFG.L | [77, 78, 72, 73, 24] |
| Inferior Frontal Gyrus (opercular part) | IFGoper.R | [78] |
| Precentral Gyrus | PreCG.L | |
| Postcentral Gyrus | PoCG.L | |
| Postcentral Gyrus | PoCG.R | |
| Median Cingulate and Paracingulate Gyri | DCG.R | [80] |
| Median Cingulate and Paracingulate Gyri | DCG.L | [80] |
| Precuneus | PCUN.L | [81] |
| Supramarginal Gyrus | SMG.L | [78] |
| Inferior Parietal Gyrus | IPL.R | [78] |
| Inferior Parietal Gyrus | IPL.L | [78, 85, 86, 31] |
| Posterior Cingulate Gyrus | PCG.L | [81, 73, 31] |
| Superior Occipital Gyrus | SOG.R | [76, 75, 74, 46] |
| Superior Occipital Gyrus | SOG.L | [76, 75, 74, 46] |
| Middle Occipital Gyrus | MOG.R | [76, 86, 13, 74] |
| Middle Occipital Gyrus | MOG.L | [76, 13, 74] |
| Cuneus | CUN.L | [75, 78] |
| Calcarine Fissure | CAL.R | [78, 73] |
| Lingual Gyrus | Ling.R | [78, 85, 46] |
| Lingual Gyrus | Ling.L | [78, 85, 46] |
| Middle Temporal Gyrus | MTG.R | [82, 83, 84, 79, 81, 73, 80, 46] |
| Middle Temporal Gyrus | MTG.L | [82, 76, 79, 83, 84, 85] |
| Inferior Temporal Gyrus | ITG.L | [82, 76, 79, 83, 84, 85] |
| Fusiform Gyrus | FFG.L | [76] |
| Vermis | Vermis | [78] |
| Caudate Nucleus | CAU.R | [73] |
| Caudate Nucleus | CAU.L | [73] |
| Olfactory Cortex | OLF | [78] |
| Gyrus Rectus | REC.G | |

# Chapter 7

## Multiple Sites, Feature Size and Interpretability

Collecting MRI data is an expensive process and has time, capacity and other constraints; therefore, each site is able to acquire a small sample size dataset. For the purpose of generalization of studies and applications in deep learning, researchers do efforts to make use of data from multiple sites to tackle the issue of smaller sample size (ADHD and ABIDE are two such datasets that were put together for form larger sample sizes). When working with multisite MRI data, one of the most significant factors that weakens the quality and performance of studies, is site variation. It refers to the differences that exist between the fMRI scans performed at different sites that are non-neural in origin. The biases can occur due to scanning parameters, preprocessing procedures, differences among vendors, field strength variation, age of subjects and subject bias etc. These biases present one of the biggest challenges in neuroimaging studies, where researchers are not able to fully benefit from the power of employing data from multiple sites, and hence a larger sample. In order to compensate for these biases, mathematical adjustment or harmonization techniques are used.

The ABIDE I dataset (containing the FC matrices of Autism and Controls subjects) was collected from 12 different sites and the data came from a total of 18 MRI machines across different sites worldwide. In this dataset we had the largest sample size of 988 subjects, the number of subjects per site was also not same. Figure 7.1 is a chart that shows the proportion of subjects from different sites.

Initially, the deep learning model shown previously for this dataset performed very unsatisfactorily with an accuracy of 58% for independent test data and 61% for cross-validation. The number of reduced features (after t-test filtering for significant group differences in the training

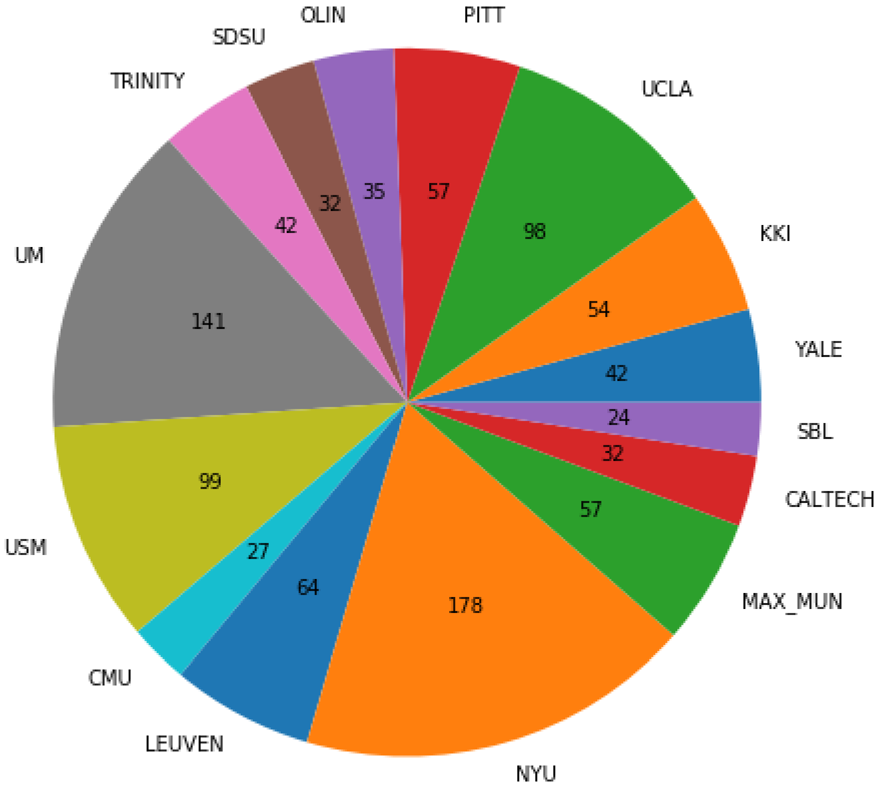


Figure 7.1: Contribution of each site to the ABIDE-I multiple sites dataset. The number in the pie chart indicates the number of subjects in a site.

set) was 1357 FC connections. Given the low accuracies, the corresponding heatmaps were not very meaningful. In fact, a consensus of the heatmaps showed at least 745 paths scattered across the brain. It is clear that the model has not learned enough about the data or the quality of data is inadequate due to site variation.

The 988 subjects could have been helpful to make a highly accurate binary classifier for autism and controls. However, due to low performance, we suspected that variability across sites was impacting classification. In order to address this issue, we used a harmonization technique called ComBat [92] to compensate for the inter-site variability in the ABIDE I dataset. With the ComBat harmonized version of the dataset, we performed a controlled random selection of subjects for the training and testing pools. The training and testing proportion was 80%

and 20% respectively. We randomly picked subjects for training and testing splits, based on the proportion of each site in the dataset. For instance: if site-1 contributed with 200 subjects, then 160 were placed in training set and 40 were placed in testing part. Furthermore, we generated the p-Matrix heatmaps using all methods and again the consensus was far from an interpretation of the disorder, and the improvement was that the decreased number of paths identified in the consensus (421). This confirmed the negative impact of site variation on model and the study itself.

Table 7.1: Comparing the classification and interpretability performance of the ABIDE-I CNN models (with 80% training set and 20% testing set).CV denotes cross-validation and Acc denotes model accuracy.

| | **Epochs** | **Train / Test Acc** | **CV Acc** | **Discriminating paths** |
|---|---|---|---|---|
| No Harmonization | 67 | 60% / 56% | 58% | 745 |
| ComBat Harmonization | 64 | 70% / 60% | 70% | 421 |

In order to take the site variation completely out of question, we trained a separate model (using all the subjects) for data obtained from each site. The training accuracies for these models varied between 56 to 94 percent. Figure 7.2 shows the training accuracies and the sample size at each site. In general, we see less accuracy for smaller sample size and higher accuracy for larger samples (YALE and USM are two sites which do not follow this trend).

We calculated a consensus of the heatmaps generated using all the methods, for each site. These consensus heatmaps across methods showed number of paths between 14 and 111 for YALE, NYU, UM, UCLA and LEUVEN. For the remaining 10 sites we did not find a consensus between the methods. Among the 15 sites data and models, we saw best training accuracy on NYU (94% and 175 subjects) and UM sites (90% and 142 Subjects). We generated the consensus explanation of these two models using CMS thresholding in the permutation test. The consensus was generated across all 10 interpretability methods. In NYU ASD/Controls explanation we saw 27, and for UM we saw 14 discriminating paths respectively. As we look at figures 7.3 and 7.4, we do not see considerable similarity between the two explanations. These are apparently two successful models (considering their training accuracy scores), yet they do not seem to agree on which brain connections are allowing them to discriminate between ASD

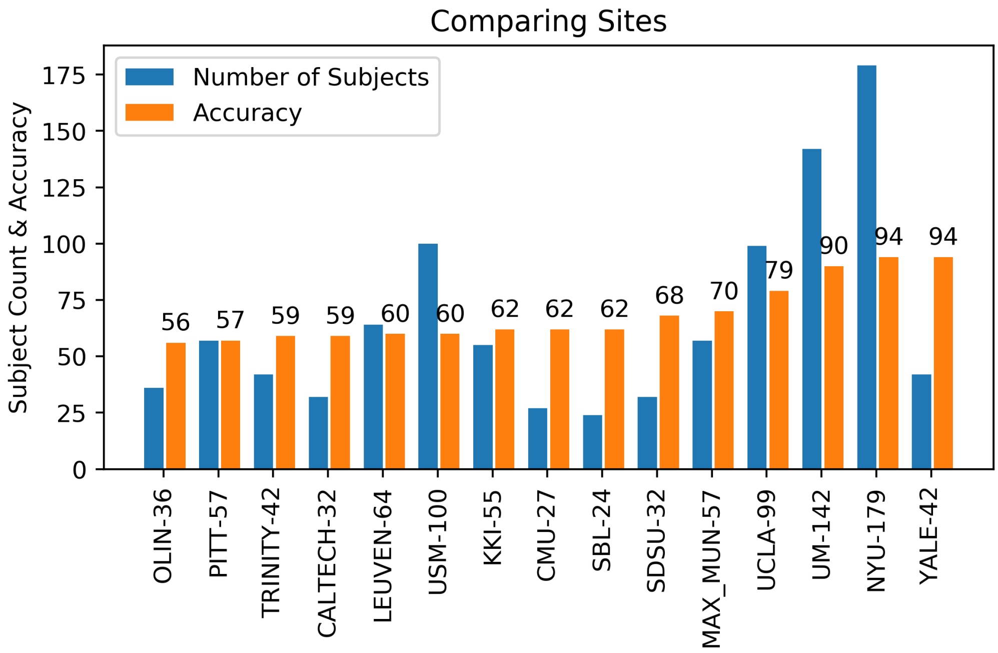


Figure 7.2: A demonstration of the accuracy values for each ABIDE-I site,Äôs model, in comparison with the number of subjects in the site.

and controls. The non-imaging measures of age and intelligence quotient of subjects were significantly different (with p-value less than 0.05) between the NYU and UM sites; this could potentially have contributed to different phenotypes or ASD sub-groups in these two sites, leading to different explanations for classifier performances in these sites.

Among all ABIDE-I sites, we found NYU to be giving the best training accuracy (94%) and it comprised of the largest sample size as mentioned in figure 7.2. So, we compared the consensus explanation of NYU site model with the existing literature to analyze its efficacy. In this consensus map for NYU site, we found several connections associated with brain areas that are known to be different in ASD and control subjects. These areas include anterior cingulate cortex, angular gyrus, precuneus, superior frontal regions, thalamus, superior temporal regions and caudate nucleus. Caudate nucleus, thalamus, precuneus and anterior cingulate cortex were previously identified as ROIs with altered connections by Lanka 2019 using exactly our same dataset. Angular gyrus, medial frontal cortex, precuneus are part of the default mode network, and were disrupted in ASD subjects according to Assaf 2010, Di Martino 2014, Monk 2009 and Washington 2014 [93, 94, 95, 96]. Anterior cingulate cortex and medial frontal regions are known to be affected in subjects with ASD Mundy 2003 [97]. used the ABIDE-I dataset and identified that superior frontal gryi, anterior cingulate cortex and thalamus are among the most affected regions in ASD subjects.

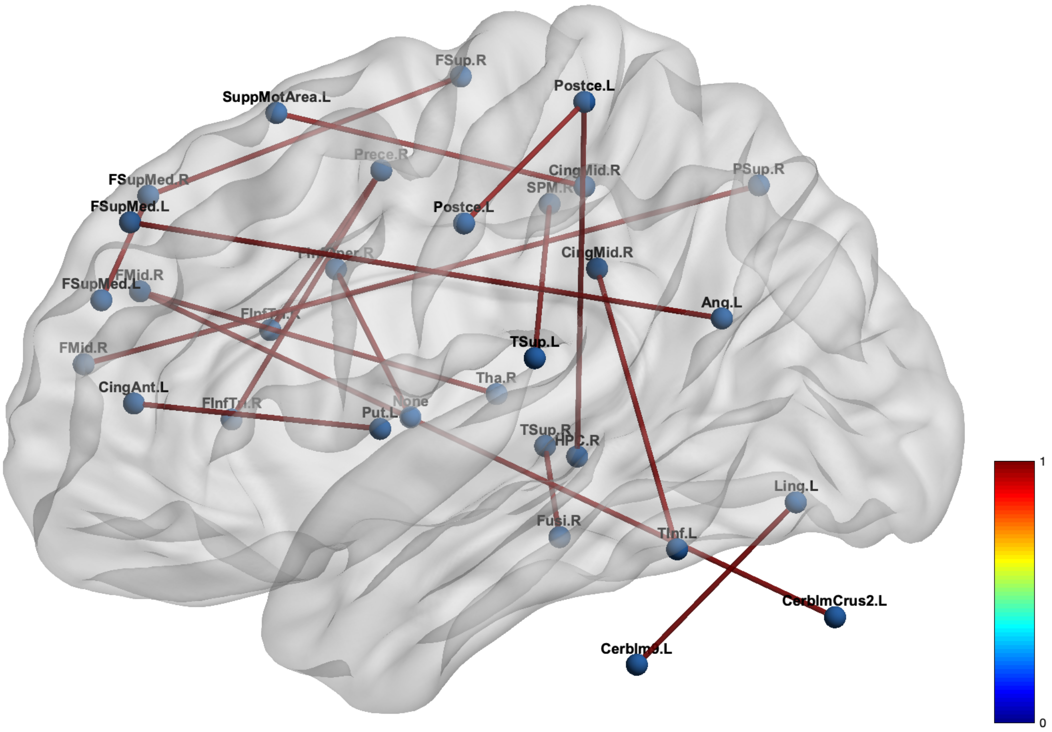


Figure 7.3: A consensus heatmap generated by calculating a consensus between 10 interpretability methods for data acquired at UM site from ABIDE-I dataset. A score of 1 means a path was identified my all interpretabilty algorithms. This is consensus is thresholded at 0.9, meaning agreement between 9 or more of the 10 interpretability methods.

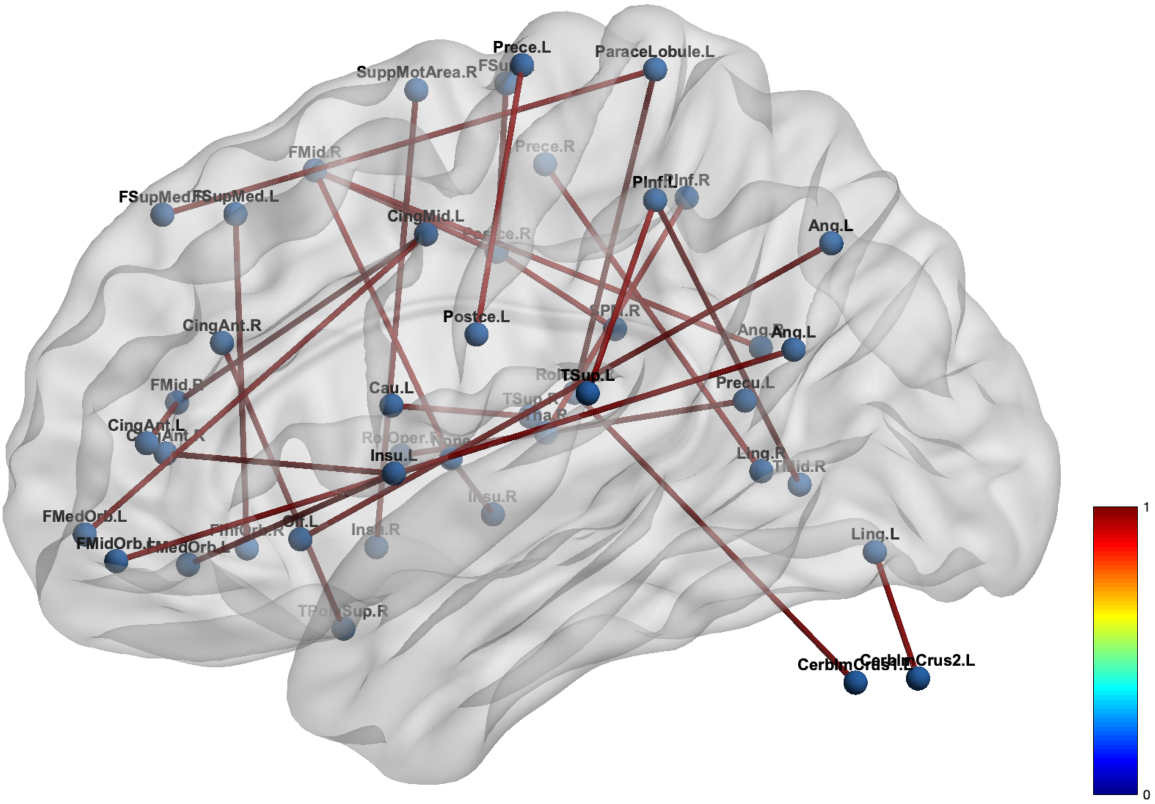


Figure 7.4: A consensus heatmap generated by calculating a consensus between 10 interpretability methods for data acquired at NYU site from ABIDE-I dataset. A score of 1 means a path was identified my all interpretabilty algorithms. This is consensus is thresholded at 0.9, meaning agreement between 9 or more of the 10 interpretability methods.

# Chapter 8

# Conclusion

This work concludes that robust interpretability of neuroimaging classifiers requires substantially higher accuracy than is currently possible with many neuroimaging datasets. Raw heatmap explanations of a deep neural network classifier did not highlight the same discriminating paths for different subjects of the same class. Neither did we see robust similarity in explanation of a model and subject using different interpretability algorithms for multi-site data. The numbers obtained in raw explanations are not constrained to any limits,it is better to using hypothesis testing methods to classify the features as important and idle. The underlying working principles of different interpretability algorithms is not same and they make different assumptions about data and model characteristics. Therefore, we have proposed a consensus approach in this thesis by investigating converging evidence from 10 different types of interpretability algorithms. Permutation tests applied on large pools of heatmaps generated using an interpretability method, helped quantify the understanding of deep learning models with significant connectivity paths. These paths were not exactly same across different interpretability algorithms. So, we calculated a consensus between the results generated by different heatmap algorithms to get to a small significant subset of totals paths. For PTSD classification model with cross-validation accuracy of 96%, we achieved consensus heatmaps, which aligned well with existing knowledge about the disorder.

For ASD versus Controls classification model for the entire multi-site ABIDE data, we did not have a good accuracy and hence the consensus also did not convey useful information. With ComBat harmonized ABIDE dataset, we saw an incremental increase in the model accuracy, however this improvement (70% cross-validation accuracy) was not good enough to produce a good explanation. Although the ABIDE-I dataset did not produce great result due to its multisite nature, we observed a better consensus explanation in successful ASD models trained on individual site data under controlled conditions.

Robust interpretability across methods and models requires substantially higher accuracy than is currently possible in many neuroimaging datasets. This should be a cautionary tale for researchers wanting to use interpretability of artificial neural networks in neuroimaging.

In future, the classification and interpretability of multiple site neuroimaging datasets may improve, by using the harmonization techniques and models that are able to discriminate between sites using the meta-information regarding the scans. Graph convolutional neural networks are another type of deep learning model that might give better results, for classification of resting state functional connectivity data of brain disorders. Brute force method of interpretability includes the methods, where a feature is removed and its effects on output probability are analyzed. This can be a technique which may even help reduce the search space in p-Matrices calculated by permutation testing.

# Bibliography

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.

[2] Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8673–8683, 2020.

[3] Klaus Möllenhoff, Ana-Maria Oros-Peusquens, and N Jon Shah. Introduction to the basics of magnetic resonance imaging. In Molecular Imaging in the Clinical Neurosciences, pages 75–98. Springer, 2012.

[4] Helen H. Shen. Core Concept: Resting-state connectivity. Proceedings of the National Academy of Sciences of the United States of America, 112(46):14115–14116, November 2015.

[5] Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magnetic Resonance in Medicine, 34(4):537–541, 1995.

[6] Marcus E. Raichle, Ann Mary MacLeod, Abraham Z. Snyder, William J. Powers, Debra A. Gusnard, and Gordon L. Shulman. A default mode of brain function. Proceedings of the National Academy of Sciences, 98(2):676–682, 2001.

[7] Michael D. Greicius, Ben Krasnow, Allan L. Reiss, and Vinod Menon. Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences, 100(1):253–258, January 2003.

[8] J. S. Damoiseaux, S. a. R. B. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann. Consistent resting-state networks across healthy subjects. Proceedings of the National Academy of Sciences, 103(37):13848–13853, September 2006.

[9] Philipp G Sämann, Renate Wehrle, David Hoehn, Victor I Spoormaker, Henning Peters, Carolin Tully, Florian Holsboer, and Michael Czisch. Development of the brain's default mode network from wakefulness to slow wave sleep. Cerebral cortex, 21(9):2082–2093, 2011.

[10] J. L. Vincent, G. H. Patel, M. D. Fox, A. Z. Snyder, J. T. Baker, D. C. Van Essen, J. M. Zempel, L. H. Snyder, M. Corbetta, and M. E. Raichle. Intrinsic functional architecture in the anaesthetized monkey brain. Nature, 447(7140):83–86, May 2007.

[11] H. Lv, Z. Wang, E. Tong, L.M. Williams, G. Zaharchuk, M. Zeineh, A.N. Goldstein-Piekarski, T.M. Ball, C. Liao, and M. Wintermark. Resting-State Functional MRI: Everything That Nonexperts Have Always Wanted to Know. American Journal of Neuroradiology, page ajnr;ajnr.A5527v1, January 2018.

[12] Karl J. Friston. Functional and Effective Connectivity: A Review. Brain Connectivity, 1(1):13–36, January 2011.

[13] Dongyang Zhang and Marcus E. Raichle. Disease and the brain's dark energy. Nature Reviews Neurology, 6(1):15–28, January 2010.

[14] Lucina Q. Uddin, Kaustubh Supekar, and Vinod Menon. Reconceptualizing functional brain connectivity in autism from a developmental perspective. Frontiers in Human Neuroscience, 7:458, 2013.

[15] Yuhui Du, Zening Fu, and Vince D. Calhoun. Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. Frontiers in Neuroscience, 12, 2018.

[16] Jill Fain Lehman. The diagnostic and statistical manual of mental disorders. Citeseer, 2000.

[17] Centers for Disease Control, Prevention, et al. Autism and developmental disabilities monitoring network, united states, 2006. Surveill Summ MMWR, 58:1–20, 2009.

[18] Alan M Leslie and Uta Frith. Autistic children's understanding of seeing, knowing and believing. British Journal of Developmental Psychology, 6(4):315–324, 1988.

[19] Serge Gauthier, Barry Reisberg, Michael Zaudig, Ronald C. Petersen, Karen Ritchie, Karl Broich, Sylvie Belleville, Henry Brodaty, David Bennett, Howard Chertkow, Jeffrey L. Cummings, Mony de Leon, Howard Feldman, Mary Ganguli, Harald Hampel, Philip Scheltens, Mary C. Tierney, Peter Whitehouse, Bengt Winblad, and International Psychogeriatric Association Expert Conference on mild cognitive impairment. Mild cognitive impairment. Lancet (London, England), 367(9518):1262–1270, April 2006.

[20] Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 7(3):270–279, May 2011.

[21] W. J. Strittmatter, A. M. Saunders, D. Schmechel, M. Pericak-Vance, J. Enghild, G. S. Salvesen, and A. D. Roses. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proceedings of the National Academy of Sciences of the United States of America, 90(5):1977–1981, March 1993.

[22] George G Glenner and Caine W Wong. Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. Biochemical and biophysical research communications, 120(3):885–890, 1984.

[23] Alzheimer's Association et al. 2018 alzheimer's disease facts and figures. Alzheimer's & Dementia, 14(3):367–429, 2018.

[24] Masaya Misaki, Raquel Phillips, Vadim Zotev, Chung-Ki Wong, Brent E. Wurfel, Frank Krueger, Matthew Feldner, and Jerzy Bodurka. Connectome-wide investigation of altered resting-state functional connectivity in war veterans with and without posttraumatic stress disorder. NeuroImage: Clinical, 17:285–296, January 2018.

[25] Charles W. Hoge, Jennifer L. Auchterlonie, and Charles S. Milliken. Mental health problems, use of mental health services, and attrition from military service after returning from deployment to Iraq or Afghanistan. JAMA, 295(9):1023–1032, March 2006.

[26] Roger K. Pitman, Ann M. Rasmusson, Karestan C. Koenen, Lisa M. Shin, Scott P. Orr, Mark W. Gilbertson, Mohammed R. Milad, and Israel Liberzon. Biological studies of post-traumatic stress disorder. Nature Reviews. Neuroscience, 13(11):769–787, November 2012.

[27] D. Rangaprakash, Michael N. Dretsch, Jeffrey S. Katz, Thomas S. Denney Jr., and Gopikrishna Deshpande. Dynamics of Segregation and Integration in Directional Brain Networks: Illustration in Soldiers With PTSD and Neurotrauma. Frontiers in Neuroscience, 13, 2019.

[28] Ziqi Tang, Kangway V. Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J. Keiser, and Brittany N. Dugger. Interpretable classification of Alzheimer,Äôs disease pathologies with a convolutional neural network pipeline. Nature Communications, 10(1):2173, December 2019.

[29] Jun Shi, Xiao Zheng, Yan Li, Qi Zhang, and Shihui Ying. Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of

Alzheimer's Disease. IEEE Journal of Biomedical and Health Informatics, 22(1):173–183, January 2018.

[30] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage, 101:569–582, November 2014.

[31] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and Plain Convolutional Neural Networks for 3D Brain MRI Classification. arXiv:1701.06643 [cs], January 2017. arXiv: 1701.06643.

[32] Armin W Thomas, Hauke R Heekeren, Klaus-Robert Müller, and Wojciech Samek. Analyzing neuroimaging data through recurrent deep learning models. Frontiers in neuroscience, 13:1321, 2019.

[33] Hongyoon Choi, Kyong Hwan Jin, Alzheimer‚Äôs Disease Neuroimaging Initiative, et al. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behavioural brain research, 344:103–109, 2018.

[34] Lizhen Shao, Donghui Zhang, Haipeng Du, and Dongmei Fu. Deep Forest in ADHD Data Classification. IEEE Access, 7:137913–137919, 2019.

[35] Atif Riaz, Muhammad Asad, Eduardo Alonso, and Greg Slabaugh. Fusion of fMRI and non-imaging data for ADHD classification. Computerized Medical Imaging and Graphics, 65:115–128, April 2018.

[36] Zhenyu Mao, Yi Su, Guangquan Xu, Xueping Wang, Yu Huang, Weihua Yue, Li Sun, and Naixue Xiong. Spatio-temporal deep learning method for ADHD fMRI classification. Information Sciences, 499:1–11, October 2019.

[37] Tianyi Wang and Sei-ichiro Kamata. Classification of Structural MRI Images in Adhd Using 3D Fractal Dimension Complexity Map. In 2019 IEEE International Conference on Image Processing (ICIP), pages 215–219, September 2019. ISSN: 1522-4880.

[38] Gopikrishna Deshpande, Peng Wang, D. Rangaprakash, and Bogdan Wilamowski. Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data. IEEE Transactions on Cybernetics, 45(12):2668–2679, December 2015.

[39] Yazhou Kong, Jianliang Gao, Yunpei Xu, Yi Pan, Jianxin Wang, and Jin Liu. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. Neurocomputing, 324:63–68, January 2019.

[40] Taban Eslami, Vahid Mirjalili, Alvis Fong, Angela R. Laird, and Fahad Saeed. ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. Frontiers in Neuroinformatics, 13, 2019.

[41] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. NeuroImage: Clinical, 17:16–23, 2018.

[42] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. Journal of neuroscience methods, 274:141–145, 2016.

[43] Kanghan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. Classification and Visualization of Alzheimer,Äôs Disease using Volumetric Convolutional Neural Network and Transfer Learning. Scientific Reports, 9(1):1–16, December 2019.

[44] Johannes Rieke, Fabian Eitel, Martin Weygandt, John-Dylan Haynes, and Kerstin Ritter. Visualizing Convolutional Networks for MRI-based Diagnosis of Alzheimer's Disease. arXiv:1808.02874 [cs], 11038:24–31, 2018. arXiv: 1808.02874.

[45] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage, 146:1038–1049, 2017.

[46] Pradyumna Lanka, D. Rangaprakash, Michael N. Dretsch, Jeffrey S. Katz, Thomas S. Denney, and Gopikrishna Deshpande. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. Brain Imaging and Behavior, November 2019.

[47] Pradyumna Lanka, D. Rangaprakash, Sai Sheshan Roy Gotoor, Michael N. Dretsch, Jeffrey S. Katz, Thomas S. Denney, and Gopikrishna Deshpande. MALINI (Machine Learning in NeuroImaging): A MATLAB toolbox for aiding clinical diagnostics using resting-state fMRI data. Data in Brief, 29:105213, April 2020.

[48] PRIME PubMed | DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI.

[49] R. Cameron Craddock, G. Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human Brain Mapping, 33(8):1914–1928, August 2012.

[50] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016.

[51] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine, 29(6):141–142, 2012.

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, November 1998. Conference Name: Proceedings of the IEEE.

[53] Rafael C Gonzalez, Richard Eugene Woods, and Steven L Eddins. Digital image processing using MATLAB. Pearson Education India, 2004.

[54] Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. arXiv:1811.11553 [cs], April 2019. arXiv: 1811.11553.

[55] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901 [cs], November 2013. arXiv: 1311.2901.

[56] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, pages 1097–1105, Red Hook, NY, USA, December 2012. Curran Associates Inc.

[57] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, pages 807–814, Madison, WI, USA, June 2010. Omnipress.

[58] Abhinav Agrawal and Namita Mittal. Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer, 36(2):405–412, 2020.

[59] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. ICLR, 2015.

[60] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, January 2014.

[61] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs], April 2014. arXiv: 1312.6034.

[62] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825, 2017.

[63] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In Proceedings of the 32Nd International

Conference on Neural Information Processing Systems, NIPS'18, pages 9525–9536, USA, 2018. Curran Associates Inc. event-place: Montr$\sqrt{}$©$al, Canada.$

[64] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 3319–3328. JMLR.org, 2017. event-place: Sydney, NSW, Australia.

[65] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In 2011 International Conference on Computer Vision, pages 2018–2025, November 2011. ISSN: 2380-7504.

[66] Douglas Curran-Everett. Explorations in statistics: permutation methods. Advances in Physiology Education, 36(3):181–187, September 2012. Publisher: American Physiological Society.

[67] Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping, 15(1):1–25, 2002.

[68] Winston Haynes. Bonferroni Correction. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, Encyclopedia of Systems Biology, pages 154–154. Springer, New York, NY, 2013.

[69] Yoav Benjamini. Discovering the false discovery rate. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):405–416, 2010. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00746.x.

[70] Alex Ing and Christian Schwarzbauer. Cluster Size Statistic and Cluster Mass Statistic: Two Novel Methods for Identifying Changes in Functional Connectivity Between Groups or Conditions. PLoS ONE, 9(6), June 2014.

[71] L. M. Shin, P. J. Whalen, R. K. Pitman, G. Bush, M. L. Macklin, N. B. Lasko, S. P. Orr, S. C. McInerney, and S. L. Rauch. An fMRI study of anterior cingulate function in posttraumatic stress disorder. Biological Psychiatry, 50(12):932–942, December 2001.

[72] Lisa M. Shin, Scott P. Orr, Margaret A. Carson, Scott L. Rauch, Michael L. Macklin, Natasha B. Lasko, Patricia Marzol Peters, Linda J. Metzger, Darin D. Dougherty, Paul A. Cannistraro, Nathaniel M. Alpert, Alan J. Fischman, and Roger K. Pitman. Regional cerebral blood flow in the amygdala and medial prefrontal cortex during traumatic imagery in male and female Vietnam veterans with PTSD. Archives of General Psychiatry, 61(2):168–176, February 2004.

[73] Qiongmin Zhang, Qizhu Wu, Hongru Zhu, Ling He, Hua Huang, Junran Zhang, and Wei Zhang. Multimodal mri-based classification of trauma survivors with and without post-traumatic stress disorder. Frontiers in neuroscience, 10:292, 2016.

[74] Kevin J Clancy, Jeremy A Andrzejewski, Jessica Simon, Mingzhou Ding, Norman B Schmidt, and Wen Li. Posttraumatic stress disorder is associated with alpha dysrhythmia across the visual cortex and the default mode network. Eneuro, 2020.

[75] Ruth A. Lanius, Peter C. Williamson, Robyn L. Bluhm, Maria Densmore, Kristine Boksman, Richard W. J. Neufeld, Joseph S. Gati, and Ravi S. Menon. Functional connectivity of dissociative responses in posttraumatic stress disorder: A functional magnetic resonance imaging investigation. Biological Psychiatry, 57(8):873–884, April 2005.

[76] J. Douglas Bremner, Meena Narayan, Lawrence H. Staib, Steven M. Southwick, Thomas McGlashan, and Dennis S. Charney. Neural Correlates of Memories of Childhood Sexual Abuse in Women With and Without Posttraumatic Stress Disorder. The American journal of psychiatry, 156(11):1787–1795, November 1999.

[77] Lei Li, Du Lei, Lingjiang Li, Xiaoqi Huang, Xueling Suo, Fenglai Xiao, Weihong Kuang, Jin Li, Feng Bi, Su Lui, Graham J. Kemp, John A. Sweeney, and Qiyong Gong. White Matter Abnormalities in Post-traumatic Stress Disorder Following a Specific Traumatic Event. EBioMedicine, 4:176–183, January 2016.

[78] Feng Liu, Bing Xie, Yifeng Wang, Wenbin Guo, Jean-Paul Fouche, Zhiliang Long, Wenqin Wang, Heng Chen, Meiling Li, Xujun Duan, Jiang Zhang, Mingguo Qiu, and Huafu Chen. Characterization of Post-traumatic Stress Disorder Using Resting-State fMRI with

a Multi-level Parametric Classification Approach. Brain Topography, 28(2):221–237, March 2015.

[79] B. T. Dunkley, S. M. Doesburg, P. A. Sedge, R. J. Grodecki, P. N. Shek, E. W. Pang, and M. J. Taylor. Resting-state hippocampal connectivity correlates with symptom severity in post-traumatic stress disorder. NeuroImage: Clinical, 5:377–384, January 2014.

[80] Hongru Zhu, Junran Zhang, Wang Zhan, Changjian Qiu, Ruizhi Wu, Yajing Meng, Haofei Cui, Xiaoqi Huang, Tao Li, Qiyong Gong, et al. Altered spontaneous neuronal activity of visual cortex and medial anterior cingulate cortex in treatment-naive posttraumatic stress disorder. Comprehensive psychiatry, 55(7):1688–1695, 2014.

[81] Du Lei, Kaiming Li, Lingjiang Li, Fuqin Chen, Xiaoqi Huang, Su Lui, Jing Li, Feng Bi, and Qiyong Gong. Disrupted Functional Brain Connectome in Patients with Posttraumatic Stress Disorder. Radiology, 276(3):818–827, April 2015. Publisher: Radiological Society of North America.

[82] Iris-Tatjana Kolassa and Thomas Elbert. Structural and functional neuroplasticity in relation to traumatic stress. Current directions in psychological science, 16(6):321–325, 2007.

[83] B. Engdahl, A. C. Leuthold, H.-R. M. Tan, S. M. Lewis, A. M. Winskowski, T. N. Dikel, and A. P. Georgopoulos. Post-traumatic stress disorder: a right temporal lobe syndrome? Journal of Neural Engineering, 7(6):066005, December 2010.

[84] Ella L. James, Michael B. Bonsall, Laura Hoppitt, Elizabeth M. Tunbridge, John R. Geddes, Amy L. Milton, and Emily A. Holmes. Computer Game Play Reduces Intrusive Memories of Experimental Trauma via Reconsolidation-Update Mechanisms. Psychological Science, 26(8):1201–1215, August 2015.

[85] Yan Yin, Changfeng Jin, Lisa T Eyler, Hua Jin, Xiaolei Hu, Lian Duan, Huirong Zheng, Bo Feng, Xuanyin Huang, Baoci Shan, et al. Altered regional homogeneity in post-traumatic stress disorder: a restingstate functional magnetic resonance imaging study. Neuroscience bulletin, 28(5):541–549, 2012.

[86] J. Douglas Bremner, Eric Vermetten, Meena Vythilingam, Nadeem Afzal, Christian Schmahl, Bernet Elzinga, and Dennis S. Charney. Neural correlates of the classic color and emotional stroop in women with abuse-related posttraumatic stress disorder. Biological Psychiatry, 55(6):612–620, March 2004.

[87] Yuan Zhong, Ruiting Zhang, Kai Li, Rongfeng Qi, Zhiqiang Zhang, Qingling Huang, and Guangming Lu. Altered cortical and subcortical local coherence in PTSD: evidence from resting-state fMRI. Acta Radiologica, 56(6):746–753, June 2015. Publisher: SAGE Publications.

[88] Lynn D. Selemon, Keith A. Young, Dianne A. Cruz, and Douglas E. Williamson. Frontal Lobe Circuitry in Posttraumatic Stress Disorder. Chronic Stress, 3, May 2019.

[89] Dmitri A. Young, Linda Chao, Thomas C. Neylan, Aoife O'Donovan, Thomas J. Metzler, and Sabra S. Inslicht. Association among anterior cingulate cortex volume, psychophysiological response, and PTSD diagnosis in a Veteran sample. Neurobiology of learning and memory, 155:189–196, November 2018.

[90] Steven H. Woodward, Danny G. Kaloupek, Chris C. Streeter, Matthew O. Kimble, Allan L. Reiss, Stephan Eliez, Lawrence L. Wald, Perry F. Renshaw, Blaise B. Frederick, Barton Lane, Javaid I. Sheikh, Wendy K. Stegman, Catherine J. Kutter, Lorraine P. Stewart, Rebecca S. Prestel, and Ned J. Arsenault. Hippocampal volume, PTSD, and alcoholism in combat veterans. The American Journal of Psychiatry, 163(4):674–681, April 2006.

[91] Noriyuki Kitayama, Sinead Quinn, and J. Douglas Bremner. Smaller volume of anterior cingulate cortex in abuse-elated posttraumatic stress disorder. Journal of affective disorders, 90(2-3):171–174, February 2006.

[92] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1):118–127, January 2007. Publisher: Oxford Academic.

[93] Michal Assaf, Kanchana Jagannathan, Vince D. Calhoun, Laura Miller, Michael C. Stevens, Robert Sahl, Jacqueline G. O'Boyle, Robert T. Schultz, and Godfrey D. Pearlson. Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. NeuroImage, 53(1):247–256, October 2010.

[94] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry, 19(6):659–667, 2014.

[95] Christopher S. Monk, Scott J. Peltier, Jillian Lee Wiggins, Shih-Jen Weng, Melisa Carrasco, Susan Risi, and Catherine Lord. Abnormalities of intrinsic functional connectivity in autism spectrum disorders. NeuroImage, 47(2):764–772, August 2009.

[96] Stuart D. Washington, Evan M. Gordon, Jasmit Brar, Samantha Warburton, Alice T. Sawyer, Amanda Wolfe, Erin R. Mease-Ference, Laura Girton, Ayichew Hailu, Juma Mbwana, William D. Gaillard, M. Layne Kalbfleisch, and John W. VanMeter. Dysmaturation of the default mode network in autism. Human Brain Mapping, 35(4):1284–1296, April 2014.

[97] Peter Mundy. Annotation: the neural basis of social impairments in autism: the role of the dorsal medial-frontal cortex and anterior cingulate system. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 44(6):793–809, September 2003.