ADHD Analysis Using Logistic Regression

Kimberly Beswick, San Diego State University
beswick.kimberly@gmail.com

Abstract - This project is a study of 730 subjects with ADHD-I and controls, numbering 105 and 625 respectively. Tests were run using logistic regression, as well as Dictionary Learning and CaniCA using the Nilearn Toolbox. Each feature was tested individually, showing that the strongest predictor of ADHD-I was the brain region partic_pos_ventral_DMN_07, the posterior ventral brain. Each feature was also tested in groups of 10 from 1000 features. The strongest feature correlation with ADHD were the four scales of ADHD: ADHD.Measure, ADHD.Index, Inattentive, and Hyper.Impulsive, which showed decreasing error as a cumulative effect. Feature selection became necessary as the addition of more features sometimes lead to increasing average and standard error rate. The phenotypic data was collected for the study conducted by Ariana Anderson, Ph.D. et al.: "Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD." One project mentioned in the aforementioned report was the ADHD-200 global ML competition. This study, in addition to using the ADHD
data about brain regions proved useful, and **phenotypic data, uses 30 of the subjects collected in that project to examine fMRI voxels to determine brain regions associated with ADHD. Both Dictionary learning and CanICA are functions of the Nilearn toolbox. The associated regions are shown anatomically in this report.**

Introduction: The problem is a binary classification, and the goal is to determine from phenotypic data whether a subject has ADHD-I or not. the subjects with ADHD-I being of class 1, and the control subjects being of class 0. Logistic regression is used with Glmfit using binomial distribution. Further, for fMRI brain scans the goal is to determine the brain region associated with ADHD and display it graphically.

Challenges: Feature selection was the primary challenge, as in some cases, increasing the number of features used caused the standard and average errors to increase. The goal became one of quality filtration; selecting the features which yielded the lowest errors. In this case, the first four ADHD measures naturally produced the most accurate result. Other single features such as IQ measure and gender also yielded low errors. The first 300 samples of best performance was from the single better than chance, but still produced a higher error than the former data. However, a single anatomical feature, partic_pos_ventral_DMN_07 did produce the lowest average error for a single feature out of any category.

Experiment: The dataset consisted of preprocessed phenotypic data from the report "Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD" which was published on http://www.journals.elsevier.com/neuroimage. The set includes 730 subjects and 1053 useable features. It also includes different types of ADHD; this report focuses on ADHD-I. Tests were run for this report by increasing the features from one to the maximum and tracking the error, by separating the features into consecutive groups of ten from one to one-thousand, by testing each feature individually, and by consciously selecting relevant features. The best result was the posterior ventral brain region for a single feature. Surprisingly, some of the worst results were obtained by testing all of the features at once.

Conclusion: Throughout testing single features, altering training to testing ratios between ⅔ and ⅘,, attempting random permutations of 50 features, and testing sequential sets of 10, the anatomical

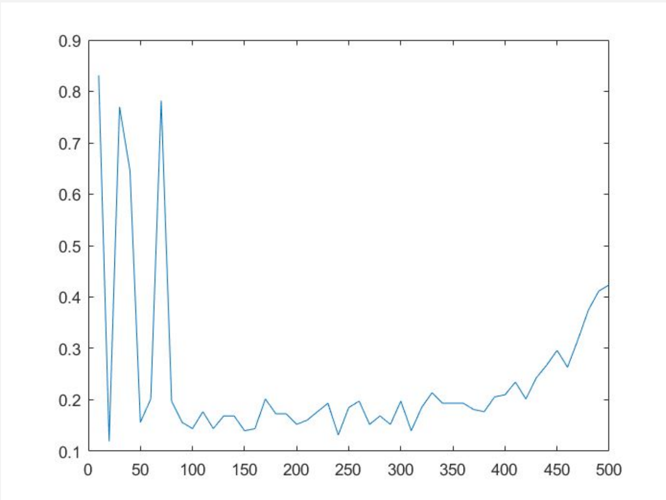
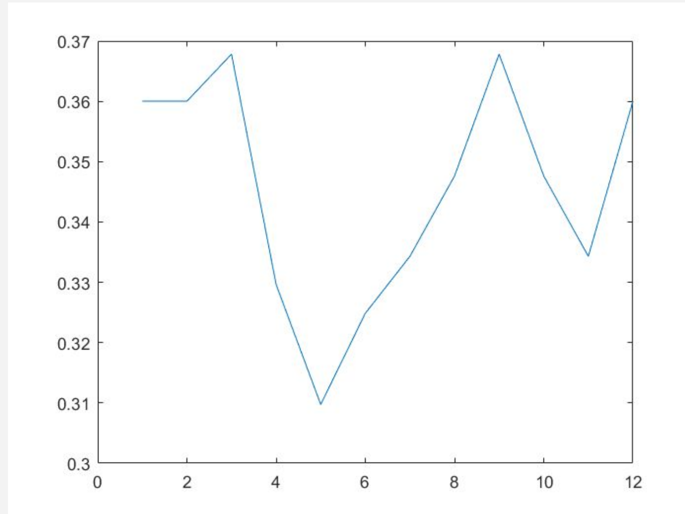
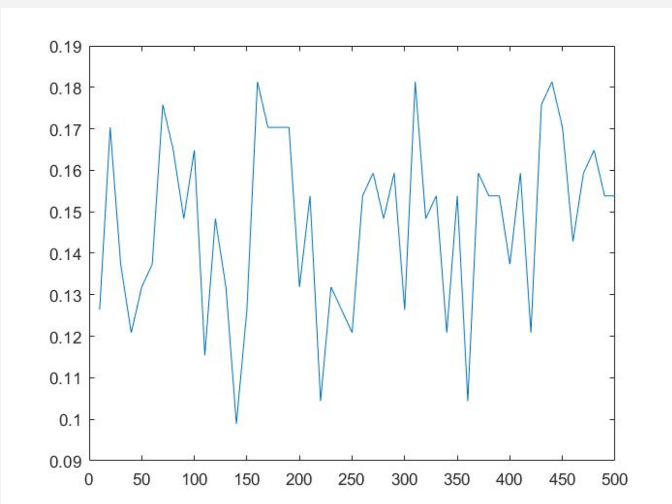
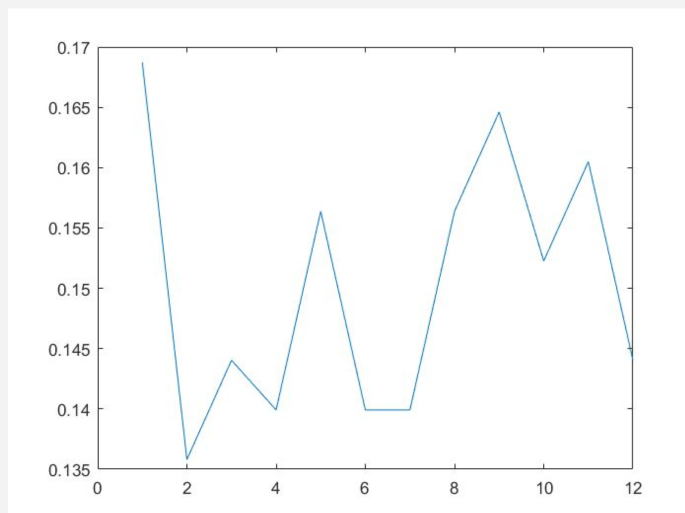
Input of the partic_pos_ventral_DMN_07, or posterior ventral default mode network, at .07 average error rate. There exists data that this region is involved in attention, as studied by Leech R et el. "The posterior cingulate cortex (PCC) is a central part of the default mode network (DMN) and part of the structural core of the brain. Although the PCC often shows consistent deactivation when attention is focused on external events, anatomical studies show that the region is not homogeneous, and electrophysiological recordings in nonhuman primates suggest that it is directly involved in some forms of attention."
(https://www.ncbi.nlm.nih.gov/pubmed/21368033) Analysis of fMRI .nii data using the Nidata toolbox showed activity in the frontal lobes and regions close to the hypothalamus and amygdala. These regions have also been shown to correlate with ADHD symptoms.

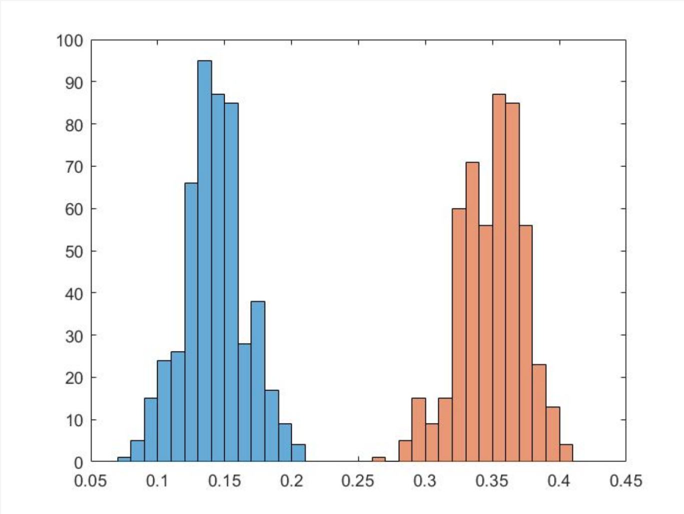
References:

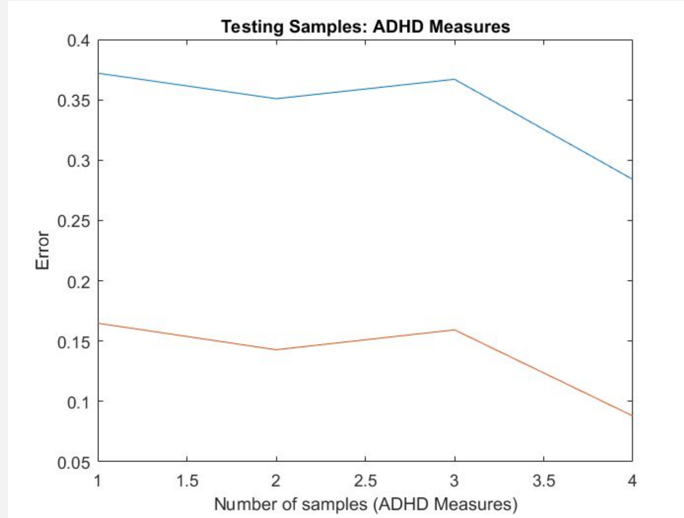
https://www.ncbi.nlm.nih.gov/pubmed/21368033
https://www.ncbi.nlm.nih.gov/pubmed/24361664
http://www.journals.elsevier.com/neuroimage
http://nilearn.github.io/modules/generated/nilearn.decomposition.CanICA.html
http://nilearn.github.io/auto_examples/03_connectivity/plot_extract_regions_dictlearning_maps.html
http://nilearn.github.io/auto_examples/index.html

ADHD-I fMRI Phenotypic Data: Study using Logistic Regression				
			Average error	Standard error
DISTR	Binomial			
Sample size	730			
Training Set:	487			
Test Set	243			
		Number of features:		
		10	0.1	0.3010772027
		20	0.1071428571	0.3104053644
		30	0.1357142857	0.3437146693
		40	0.1571428571	0.3652421799
		400	0.1857142857	0.3902722585
		500	0.2642857143	0.44253544
		600	0.4	0.4916570133
		700	0.4214285714	0.4955609734
		800	0.45	0.4992800572
		900	0.4214285714	0.4955609734
		1000	0.5071428571	0.5017441317
		Single Features:		
**	Best Performing	partic_pos_ventral_DMN_07	0.07143	
		IQ.Measure		
		Training size: 584/730	0.143836	0.352131
		Training size: 547/730	0.126374	0.333187
		Training size: 486/730	0.127572	0.334301
		Male Vs. Female		
		Male % ADHD occurance	0.1776315789	
		Female % ADHD occurance	0.08791208791	
		Female	0.123457	0.32964
		Male	0.152263	0.360018
		M1-M12	0.168724	0.375281

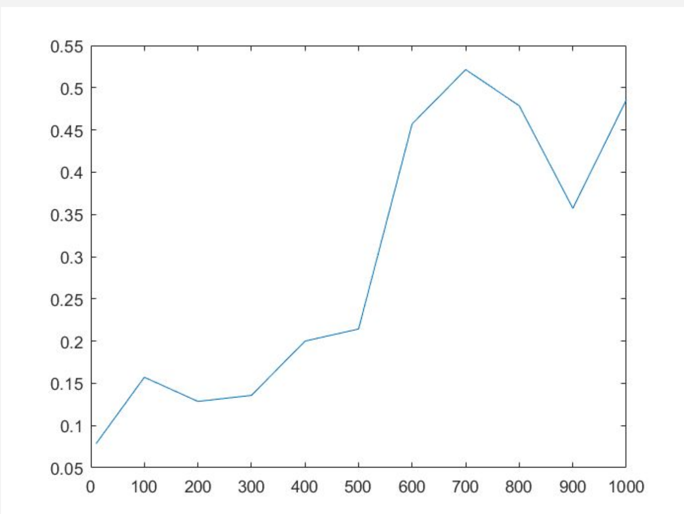
		Anatomical (first 300)		
		300	0.308642	0.462886
			Confusion Matrix	
			154	55
			20	14
			Accuracy: .6914	
		ADHD Measures 1 - 4		
	1	ADHD.Measure	0.1648351648	0.3720552523
	2	ADHD.Index	0.1428571429	0.3508924241
	3	Inattentive	0.1593406593	0.3670031002
	4	Hyper.Impulsive	0.08791208791	0.2839481568
			Confusion Matrix	
			153	8
			8	13
			Accuracy: .9396	
		Random 50 Permutations Features 1 - 1057		
		10 Iterations	0.131868131868	0.339280721422
		50 iterations	0.12637362637	0.333186578505
		90 features	0.	

<table>
<tr><th>All Features: 1 - 500</th><th>M1-M12</th></tr>
<tr><td>First 500 features, increasing from 1 to the specified x value. Average Error:</td><td>M1-M12: Standard error</td></tr>
<tr><td></td><td></td></tr>
<tr><td>First 500 features, groups of 10. Average Error:</td><td>M1-M12: Average Error:</td></tr>
<tr><td></td><td></td></tr>
<tr><td></td><td></td></tr>
<tr><td>First 500 single features: Average error and standard error.</td><td>ADHD Average Error Samples 1 - 4</td></tr>
</table>





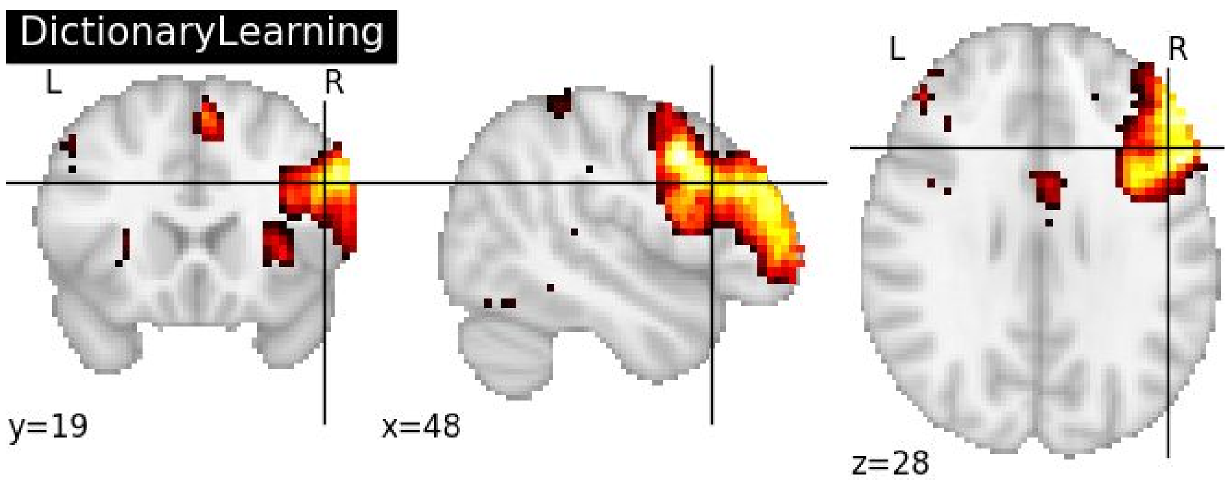
Feature error 1 - 1000, ranging from 1 to N



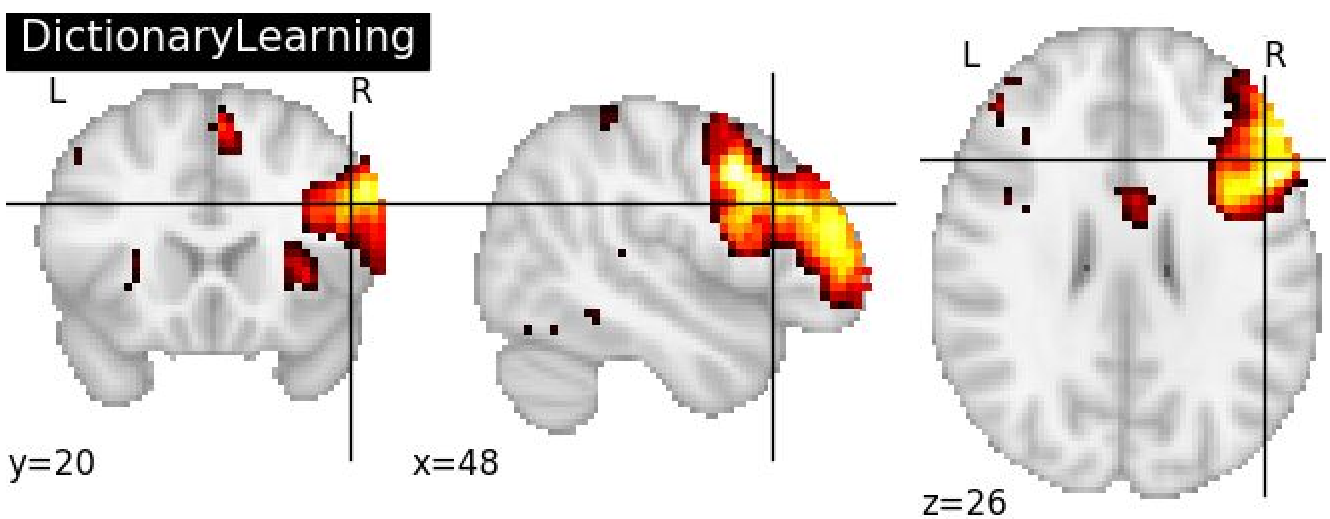
fMRI Data Extraction Using Nilearn Toolbox

Dictionary learning

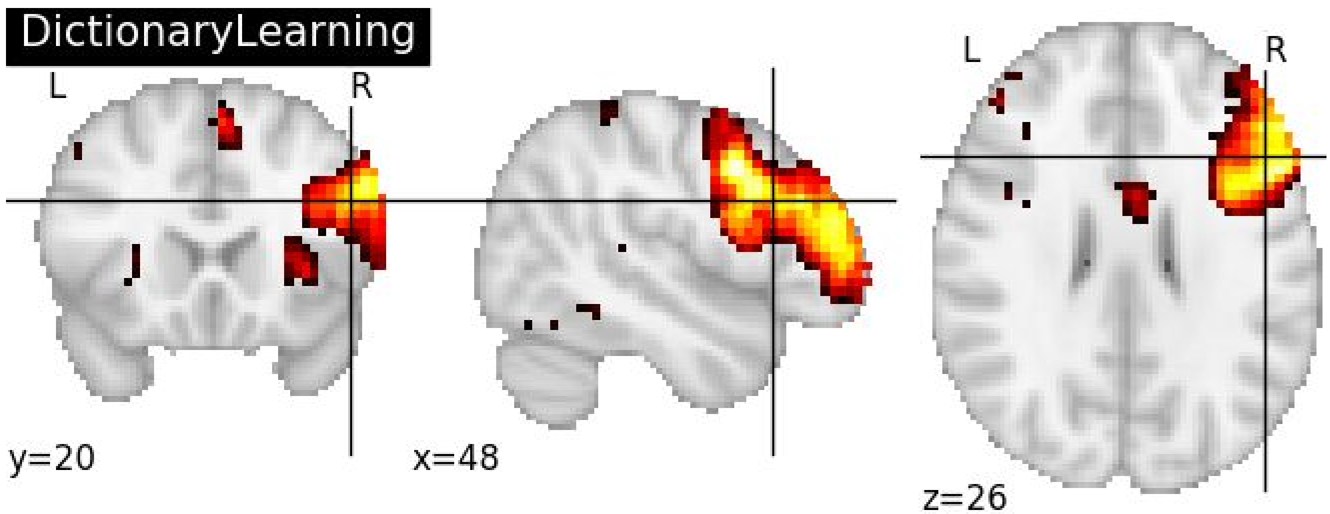
Dataset: Nitrc ADHD-200 Preprocessed
Subjects: 30
Model: Dictionary Learning
Number of epochs: **1**



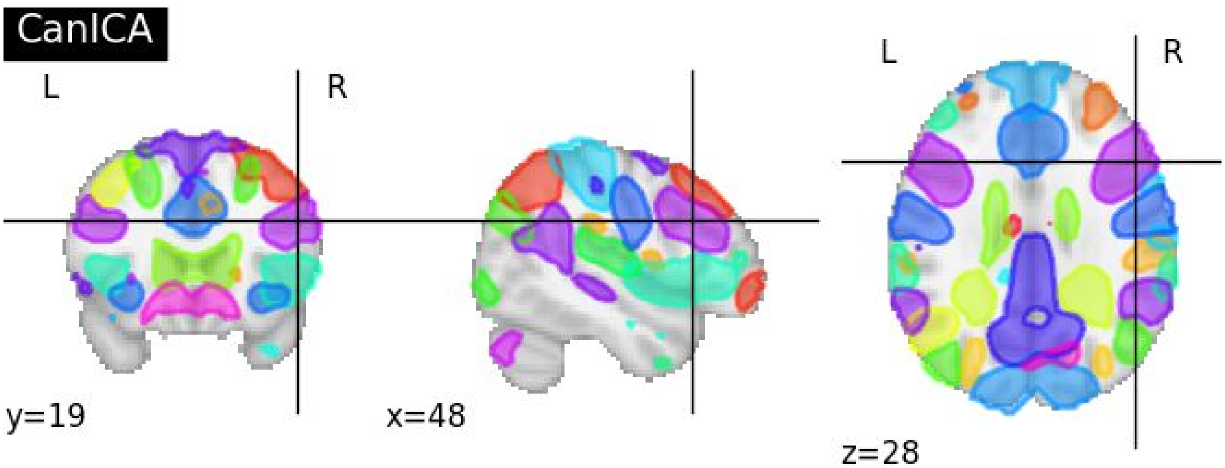
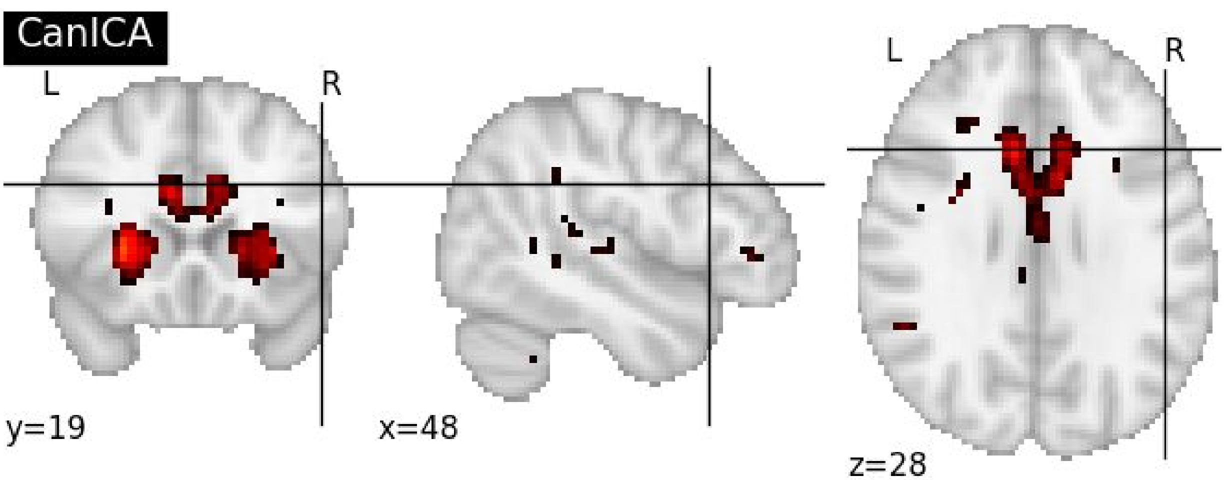
Dataset: Nitrc ADHD-200 Preprocessed
Subjects: 30
Model: Dictionary Learning
Number of epochs: **3**

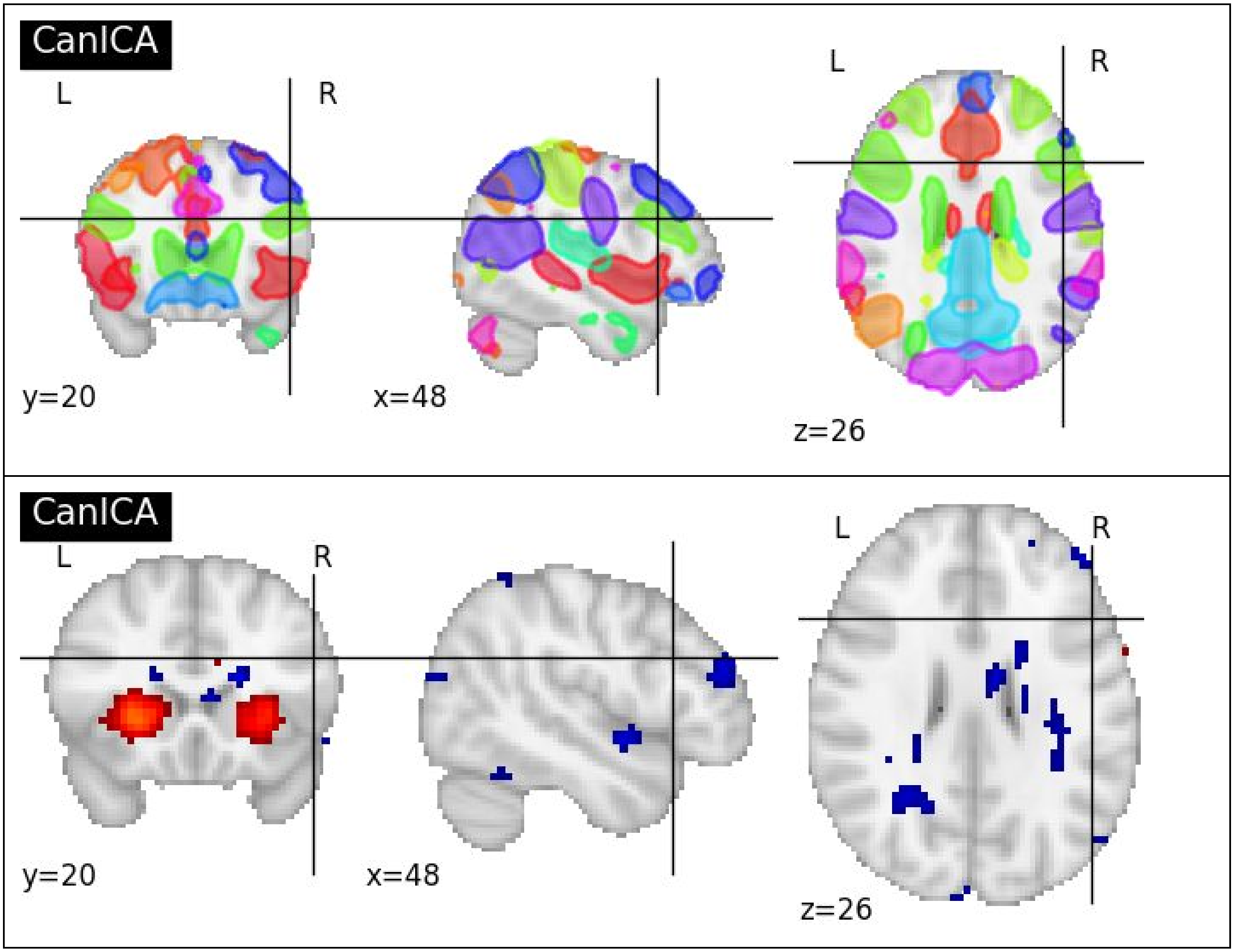


Dataset: Nitrc ADHD-200 Preprocessed
Subjects: 30
Model: Dictionary Learning
Number of epochs: **5**



CanICA

Dataset: Nitrc ADHD-200 Preprocessed Subjects: 30 Model: CaniCA Threshold: **2 x n_voxels**


Dataset: Nitrc ADHD-200 Preprocessed Subjects: 30 Model: CaniCA Threshold: **5 x n_voxels**


CanICA
L
R
y=20
x=48
z=26
CanICA
L
R
y=20
x=48
z=26