Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the worldwide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yujie Zhao Date

Ordinal Support Vector Classifier for Clinical Staging of Major Depression Using Multimodal Imaging

By

Yujie Zhao
Master of Public Health

Department of Biostatistics and Bioinformatics

_____

Ying Guo
Thesis Advisor

_____

Ki Sueng Choi
Thesis Reader

Ordinal Support Vector Classifier for Clinical Staging of Major Depression Using Multimodal Imaging

By

Yujie Zhao

B.S.,
Peking University
2015

Thesis Advisor: Ying Guo, Ph.D.
Thesis Reader: Ki Sueng Choi, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2017

Abstract

Ordinal Support Vector Classifier for Clinical Staging of Major Depression Using Multimodal Imaging
By Yujie Zhao

Introduction: Major depressive disorder (MDD) is a highly widespread, disabling, and pricey illness. Diagnosis and treatment of MDD is considered as a complex problem, because MDD results from a comprehensive interaction of social, psychological and biological factors. Patients received ineffective initial treatment would have significant personal and social costs as well as continued suffering. Identification of biomarkers of MDD in neuroimaging studies is a feasible method to improve diagnostic accuracy and will be helpful to guide treatment selection for individual patients. This study aimed to provide a neurobiological support for practical MDD diagnosis model – clinical staging model by establishing algorithms that discriminate clinical staging subtypes using machine-learning methodology and define most interesting features related to MDD clinical staging model to improve precision of diagnosis and treatment selection for MDD patients.

Methods: Three different treatment status (treatment naïve, treatment responsive recurrent, treatment resistant) and control group were treated as a surrogate of clinical stage for MDD. Two ordinal multiclass support vector classifiers (SVM) were developed to classify subjects into these four clinical stages using functional magnetic resonance imaging data and diffusion tensor imaging data comparing to two traditional multiclass SVM classifiers. SVM recursive feature elimination (SVM-RFE) was applied after each model to select most significant features in this study.

Results: The result of cross-validation indicated that SVM models built on multimodal data have much better classification accuracy than those built on single neuroimaging modal. All-subset ordinal SVM model was more sensitive to ordinal features as well as similar classification accuracy compared to traditional one-to-one SVM model.

Discussion: With the hypothesis of ordinal trend in four clinical stage, all-subset ordinal model is more capable of defining most interesting features related to MDD clinical staging model. Therefore, this model could provide a new strategy using selected significant features for MDD early diagnosis and patients individualized treatment selection.

Ordinal Support Vector Classifier for Clinical Staging of Major Depression Using Multimodal Imaging

By

Yujie Zhao

B.S.,
Peking University
2011

Thesis Advisor: Ying Guo, Ph.D.
Thesis Reader: Ki Sueng Choi, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2017

Table of Contents

1. INTRODUCTION AND REVIEW OF THE LITERATURE

Major depressive disorder (MDD), also known simply as depression, is a common illness worldwide. There are more than 300 million people of all ages suffer from depression all over the world. MDD affects individual's attitude towards life and make person suffer greatly. Nearly 800,000 people die due to suicide every year (World Health Organization, 2017). Although there are many known treatments for depression, effectiveness of these treatments is mainly dependent of patients themselves. Because depression results from a complex interaction of social, psychological and biological factors, there are no definitive algorithms that can directly determine or predict the sufficient and necessary treatment for individual patients. Fewer than 40% of patients achieve remission with initial treatment (C. L. McGrath et al., 2013). Given the public health consequence of ineffective treatment, new perspectives are needed on the biological characteristics of depressive disorders with relationship to disease diagnosis and treatment strategies. Identification of biomarkers was evaluated to have an improvement to guide treatment (H. S. Mayberg, 2003 & R. C. Craddock, 2009). It can be predicted that in near future quantitative measures of brain function will be an essential process to establish optimal treatment for a given patient with major depressive disorder.

In previous neuroimaging studies of MDD patients, biomarkers are defined with a direct association with improvement of a single treatment response (C. L. McGrath et al., 2013 & S. Haller et al., 2014). However, diagnosis and treatment of MDD is a more systematic and comprehensive procedure. New relationship needs to be established

between neurobiological features and practical diagnosis-treatment model of MDD. Clinical staging models have been first proposed as a reliable model for MDD diagnosis over two decades ago (G. A. Fava et al., 1993). It defined the extent of progression of disease at a time point and focused on detailed description of where a person lies currently along the continuum of the course of illness (P. D. McGorry et al., 2006). With construction and extensive research on clinical staging model, this model provided a strategy for development and evaluation for clinical interventions of MDD. However, these was little biological evidence for such classifications.

In recent few years, there have been growing interests in use of machine learning methodology for analyzing neuroimaging data (F. Pereira et al., 2009 & A. Cerasa et al., 2015 & J. R. Sato et al, 2012). Studies have shown that machine learning methods could mine new information from neuroimaging data comparing to traditional approaches (C. Chu et al, 2015). These novel methods, including classifier methods, facilitated analysis of high-dimensional datasets and were less sensitive to noise (I. Guyon et al, 2003). Support vector machine (SVM), which is the most popular multivariate machine learning feature selection method, has been successfully applied to neuroimaging data in many situations. Recent research showed SVM classifiers had several advantages over common univariate methods and could identify features contributes most to subject classification (S. R. Gunn, 1998 & A. M. Andrew, 2000).

In this study, new extensive method based on SVM was developed to classify MDD patients using clinical staging model incorporating neuroimaging biomarkers. Functional magnetic resonance imaging (fMRI) and Diffusion tensor tractography imaging (DTI or DTT) were examined to distinguish significant biological features with association clinical staging. Establishment of multivariate classification methods that

can discriminate MDD stages using a combination of diffusion and functional MRI imaging techniques will provide a new strategy to stratify patients in order to select their optimal treatment at any given time point.

1. METHODOLODY

1.1. Study Overview

In this study, three treatment status (treatment naïve, treatment responsive recurrent, treatment resistant) and control group were treated as a surrogate of clinical stage for MDD and assumed that these four stages have intrinsic ordinal trend from the slightest to the severest on MDD status. Multiclass SVM classifiers were trained to classify individual subjects into four different stages of MDD using functional magnetic resonance imaging data and diffusion tensor imaging data and selected most significant features related to this ordinal clinical staging models.

1.2. Objectives

This study considered two objectives: to develop algorithms that discriminate stage subtypes using multimodal neuroimaging data and to define neurobiological biomarkers of MDD staging modal.

1.2.1. Objective 1: To Train Multiclass Classifiers for Four Ordinal Groups of Outcome

Given that SVM method is a binary classifier, there is no ideal SVM-based methods for multiclass classification (S. R. Gunn, 1998). In this research, two extensive multiclass SVM classifier was developed based on ordinal hypothesis of MDD clinical stages comparing two traditional multiclass approaches. New approaches focused on ordinal features and provided comparable prediction accuracy considering ordinal clinical stages.

2.2.2. Objective 2: To Define Most Significant Features Related to Four-Group Classification

Feature selection are generally including two different processes, filter methods and wrapper methods (I. Guyon et al., 2002). Filter methods treat feature selection as a preprocessing step and remove features based on some criterion. Wrapper methods consider feature selection as an optimization problem and select features with minimal prediction error. In this study, univariate ANOVA, commonly used filter method, was used to test correlation between each feature and outcome and selection features by a threshold. Then multiclass SVM classifiers recursive feature elimination method, a nested iterative wrapper based method, was utilized to find feature subset with minimal prediction error. Finally, since I trained four multiclass SVM classifiers in our analysis, I compared feature subsets selected by all four models and ordinal characteristic of selected features.

2.3. *Statistical learning methods*

2.3.1. *Support Vector Machine (SVM)*

Support vector classification is a statistical learning theory to solve binary classification problems. Given a dataset with N observation each of p input features, SVM classification maps every observation into a point of p-dimensional hyperspace and performs a hyperplane in this hyperspace to discriminate all observations into two classes with the maximal distance between the hyperplane and the nearest observation in either class. In detail, for $X \in R^p$ with corresponding class labels $y \in \{-1,1\}$, SVM define a hyperplane

$$y(X) = w^T X + b$$

Where the parameter w is the normal vector to the hyperplane and $\frac{b}{\|w\|}$ determines the offset of the hyperplane from the origin along the normal vector w. The distance between two hyperplanes

$$w^T X + b = 1 \; and \; w^T X + b = -1$$

is $\frac{2}{\|w\|}$, which is called 'margin'. To maximize the margin, the hyperplane is determined by solving the convex quadratic programming optimization problem

$$\min C \sum_i \xi_i + \frac{1}{2}\|w\|^2$$

Subject to

$$y_i(w^T X + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Where ξ_i is the distance of the i^{th} misclassified observation from its correct side of the margin and the box constraint $C > 0$ controls the degree to which the misclassified data points affect the solution. This problem is solved by adding Lagrange multipliers to

obtain its dual problem. However, more specific details will not be discussed in this work. Once this hyperplane is determined, a classification rule induced is

$$y(X) = sign(w^T X + b)$$

The performance of SVM is evaluated by cross validation, which will be described below.

2.3.2. *Multiclass SVM*

SVM is a binary classifier only works for two-class classification. There are two traditional approaches for multiclass classification based on SVM methods as following (A. Govada, 2015).

1) One versus One: Every binary classifier is built to differentiate between each pair of two classes, while discarding the rest of the classes, which requires constructing N(N-1)/2 binary classifiers. When testing a new object, a voting is performed among the classifiers and the class, and the classifier with the maximum output will be considered as the best choice.

2) One versus All: This simple approach is to decompose the problem of classifying N classes into N binary problems, where each problem differentiates a class versus all other classes. In this approach, I require N binary classifiers, where the Kth classifier is trained with positive examples belonging to class K and negative examples belonging to the other N-1 classes. When to predict a new test subject, the classifier with the maximum output is considered as the best choice, and the corresponding class label is assigned to that test object.

Following are ordinal methods I am working on.

1) In this study, as groups of outcomes have ordinal scale, one intuitive thought is to use an ordinal SVM method. This problem can be decomposed into N-1 binary problems, where the K^{th} classifier is trained with positive examples belong to class 1 to class K and negative examples belong to class K+1 to class N. When to test a new object by these N-1 binary classifiers, the results will become a list of numbers, positive means it belongs to positive group in this binary classifier and negative means it belongs to negative group in this binary classifier. I find the first number S that number in S^{th} position and $(S+1)^{th}$ position is different, and last number T that number in T^{th} position and $(T+1)^{th}$position are different. This object is thought belonging to an open set between group S and group T+1.

2) For more precise results, I consider an iteration for ordinal SVM method showed above. As this is a N-class classification problem, I tried my algorithm as following. For every K in (1, 2, 3, …, N), I choose every possible combination of K groups from N groups and run ordinal SVMs to get results of open sets. Then I combine all results together to find out the group that this new object has the highest probability to be located in.

In this study, the model I am most interested in is all-subset ordinal model. I performed comparisons among all four model, especially the comparison between all-subset ordinal model and most common-used multiclass SVM method – one to one model.

2.3.3. *SVM Recursive Feature Elimination (SVM-RFE)*

SVM Recursive Feature Elimination (SVM-RFE) is commonly used method in genetics to find out the most significant genes based on SVM method (I. Guyon, 2002). It's a stepwise feature selection method using weight magnitude of SVM as ranking criterion.

When using SVM classification, a hyperplane $w^T X + b = 0$ is generated to separate two classes of objects into largest marginal distance. This weight vector $w = \sum \alpha_i Y_i X_i$ is also trained to decide the weight of each features in SVM model (X. Zhou, 2007). Intuitively, those features with the largest weights w^2 is believed most significant. The SVM-RFE method is an iteration training a SVM classifier by using X, Y and selecting the lowest absolute value in ω, then recording its feature index, removing this feature from X, and starting to train a new SVM classifier. The result of this iteration is a feature rank list and most interesting features can be selected in this list.

This method works for a single SVM classifier. To implement to multiclass SVM, I grade every feature in feature rank list, sum up their grade for all binary classifiers and sort these features based on their grade.

2.3.4. *Cross-validation*

In the study, I ran 10-fold cross-validation for 100 times. In detail, I partitioned objects into 10 groups, each round of cross-validation I trained multiclass SVM models based 9 groups and used the last group to validate my result. This process was repeated 10 times (10 folds) in a round. And I did 100 rounds of cross validation by using different partitions. Accuracy was evaluated by the mean proportion of objects classified into correct group.

2.4. *Simulation Study*

At first, 80 datasets were generated (4 settings by 20 replicates in each setting), and each dataset has 400 observations (100 observations in each group) multiplied by 50 features. Every dataset X is trained as following,

$$X = \omega \cdot \beta + \varepsilon, \qquad \omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \end{pmatrix}, \quad \beta = (\beta_1 \quad \beta_2), \quad \varepsilon \sim Multivariate\ Normal(0, \Sigma)$$

Here dataset X is a 400 by 50 matrix, each row represents an observation and each column represents a feature. ω is a 400 by 1 vector showing the setting of ordinal scale observations on all these features, so I used four numbers $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ indicating means of every feature in four groups and each number was duplicated 100 times because of the number of observations in each group. There are 4 different settings. First is linear using {1, 2, 3, 4}; second is ordered but not linear using four sorted random number between 1 and 4; third is not ordered using four random number between 1 and 4; last is combination of all above using one-third of feature from first setting, one-third from second setting and one-third from third setting. β is a 1 by 50 vector representing characteristic of features. It can be separated into two parts $(\beta_1, \ \beta_2)$, where $\beta_1 \sim N(0, \Sigma)$ means all the features in β_1 is related to our outcome and $\beta_2 = 0$ means features in β_2 is not related to our outcome. In this simulation, I chose 6, 12, 18, 24, respectively as number of features associated with our outcome. ε is noise, generated by Normal distribution. At last, all datasets were standardized to Z-score. Our outcome Y is a 400 by 1 vector, and I used {1,2,3,4} as the class label of each group.

2.5. *Experimental Study*

2.5.1. *Dataset*

In this study, I got functional magnetic resonance imaging (fMRI) and diffusion tensor tractography imaging (DTI or DTT) dataset with our features of interest the connectivity between different regions in brain. These datasets were obtained from patients in three major depressive disorder (MDD) treatment status (treatment-naïve, treatment responsive-recurrent and treatment-resistant) as well as control group. In terms of severity of depression, I assumed there is an order, objects in control group (CON) have the slightest MDD status; objects in treatment naïve group (CIDAR) have a slighter MDD status; objects in treatment responsive recurrent group (R01) have a severer MDD status; and objects in treatment resistant group (DBS) have the severest MDD status. Since the exact gradient of four ordinal MDD status wasn't clarified, I set value of {1,2,3,4} as the outcome of CON, CIDAR, R01, DBS group.

In our datasets, connectivity between 85 different brain regions were presented. fMRI dataset for every object is a Z-score 85 by 85 correlation matrix. Since this matrix was symmetric, I treated every cell in upper triangular matrix as a feature, and there were totally 85*84/2 = 3570 features. DTI dataset for every subject was an 85 by 85 matrix, and each cell was the number of streamlines that are not rejected from row region to column region. In every cell the proportion of streamlines in its row was calculated and this proportion of streamlines from row region to column region and that from column region to row region were averaged and treated as a feature. Like fMRI

data, DTI data also had 85*84/2 = 3570 features. As there were many zero in DTI data, logit transformation was not proper for this dataset. For using multimodal data, I directly combined fMRI and DTI data together, so the number of feature in multimodal dataset was 2*3570 = 7140.

2.5.2. *Screening*

Univariate ANOVA was used to compute correlation coefficient between every feature and outcome. Features with P-value < 0.2 was retained for further feature selection step. fMRI and DTI data selected were combined as our final dataset.

3. RESULTS

3.1. *Simulation Study*

3.1.1. *Classification Accuracy*

Table1 Mean classification accuracy of cross-validation for 4 models and 4 feature settings

6 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.6263	0.5258	0.5259	0.6349
One to all	0.5779	0.5109	0.4998	0.6256
Ordinal	0.6444	0.5477	0.4001	0.6365
All-subset ordinal	0.6387	0.5445	0.5081	0.6496

12 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature

One to one	0.7831	0.6035	0.6554	0.7824
One to all	0.6641	0.5552	0.6019	0.7470
Ordinal	0.7883	0.6162	0.4766	0.7807
All-subset ordinal	0.7869	0.6140	0.6185	0.7933

18 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.8627	0.7342	0.6511	0.8525
One to all	0.7036	0.6403	0.5876	0.7927
Ordinal	0.8632	0.7413	0.4618	0.8405
All-subset ordinal	0.8627	0.7402	0.6067	0.8531

24 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.8878	0.7069	0.6946	0.9064
One to all	0.7092	0.6221	0.6043	0.8381
Ordinal	0.8884	0.7164	0.5015	0.8960
All-subset ordinal	0.8881	0.7164	0.6571	0.9070

The results of cross-validation in Table 1 showed that in linear and ordinal feature setting, all-subset ordinal SVM methods had a similar accuracy as one to one method. In not ordinal situation, original ordinal method didn't work very good, however, all-subset ordinal SVM had a comparable results as one-to-one method. In most conditions, one-to-all method was not a good choice.

When the number of features related to outcome increased (compared rows in *figure 1*), the accuracy of cross-validation in all four models also increased. This result indicated that SVM models were quite sensitive for pre-step of feature selection, and additional screening step should be added for real data analysis before utilizing SVM models.

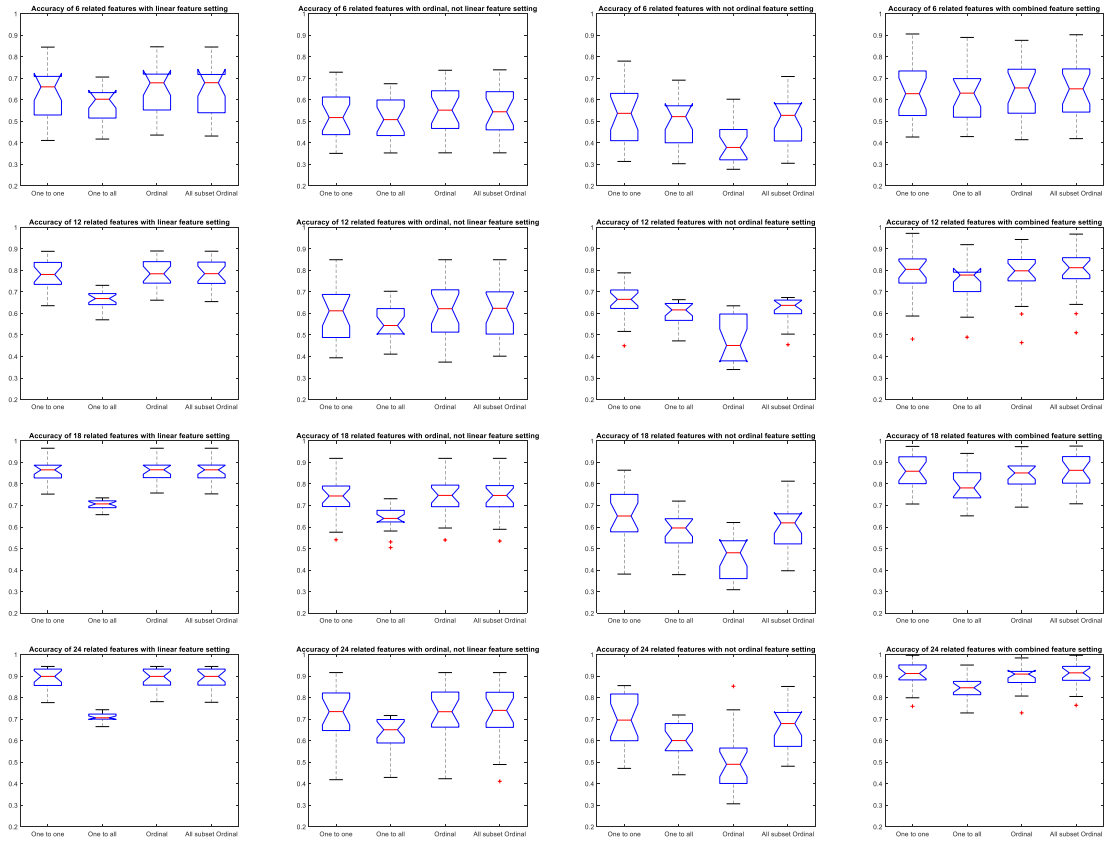

Figure 1 Boxplots for four different number {6, 12, 18, 24} of related features are presented by rows and four different feature settings {linear, ordinal not linear, not ordinal, combined} are presented by columns. In each plot, results of four multiclass SVM {One to one, One to all, Ordinal, All subset ordinal} are presented in order.

3.1.2. *Classification Accuracy*

Because some of features were generated with association to outcome, I considered these features were most significant features that I needed while the other features were treated to be unrelated to outcome. I calculated the proportion whether these significant features were selected from feature space by SVM-RFE. For example, for datasets with 6 significant features and 44 unrelated features, I counted the

proportion of these 6 features presented in the top 6 significant elements of feature space generated by SVM-RFE.

Table 2 Proportion that significant features are selected in feature subset by SVM-RFE on condition of four feature settings

6 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.7917	0.6444	0.6056	0.7306
One to all	0.6542	0.5583	0.5542	0.6292
Ordinal	0.8722	0.7889	0.4667	0.7333
All-subset ordinal	0.8422	0.7265	0.5520	0.7294

12 of 50 features are related to outcome	Feature settings			
	linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.7708	0.6833	0.6764	0.7069
One to all	0.6333	0.6125	0.6229	0.6250
Ordinal	0.8333	0.7722	0.6472	0.7694
All-subset ordinal	0.8064	0.7324	0.6539	0.7348

18 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.8287	0.7602	0.7944	0.7870
One to all	0.7361	0.6931	0.7264	0.7194
Ordinal	0.8574	0.8278	0.7685	0.8093
All-subset ordinal	0.8523	0.7938	0.7791	0.7997

24 of 50 features are related to outcome	Feature settings			
	Linear Feature	Ordinal, not linear Feature	Not ordinal Feature	Combined Feature
One to one	0.8438	0.7792	0.7708	0.8056
One to all	0.7677	0.7135	0.7167	0.7427
Ordinal	0.8778	0.8264	0.7347	0.8278
All-subset ordinal	0.8659	0.8083	0.7475	0.8167

According to results in Table 2, it showed that for those linear and ordinal features, ordinal SVM methods had higher proportion to select these features from

datasets than traditional methods. However, ordinal SVM model could only select the lowest proportion non-ordinal features and in combined feature settings, the result of this model was similar to result from one to one model.

All-subset ordinal SVM model is what I am interested in. This model got a good chance to select linear and ordinal features, just a little lower than ordinal SVM model, and also a good result to select non-ordinal features which was much better than ordinal method. Compared to one to one model, this all-subset ordinal model still had an advantage in ordinal feature selection and a disadvantage in non-ordinal selection. I will discuss this further in discussion section.

3.2. *Experimental Study*

3.2.1. *Screening*

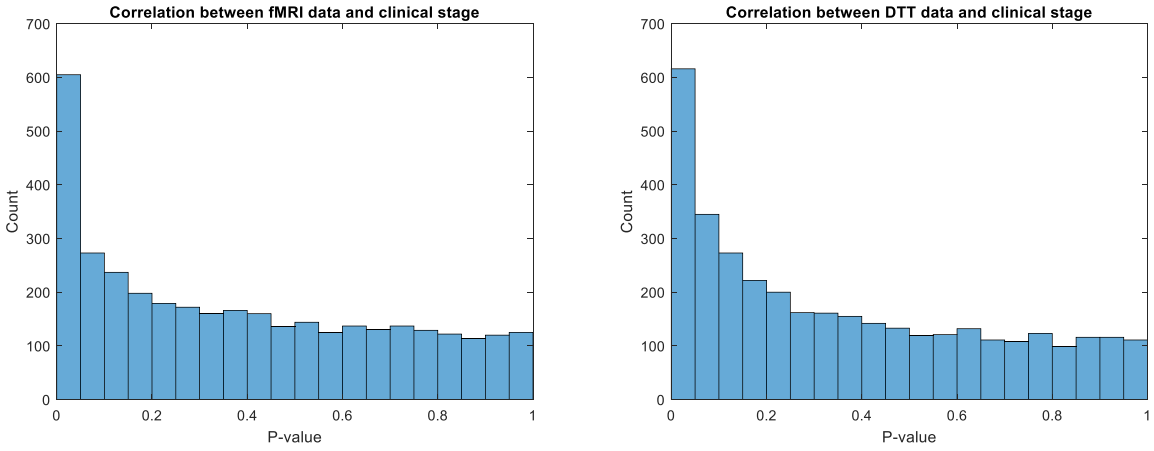

Figure 2 histogram of feature correlation between real data and clinical stage

In all 3570 features of fMRI and DTT data, the correlation between feature and clinical stage is showed in *Figure 2*. There are 1456 (40.8%) features in DTT dataset and

1313 (36.8%) features in fMRI dataset have P-value < 0.2. There are 961 (26.9%) features in DTT dataset and 878 (24.6%) features in fMRI dataset have P-value < 0.1. There are 616 (17.3%) features in DTT dataset and 605 (16.9%) features in fMRI dataset have P-value < 0.05.

3.2.2. *Classification Accuracy*

There was not much difference in mean accuracy when I chose univariate ANOVA P-value = 0.2 or 0.1 or 0.05 as screening threshold see in *Figure 3*. The mean accuracy didn't increase when screening threshold decreased. In this study, I chose 0.2 as my threshold for further feature selection.

Results in Table 3 also indicated that ordinal model didn't work very well in fMRI data, giving that ordinal model are in favor of ordinal features, it showed in fMRI data ordinal features are not dominated. In the other side, the mean accuracy of ordinal method in DTI data is high against one to one model. So there are probably more ordinal features in DTI dataset. The result in multimodal dataset was much higher than the other two groups. Owing that the number of features in multimodal dataset was the combination of fMRI and DTI dataset, this result couldn't directly show the advantage of multimodal method. However, compared to fMRI and DTI dataset separately, it is certain that multimodal data could receive a better prediction accuracy in MDD clinical staging.

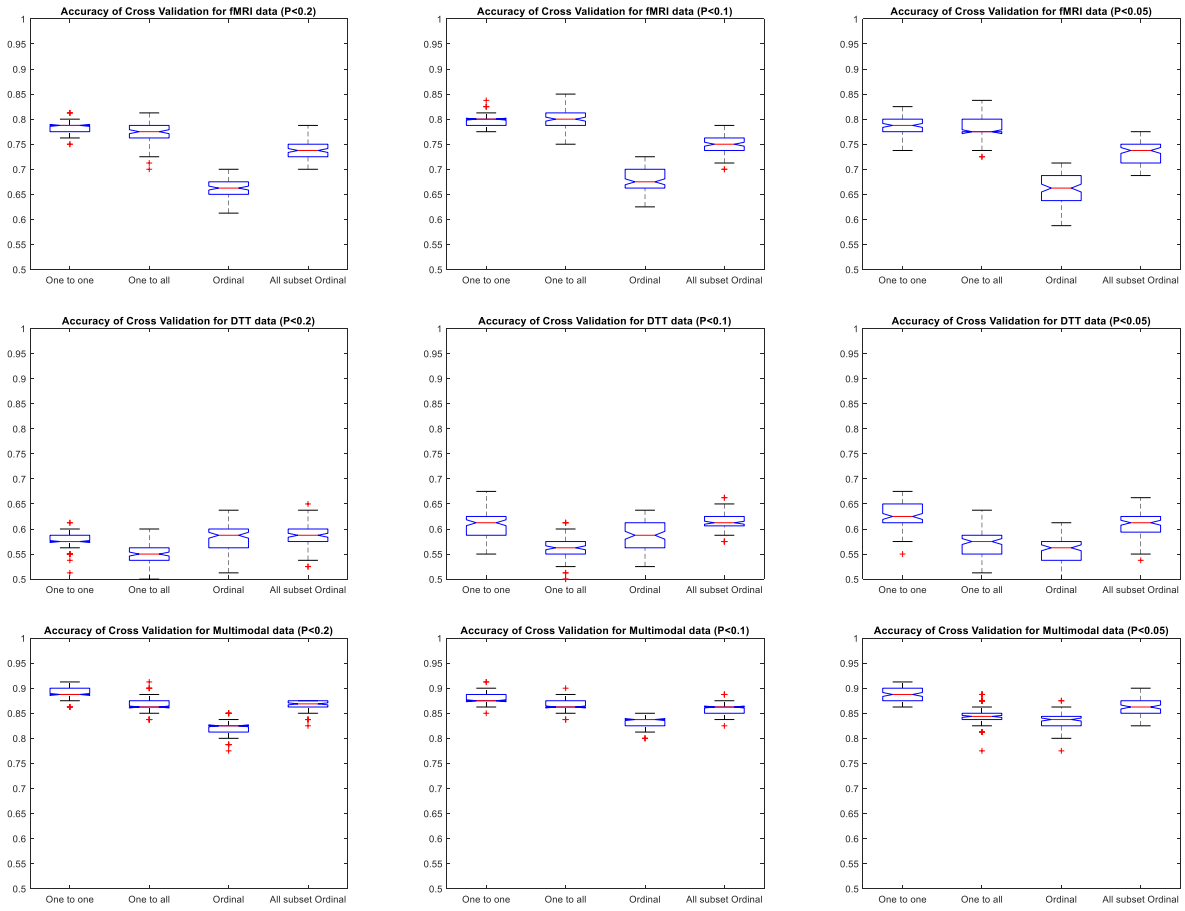

Figure 3 Prediction accuracy of Cross Validation for fMRI data, DTT data and multimodal data are presented by rows after screening. In each plot, results of four multiclass SVM {One to one, One to all, Ordinal, All subset ordinal} are presented in order.

Table 3 Mean prediction accuracy of Multiclass SVM Cross-Validation in 100 times

fMRI	One to one	One to all	Ordinal	All-subset Ordinal
P<0.2	0.7841	0.7714	0.6670	0.7412
P<0.1	0.8006	0.8029	0.6792	0.7494
P<0.05	0.7824	0.7846	0.6616	0.7309

DTT	One to one	One to all	Ordinal	All-subset Ordinal
P<0.2	0.5789	0.5515	0.5876	0.5856
P<0.1	0.6151	0.5575	0.5886	0.6228
P<0.05	0.6290	0.5676	0..5673	0.6141

Multimodal	One to one	One to all	Ordinal	All-subset Ordinal
P<0.2	0.8915	0.8681	0.8209	0.8661
P<0.1	0.8798	0.8644	0.8316	0.8620
P<0.05	0.8901	0.8480	0.8366	0.8631

3.2.3. *Significant Features selection*

In each SVM model, a 100-feature subset was generated by SVM-RFE and these 100 features were considered the most significant features selected by these models (see graph of feature lists in Appendix *Figure 1*). With the hypothesis that there is an ordinal trend in MDD staging, I am very interested in ordinal data among groups. I compared the mean value of selected features among groups and Table 4 shows the numbers of ordinal features in top 50 significant features. The feature selected in four models were very different, it's more likely to select ordinal features in ordinal model and all-subset model than one to one model. These results should be double-checked by its clinical significance.

In this study, I compared results from fMRI data, DTI data and their combination. Results indicated the classification accuracy of fMRI data is better than that of DTI data. Among top 200 significant features selected in multimodal data showed in Table 5, most features are from fMRI, which means fMRI data are more sensitive to MDD staging model.

Table 4 Number of ordinal features in top 50 significant features selected by four models

	One to one	One to all	Ordinal	All-subset
fMRI	1	2	21	17
DTI	13	7	24	17
Multimodal	3	2	21	15

Table 5 Number of features from fMRI data and DTI data in top 200 significant features selected by four models in multimodal dataset

	One to one	One to all	Ordinal	All-subset
fMRI	171	191	144	142
DTI	29	9	56	58
Multimodal	200	200	200	200

4. DISCUSSION

Although many clinical staging models of major depression have been proposed in recent papers, there were few studies focusing on the relationship between neurobiological markers and these classification models. However, development of biomarkers that can be used to identify MDD progression is still an important clinical goal for patients' early diagnose and individualized treatment selection. The purpose of this report was to define neurobiological biomarkers of major depressive disorder clinical staging using multimodal neuroimaging dataset. It's a new perspective to utilize machine learning analyses on multimodal data to identify brain networks contributed most to stage classifications. Multiple multiclass SVM classifiers were applied to neuroimaging data in the classification of MDD, including two traditional methods one-to-one model and one-to-all model as well as two novel algorithms developed in this study ordinal model and all-subset ordinal model. Results showed that classification accuracy of multimodal data was higher than both ordinal model and all-subset ordinal model have higher chance to identify ordinal features and all-subset ordinal model had a quite good cross-validation result compared to one-to-one model. Classification accuracy of multimodal data was much better than results from fMRI and DTI data separately.

Support vector machine is a supervised machine learning algorithm, which means it needs an output value in every observation for classification. In this study, there is a hypothesis that four group of MDD patients (control group, treatment naïve group, treatment responsive recurrent group and treatment resistant group) have a

trend in severity, and I set label value as the outcome of these four groups. further research could treat MDD clinical episodes of patients as new outcome, which is probably more accurate than label value used here. Additionally, if this hypothesis needs to be verified, an unsupervised machine learning algorithm should be applied to datasets, because unsupervised machine learning algorithms, for example clustering and Gaussian mixture models, can draw inferences without outcome.

In this study, four different SVM models were applied to same datasets. Ordinal methods and All-subset ordinal methods were two models built for selecting ordinal features. However, these two models were similar to one-to-one model and one-to-all model. All models divided a multiclass classification problem into several binary classification problems. These models are combinations of local optimization problems rather than a single global optimization problem. Ordinal models and All-subset models which had restrictions that only evaluate ordinal separations are reliable to get a result not good as one-to-one model. Therefore, more multiclass SVM models, especially those use global optimizations, could be utilized in our study to improve classifications results. Although those methods developed in recent research still have problems in computation complexity, it is feasible to extend these models into ordinal form for ordinal feature selection.

REFERENCES

[1] World Health Organization (2017). Depression.

[2] McGorry, P. D., Purcell, R., Hickie, I. B., Yung, A. R., Pantelis, C., & Jackson, H. J. (2007). Clinical staging: a heuristic model for psychiatry and youth mental health. *Med J Aust*, *187*(7 Suppl), S40-S42.

[3] Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1), S199-S209.

[4] Craddock, R. C., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic resonance in Medicine*, *62*(6), 1619-1628.

[5] Zhou, X., & Tuck, D. P. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, *23*(9), 1106-1114.

[6] Mayberg, H. S. (2003). Modulating dysfunctional limbic-cortical circuits in depression: towards development of brain-based algorithms for diagnosis and optimised treatment. *British medical bulletin*, *65*(1), 193-207.

[7] McGrath, C. L., Kelley, M. E., Holtzheimer, P. E., Dunlop, B. W., Craighead, W. E., Franco, A. R., ... & Mayberg, H. S. (2013). Toward a neuroimaging treatment selection biomarker for major depressive disorder. *JAMA psychiatry*, *70*(8), 821-829.

[8] McGorry, P. D., Hickie, I. B., Yung, A. R., Pantelis, C., & Jackson, H. J. (2006). Clinical staging of psychiatric disorders: a heuristic framework for choosing earlier, safer and more effective interventions. *Australian and New Zealand Journal of Psychiatry*, *40*(8), 616-622.

[9] Fava, G. A., & Kellner, R. (1993). Staging: a neglected dimension in psychiatric classification. *Acta Psychiatrica Scandinavica*, *87*(4), 225-230.

[10] Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*, 85-86.

[11] Andrew, A. M. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods by Nello Christianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000, xiii+ 189 pp., ISBN 0-521-78019-5 (Hbk,£ 27.50).

[12] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422.

[13] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

[14] HERNÁNDEZ, J. Partial Discharge feature selection and evaluation using an enhanced recursive feature elimination (RFE) algorithm.

[15] Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415-425.

[16] Cerasa, A., Castiglioni, I., Salvatore, C., Funaro, A., Martino, I., Alfano, S., ... & Quattrone, A. (2015). Biomarkers of Eating Disorders Using Support Vector Machine Analysis of Structural Neuroimaging Data: Preliminary Results. *Behavioural neurology*, *2015*.

[17] Haller, S., Lovblad, K. O., Giannakopoulos, P., & Van De Ville, D. (2014). Multivariate pattern recognition for diagnosis and prognosis in clinical neuroimaging: state of the art, current challenges and future trends. *Brain topography*, *27*(3), 329-337.

[18] Chu, C., Lagercrantz, H., Forssberg, H., & Nagy, Z. (2015). Investigating the Use of Support Vector Machine Classification on Structural Brain Images of Preterm–Born Teenagers as a Biological Marker. *PloS one*, *10*(4), e0123108.

[19] Goodwin, D., Bleymaier, T., & Bhal, S. Identification of Neuroimaging Biomarkers.

[20] Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., & Gotlib, I. H. (2015). Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Frontiers in psychiatry*, *6*, 21.

[21] Sato, J. R., Rondina, J. M., & Mourão-Miranda, J. (2012). Measuring abnormal brains: building normative rules in neuroimaging using one-class support vector machines. *Frontiers in neuroscience*, *6*, 178.

[22] Govada, A., Gauri, B., & Sahay, S. K. (2015, August). Centroid based Binary Tree Structured SVM for multi classification. In *Advances in Computing, Communications and Informatics (ICACCI)*, 2015 International Conference on (pp. 258-262). IEEE.

APPENDICES

Figure 1.1.1 – 1.12.1 are graphs I generated based on top 100 significant features. First 4 graphs are results of fMRI dataset with One-vs-one method, One-vs-all method, Ordinal method and All-subset Ordinal method. Next 4 graphs are results of DTT dataset with One-vs-one method, One-vs-all method, Ordinal method and All-subset Ordinal method and Last 4 graphs results of Multimodal dataset with One-vs-one method, One-vs-all method, Ordinal method and All-subset Ordinal method.

Table 1.1 – 1.12 are regions with most frequency in top 100 significant features.

Figure 1.1.2 – 1.12.2 are plot of mean and standard error of top 50 significant features in four groups

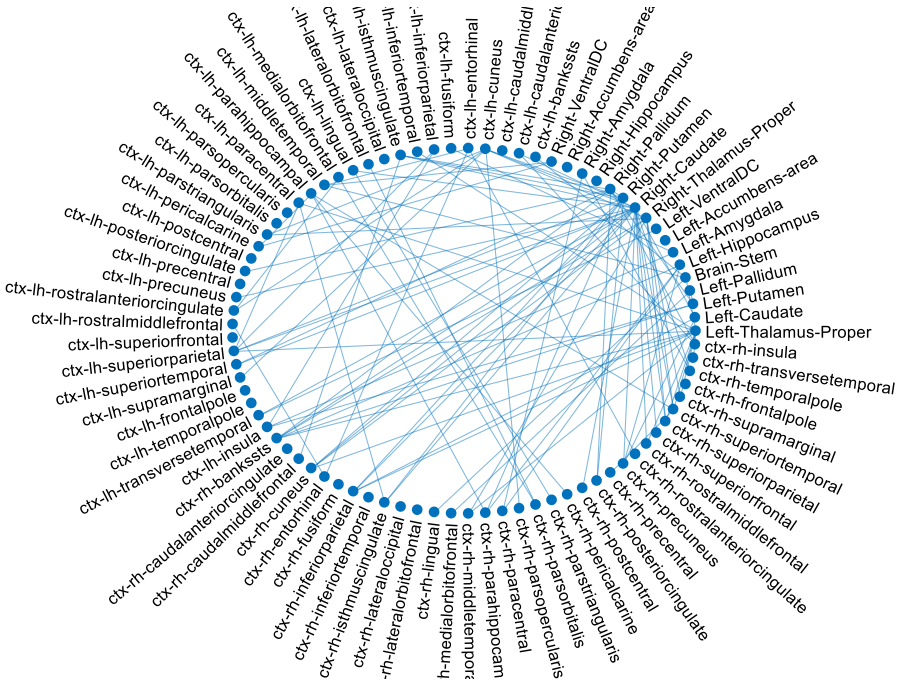

Figure 1.1.1 Graph of top 100 significant features in fMRI data selected by one to one SVM model

Table 1.1 Regions with most frequency in top 100 significant features in fMRI data

fMRI one to one	
Region	Frequency in Top 100 edges
'Right-Caudate'	20
'Left-Thalamus-Proper'	13
'Right-Putamen'	9
'Right-Thalamus-Proper'	8
'Right-Pallidum'	8
'ctx-lh-cuneus'	7
'ctx-rh-bankssts'	6
'ctx-lh-superiorparietal'	5
'ctx-rh-superiortemporal'	5
'Left-Putamen'	4
'Brain-Stem'	4
'Left-Hippocampus'	4
'ctx-lh-isthmuscingulate'	4
'ctx-lh-parahippocampal'	4
'ctx-rh-cuneus'	4
'ctx-rh-inferiorparietal'	4
'Left-Caudate'	3
'ctx-lh-entorhinal'	3
'ctx-lh-middletemporal'	3
'ctx-lh-paracentral'	3

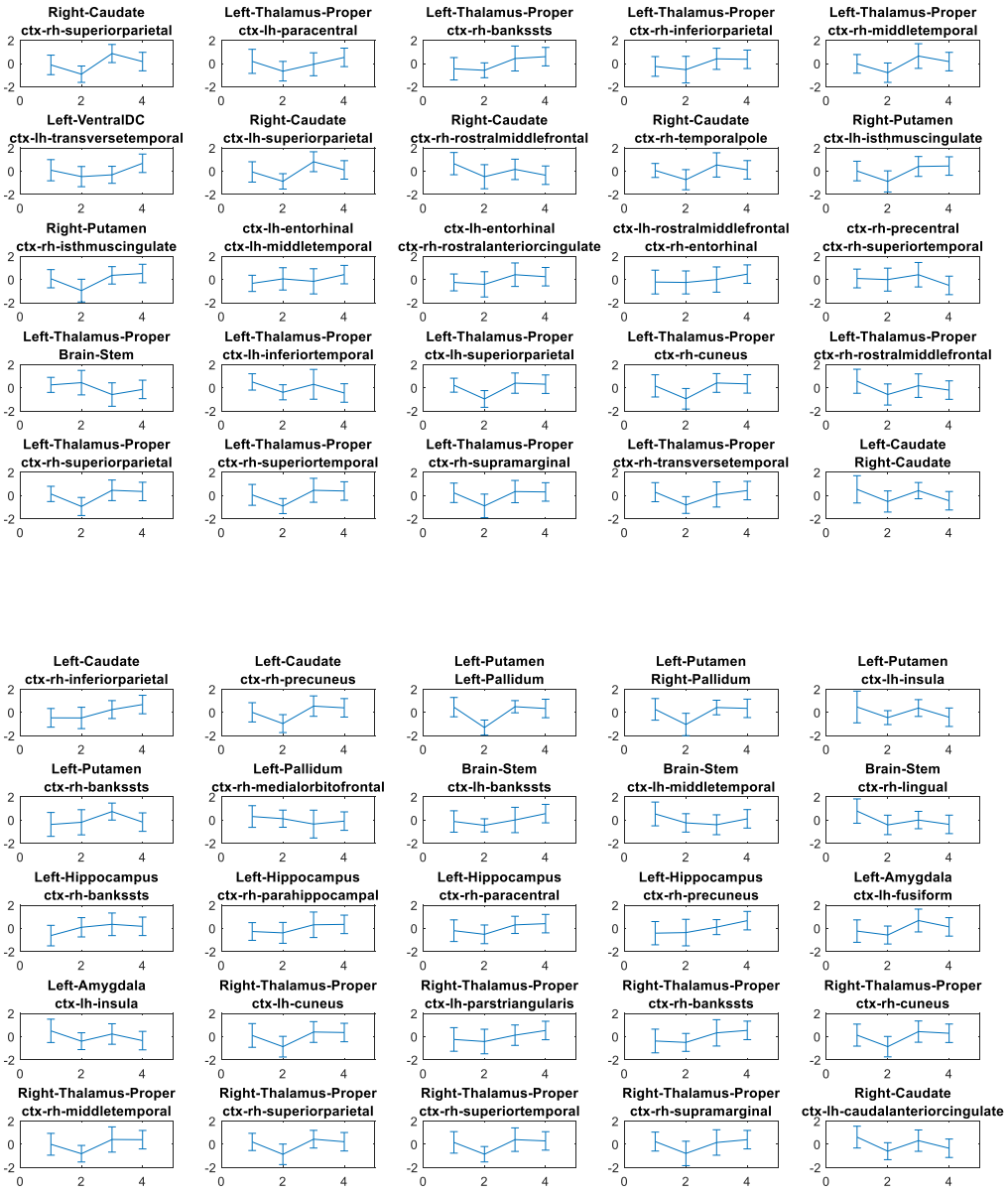

Figure 1.1.2 Plot of top 50 significant features in fMRI data selected by one to one method

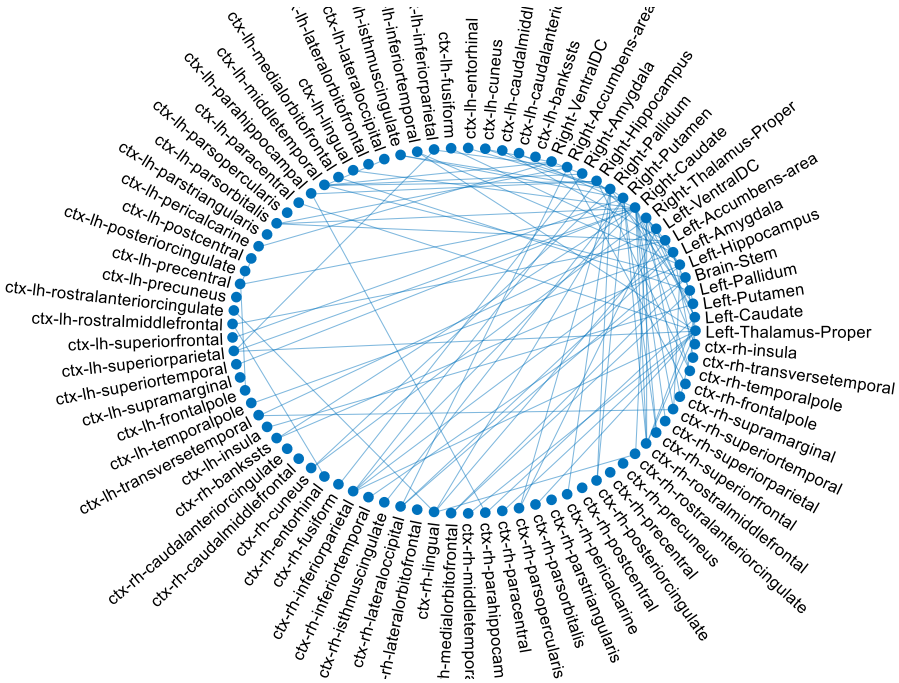

Figure 1.2.1 Graph of top 100 significant features in fMRI data selected by one to all SVM model

Table 1.2 Regions with most frequency in top 100 significant features in fMRI data

fMRI one to all	
Region	Frequency in Top 100 edges
'Right-Caudate'	15
'Left-Thalamus-Proper'	11
'Right-Pallidum'	9
'Right-Thalamus-Proper'	8
'Right-Putamen'	8
'Left-Accumbens-area'	7
'Left-Hippocampus'	6
'Left-Amygdala'	6
'Right-Hippocampus'	5
'ctx-lh-middletemporal'	5
'ctx-rh-rostralmiddlefrontal'	5
'Left-Caudate'	4
'Left-Pallidum'	4
'Brain-Stem'	4
'Left-VentralDC'	4
'Right-Accumbens-area'	4
'Right-VentralDC'	4
'ctx-rh-inferiorparietal'	4
'ctx-rh-lingual'	4
'Left-Putamen'	3

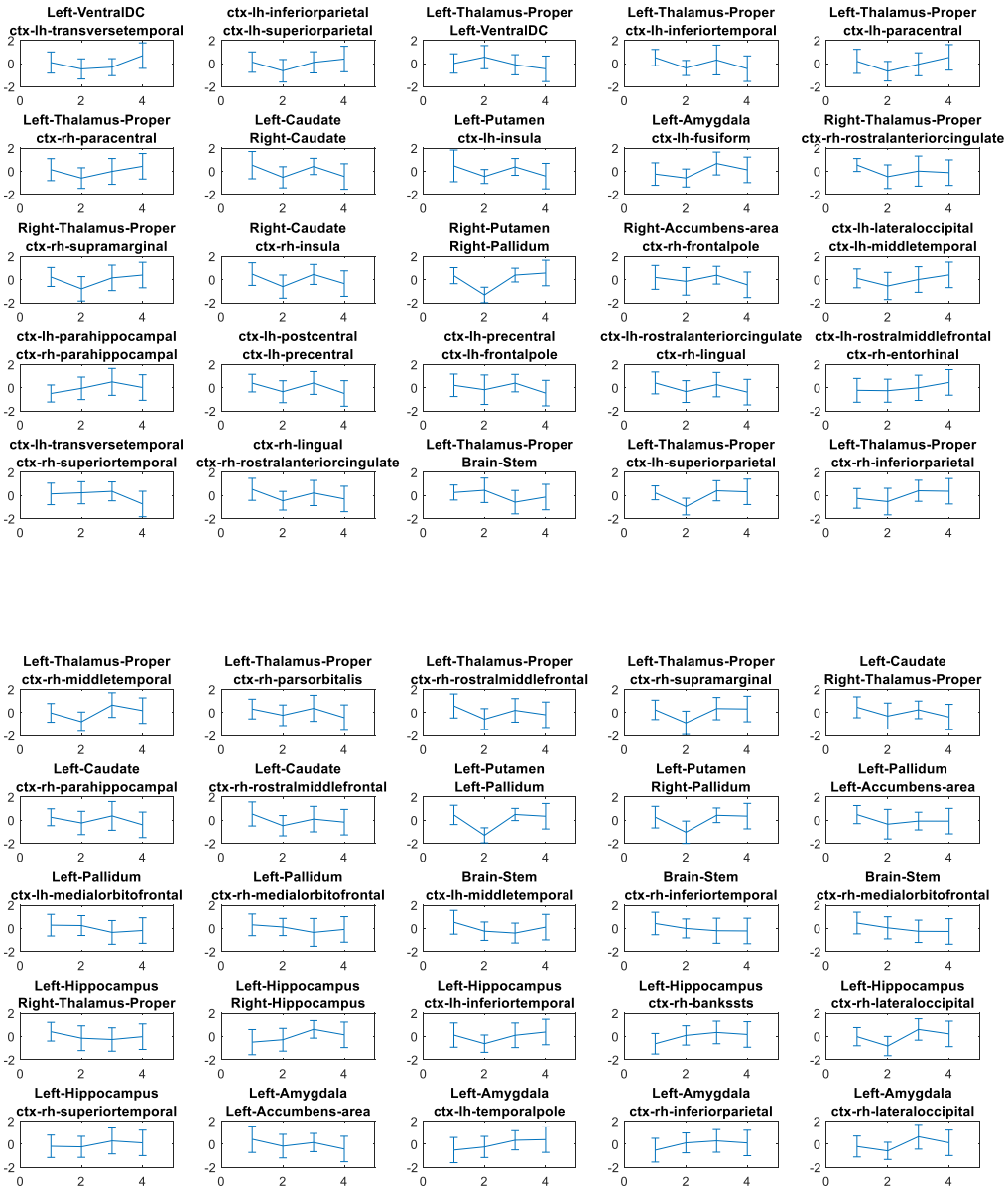

Figure 1.2.2 Plot of top 50 significant features in fMRI data selected by one to all method

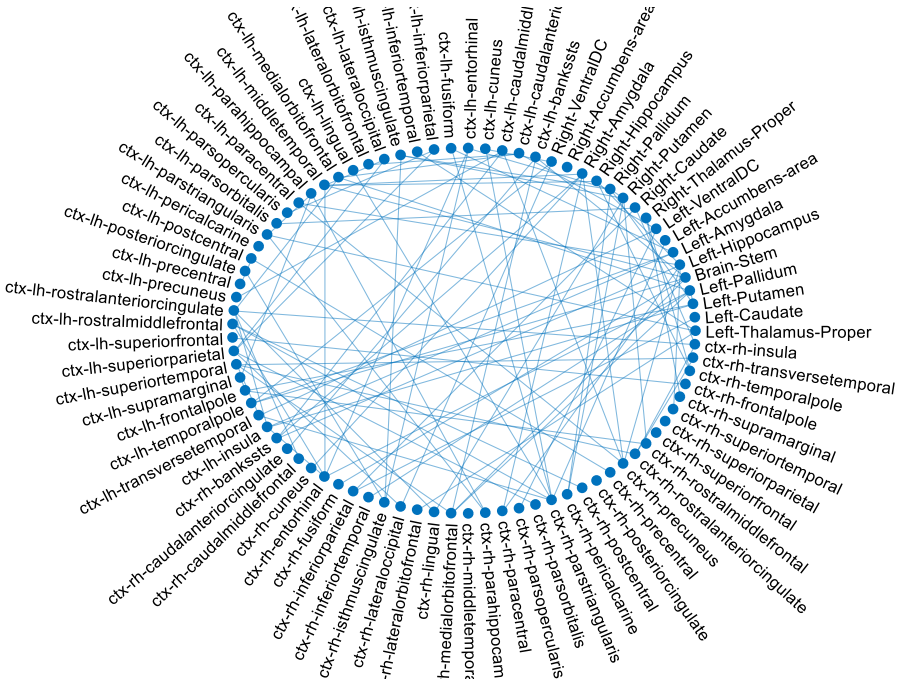

Figure 1.3.1 Graph of top 100 significant features in fMRI data selected by ordinal SVM model

Table 1.3 Regions with most frequency in top 100 significant features in fMRI data

fMRI ordinal SVM model	
Region	Frequency in Top 100 edges
'Left-Hippocampus'	6
'ctx-rh-parstriangularis'	6
'Left-Pallidum'	5
'Brain-Stem'	5
'Right-Putamen'	5
'Right-Pallidum'	5
'Right-Amygdala'	5
'ctx-lh-middletemporal'	5
'ctx-lh-rostralanteriorcingulate'	5
'ctx-rh-entorhinal'	5
'Left-Putamen'	4
'ctx-lh-superiorparietal'	4
'ctx-lh-superiortemporal'	4
'ctx-lh-insula'	4
'ctx-rh-medialorbitofrontal'	4
'ctx-rh-precuneus'	4
'Left-Thalamus-Proper'	3
'Left-Amygdala'	3
'Left-VentralDC'	3
'Right-Caudate'	3

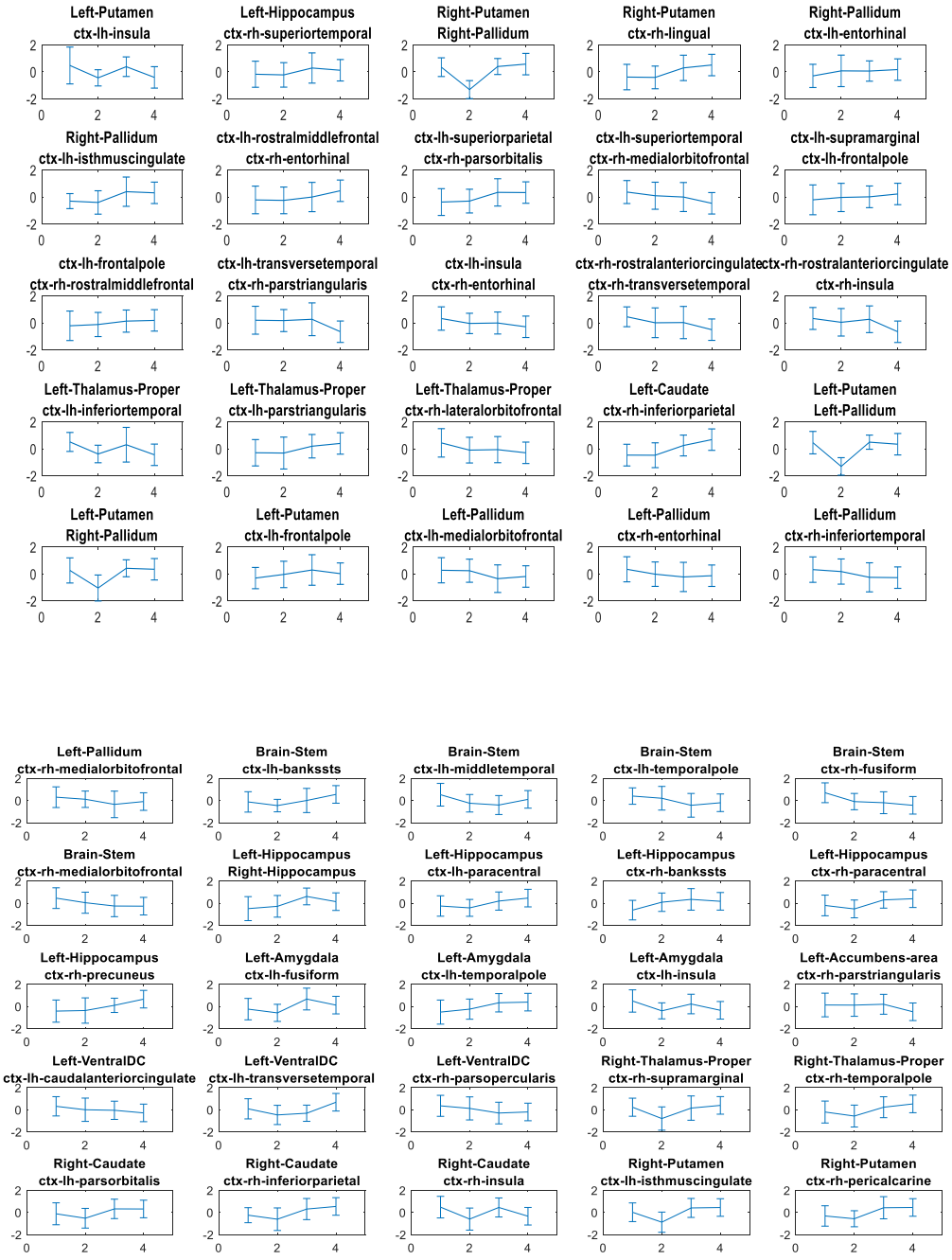

Figure 1.3.2 Plot of top 50 significant features in fMRI data selected by ordinal method

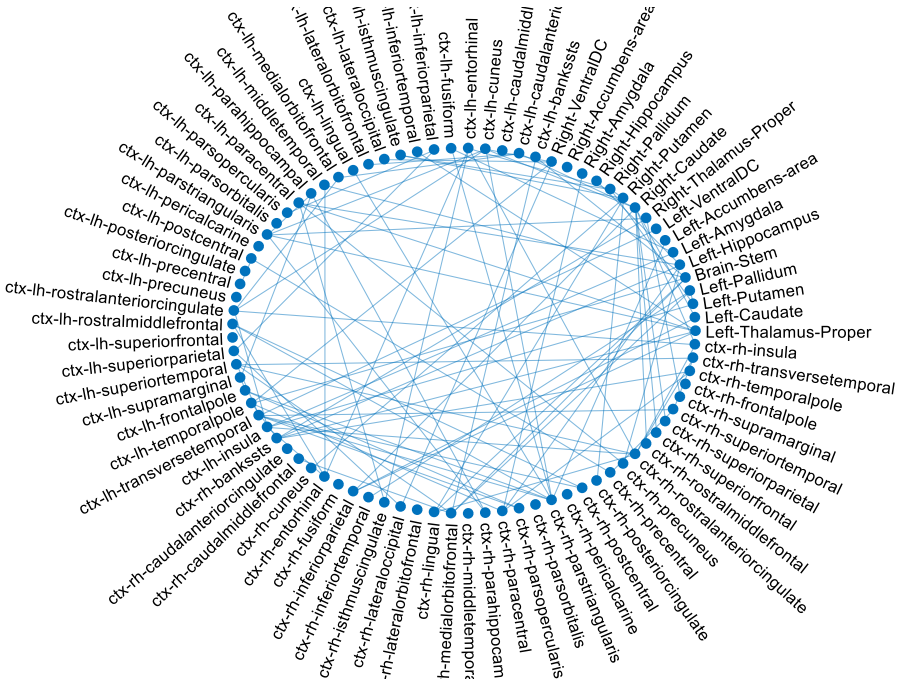

Figure 1.4.1 Graph of top 100 significant features in fMRI data selected by all-subset ordinal model

Table 1.4 Regions with most frequency in top 100 significant features in fMRI data

fMRI all-subset ordinal	
Region	Frequency in Top 100 edges
'Right-Caudate'	8
'ctx-rh-parstriangularis'	7
'ctx-lh-transversetemporal'	6
'ctx-rh-rostralanteriorcingulate'	6
'Left-Thalamus-Proper'	5
'Brain-Stem'	5
'Left-Hippocampus'	5
'Right-Putamen'	5
'ctx-lh-cuneus'	5
'ctx-lh-parstriangularis'	5
'ctx-lh-insula'	5
'ctx-rh-bankssts'	5
'ctx-rh-medialorbitofrontal'	5
'Right-Pallidum'	4
'Right-VentralDC'	4
'ctx-lh-entorhinal'	4
'ctx-lh-paracentral'	4
'ctx-lh-rostralanteriorcingulate'	4
'ctx-lh-superiortemporal'	4
'ctx-lh-temporalpole'	4

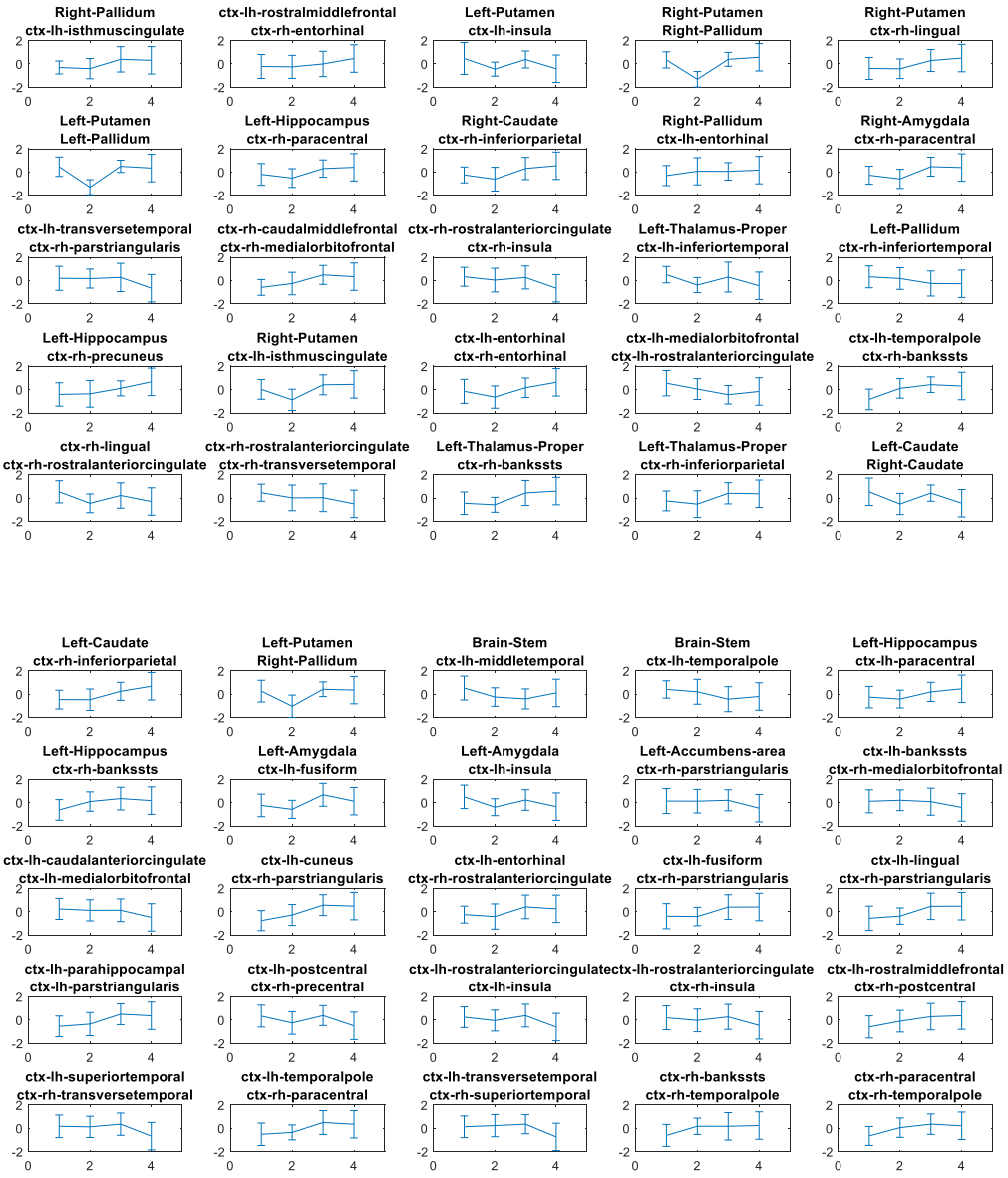

Figure 1.4.2 Plot of top 50 significant features in fMRI data selected by all-subset method

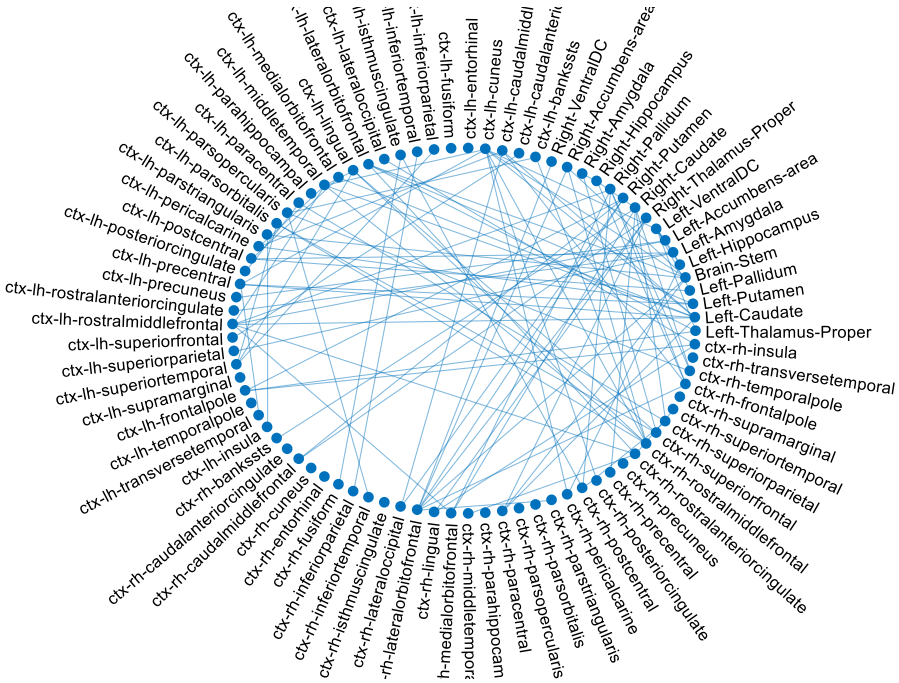

Figure 1.5.1 Graph of top 100 significant features in DTI data selected by one to one model

Table 1.5 Regions with most frequency in top 100 significant features in DTI data

DTT one to one	
Region	Frequency in Top 100 edges
'ctx-lh-cuneus'	8
'Left-Caudate'	7
'ctx-rh-lateralorbitofrontal'	7
'Brain-Stem'	6
'Left-VentralDC'	6
'Right-Caudate'	6
'ctx-lh-middletemporal'	6
'ctx-lh-rostralmiddlefrontal'	6
'ctx-rh-rostralmiddlefrontal'	6
'Left-Amygdala'	5
'Right-Pallidum'	5
'Right-Thalamus-Proper'	4
'Right-Putamen'	4
'ctx-lh-caudalmiddlefrontal'	4
'ctx-lh-lateralorbitofrontal'	4
'ctx-lh-lingual'	4
'ctx-lh-precentral'	4
'ctx-rh-rostralanteriorcingulate'	4
'ctx-rh-frontalpole'	4
'Left-Thalamus-Proper'	3

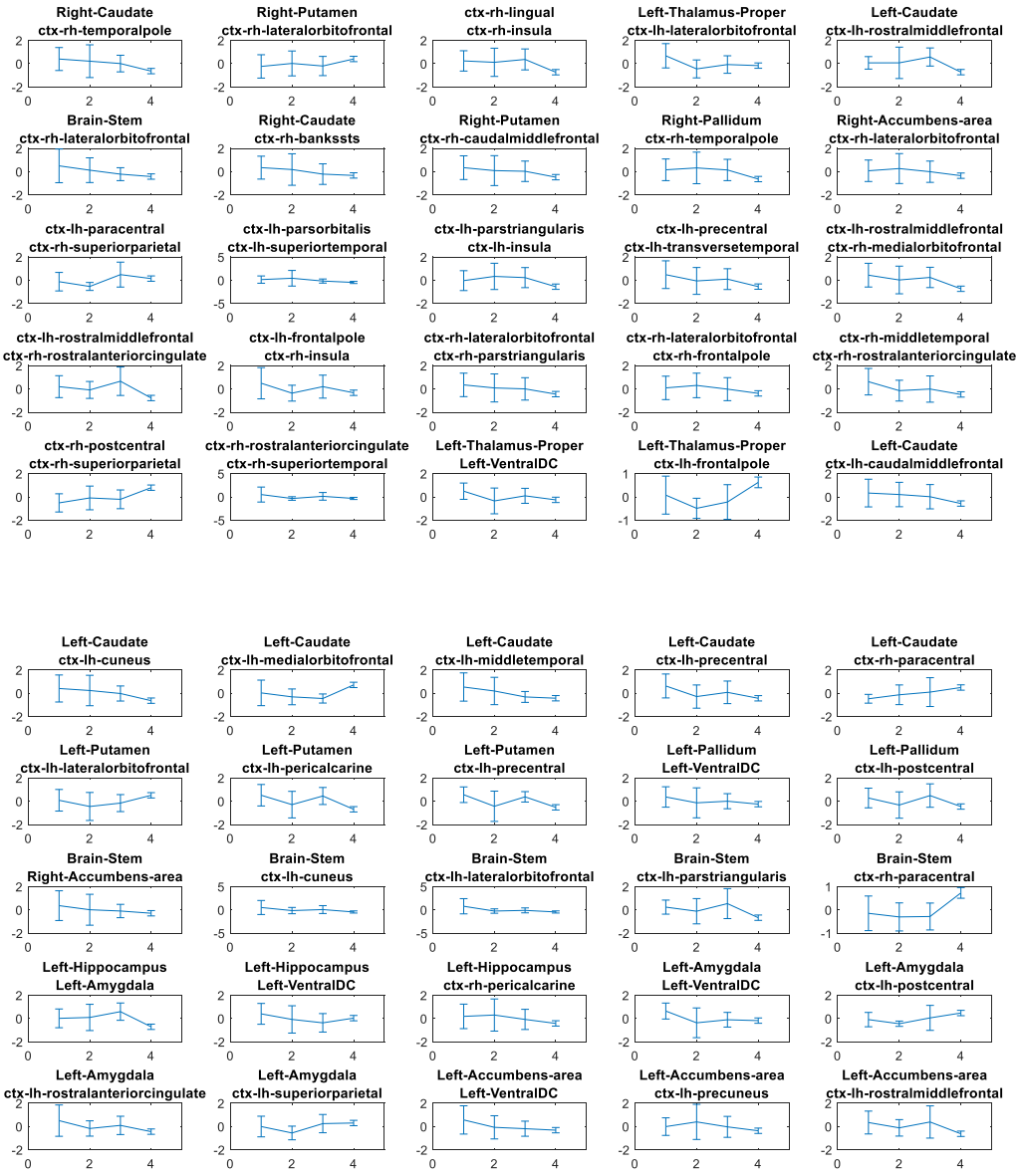

Figure 1.5.2 Plot of top 50 significant features in DTI data selected by one to one method

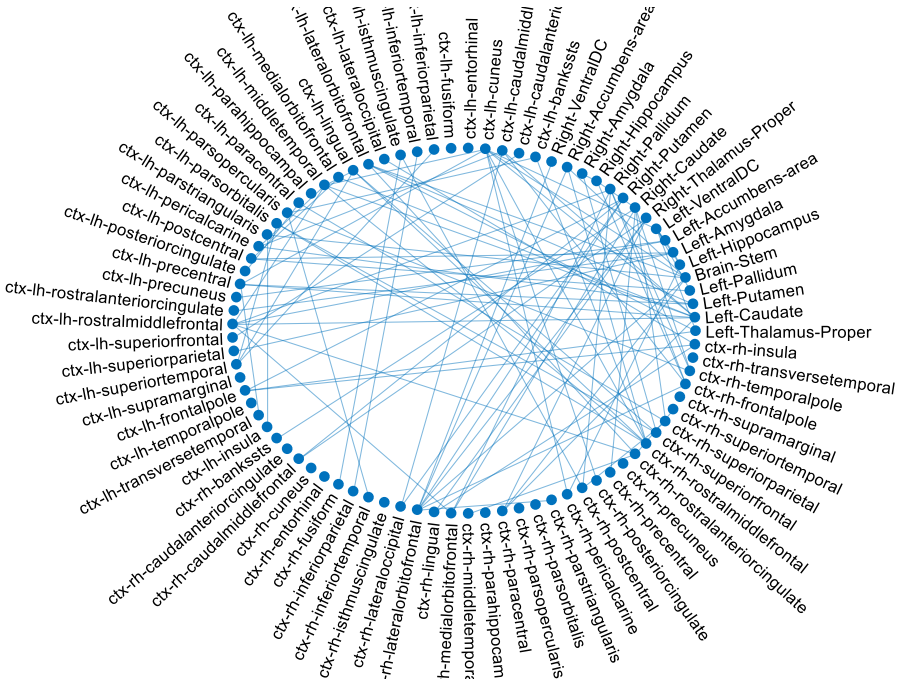

Figure 1.6.1 Graph of top 100 significant features in DTI data selected by one to all model

Table 1.6 Regions with most frequency in top 100 significant features in DTI data

DTT one to all	
Region	Frequency in Top 100 edges
'Left-Putamen'	8
'Left-Amygdala'	8
'Right-Caudate'	7
'Left-Thalamus-Proper'	6
'Left-Caudate'	6
'Brain-Stem'	6
'Left-VentralDC'	5
'Right-Pallidum'	5
'Right-VentralDC'	5
'ctx-lh-lateraloccipital'	5
'ctx-lh-lateralorbitofrontal'	5
'ctx-lh-parstriangularis'	5
'ctx-rh-lateralorbitofrontal'	5
'Left-Pallidum'	4
'Left-Hippocampus'	4
'Right-Putamen'	4
'Right-Hippocampus'	4
'Right-Amygdala'	4
'ctx-lh-cuneus'	4
'ctx-lh-rostralmiddlefrontal'	4

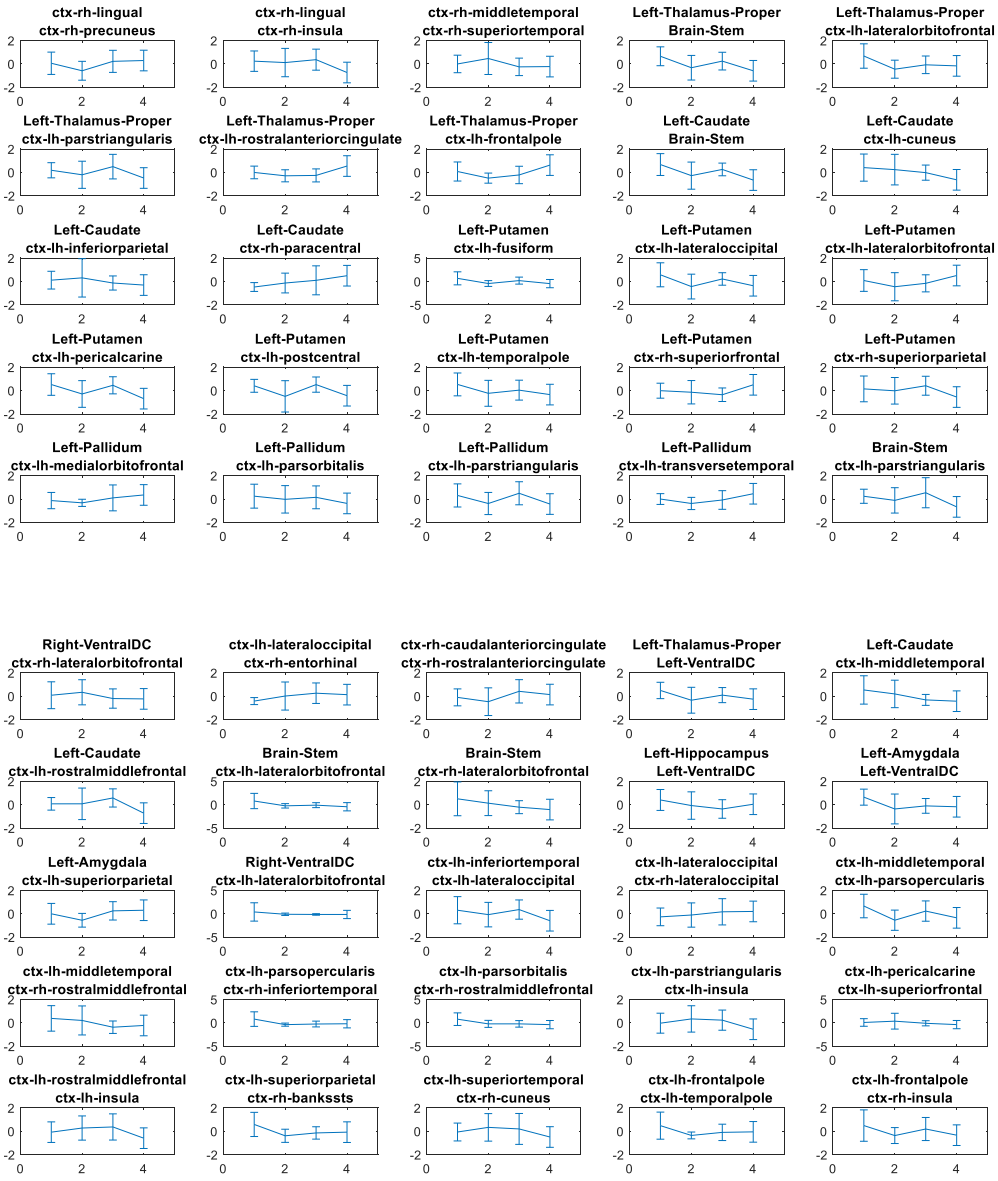

Figure 1.6.2 Plot of top 50 significant features in DTI data selected by one to all method

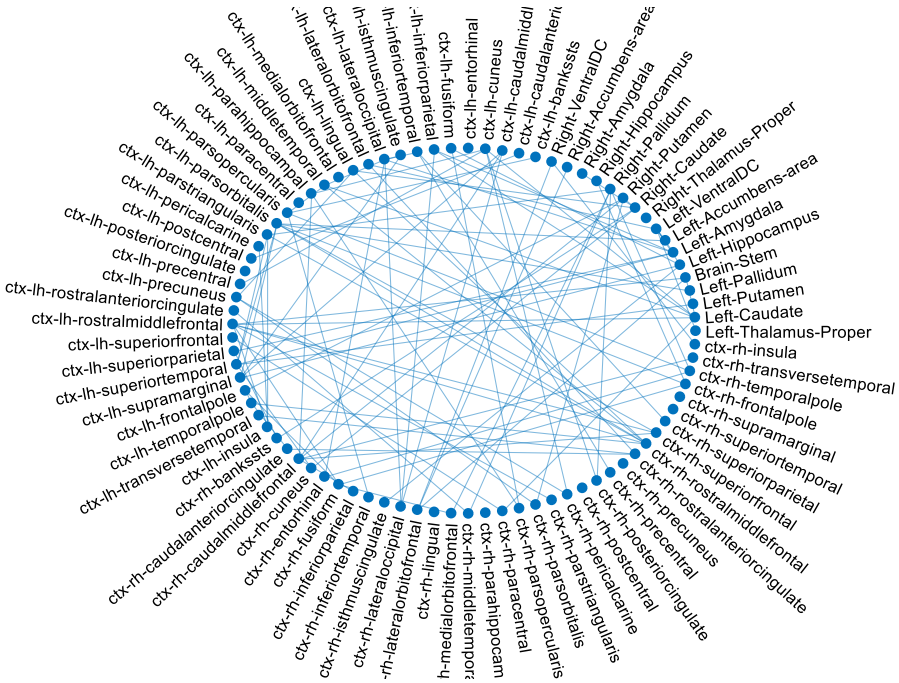

Figure 1.7.1 Graph of top 100 significant features in DTI data selected by ordinal model

Table 1.7 Regions with most frequency in top 100 significant features in DTI data

DTT ordinal	
Region	Frequency in Top 100 edges
'ctx-lh-parsorbitalis'	7
'ctx-lh-rostralmiddlefrontal'	7
'Left-Caudate'	6
'ctx-lh-lateraloccipital'	6
'ctx-rh-caudalmiddlefrontal'	6
'ctx-rh-rostralmiddlefrontal'	6
'Left-Amygdala'	5
'Left-VentralDC'	5
'Right-Caudate'	5
'Right-Pallidum'	5
'ctx-lh-cuneus'	5
'ctx-lh-superiortemporal'	5
'Left-Hippocampus'	4
'Right-Putamen'	4
'ctx-lh-middletemporal'	4
'ctx-lh-parsopercularis'	4
'ctx-lh-parstriangularis'	4
'ctx-lh-supramarginal'	4
'ctx-rh-fusiform'	4
'Brain-Stem'	3

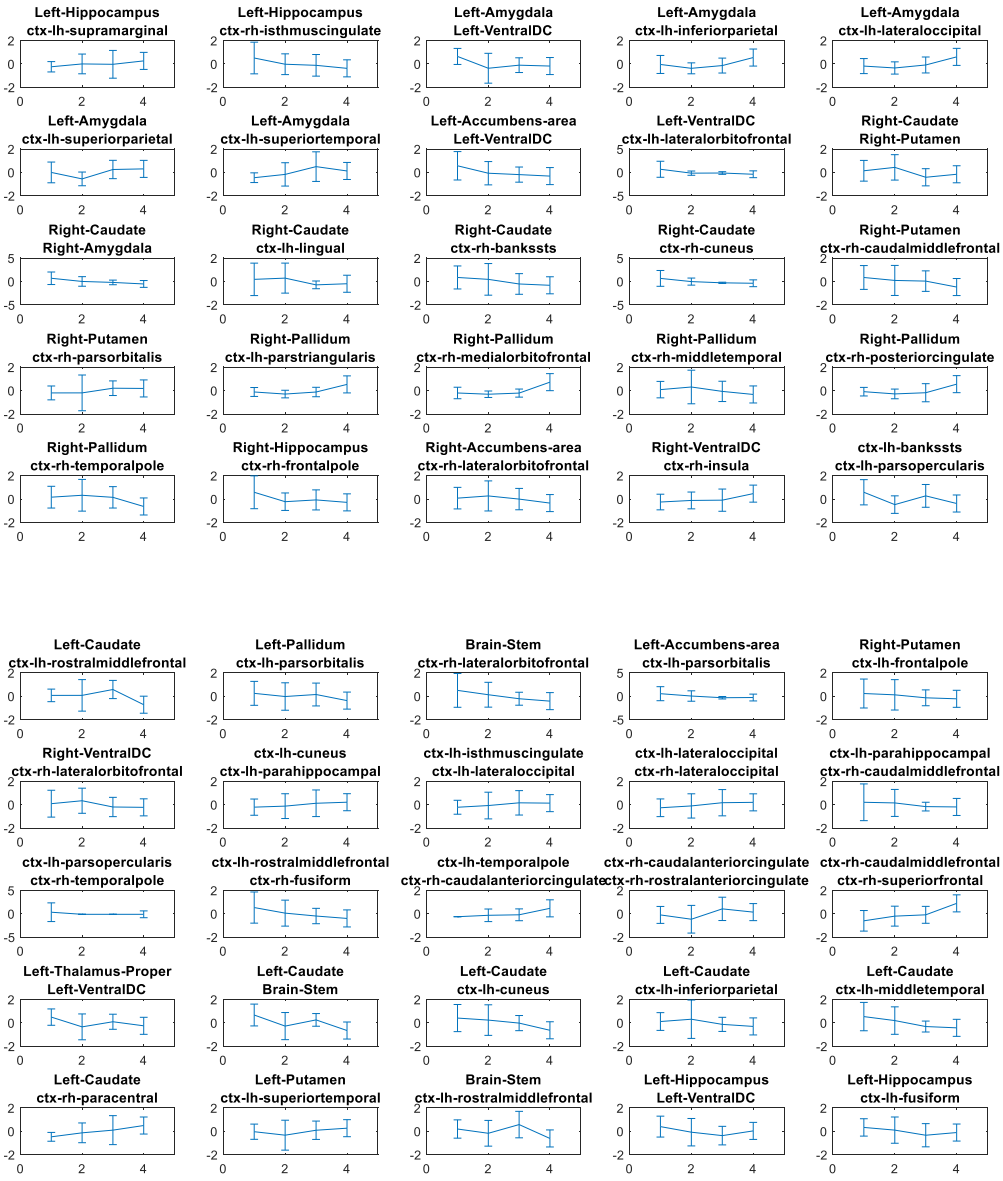

Figure 1.7.2 Plot of top 50 significant features in DTI data selected by ordinal method

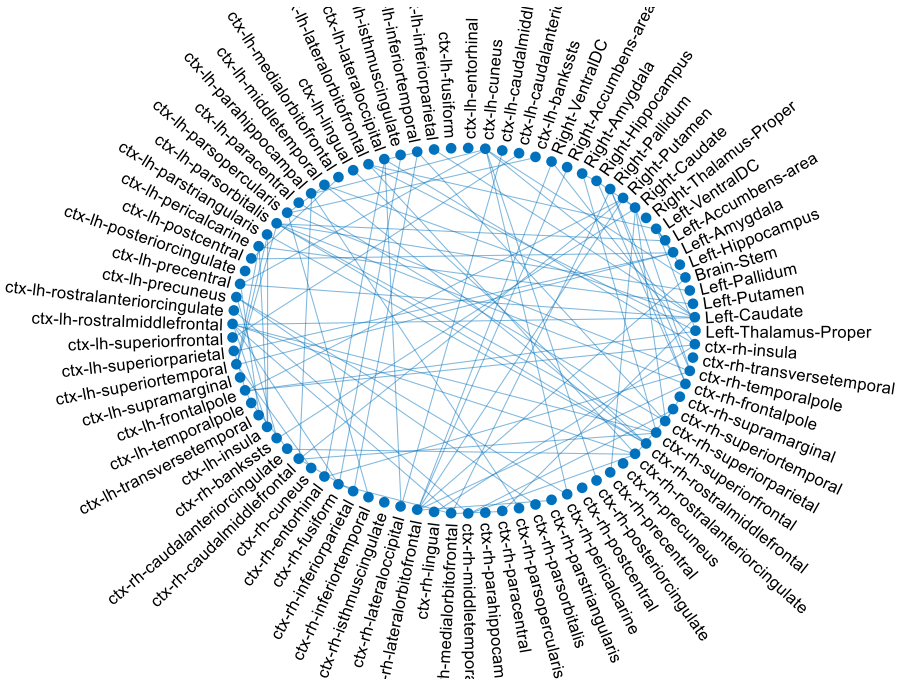

Figure 1.8.1 Graph of top 100 significant features in DTI data selected by all-subset model

Table 1.8 Regions with most frequency in top 100 significant features in DTI data

DTT all-subset ordinal	
Region	Frequency in Top 100 edges
'Left-Caudate'	6
'Right-Caudate'	6
'ctx-lh-cuneus'	6
'ctx-lh-parsorbitalis'	6
'ctx-lh-rostralmiddlefrontal'	6
'ctx-rh-lateralorbitofrontal'	6
'Left-VentralDC'	5
'ctx-lh-superiortemporal'	5
'ctx-lh-frontalpole'	5
'ctx-rh-rostralanteriorcingulate'	5
'Left-Amygdala'	4
'Right-Putamen'	4
'ctx-lh-lateraloccipital'	4
'ctx-lh-parsopercularis'	4
'ctx-rh-caudalmiddlefrontal'	4
'ctx-rh-medialorbitofrontal'	4
'ctx-rh-rostralmiddlefrontal'	4
'ctx-rh-superiorfrontal'	4
'ctx-rh-superiorparietal'	4
'ctx-rh-frontalpole'	4

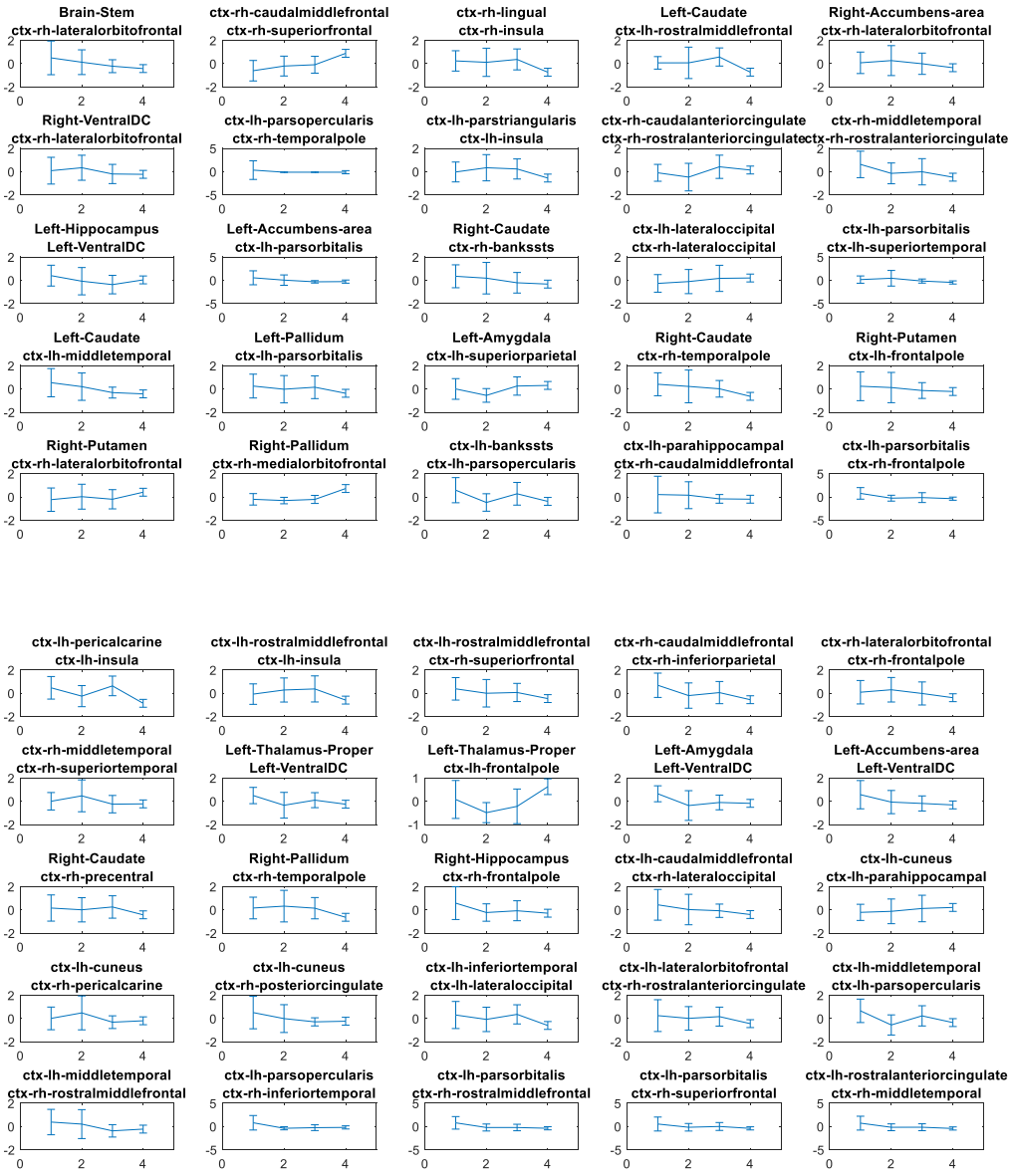

Figure 1.8.2 Plot of top 50 significant features in DTI data selected by all-subset method

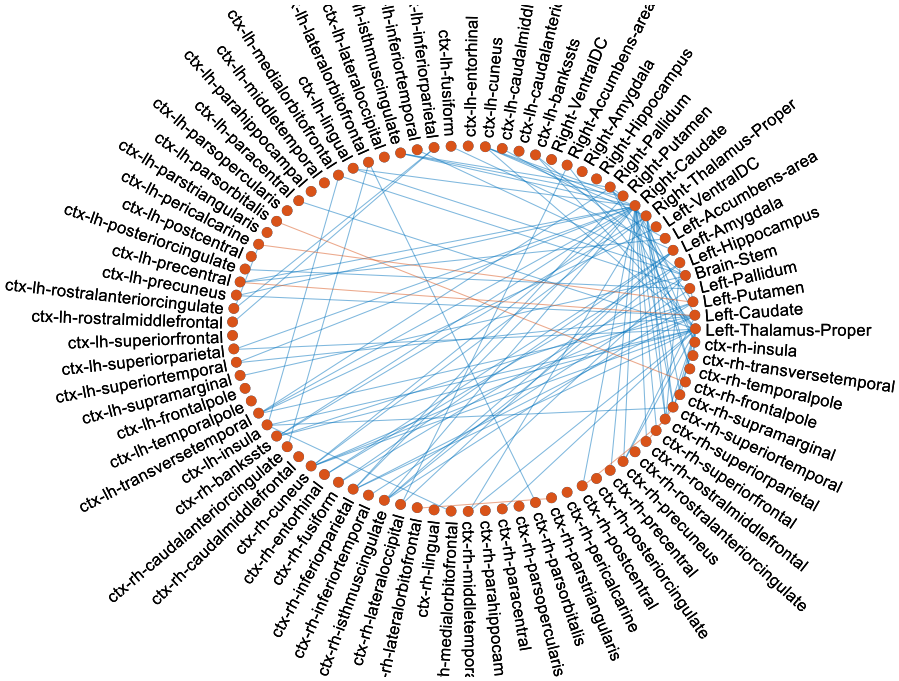

Figure 1.9.1 Graph of top 100 significant features in multimodal data selected by one to one model

Table 1.9 Regions with most frequency in top 100 significant features in multimodal data

Multimodal one to one (fMRI features are in blue, DTT features are in red)	
Region	Frequency in Top 100 edges
'Right-Caudate'	22
'Left-Thalamus-Proper'	16
'Left-Caudate'	8
'Right-Thalamus-Proper'	8
'Left-Putamen'	6
'Brain-Stem'	6
'ctx-rh-superiortemporal'	6
'Left-Amygdala'	5
'Right-Pallidum'	5
'ctx-rh-cuneus'	5
'ctx-rh-superiorparietal'	5
'Left-Pallidum'	4
'Right-Putamen'	4
'ctx-lh-cuneus'	4
'ctx-lh-transversetemporal'	4
'ctx-lh-insula'	4
'ctx-rh-bankssts'	4
'ctx-rh-inferiorparietal'	4
'ctx-rh-isthmuscingulate'	4
'Left-Hippocampus'	3

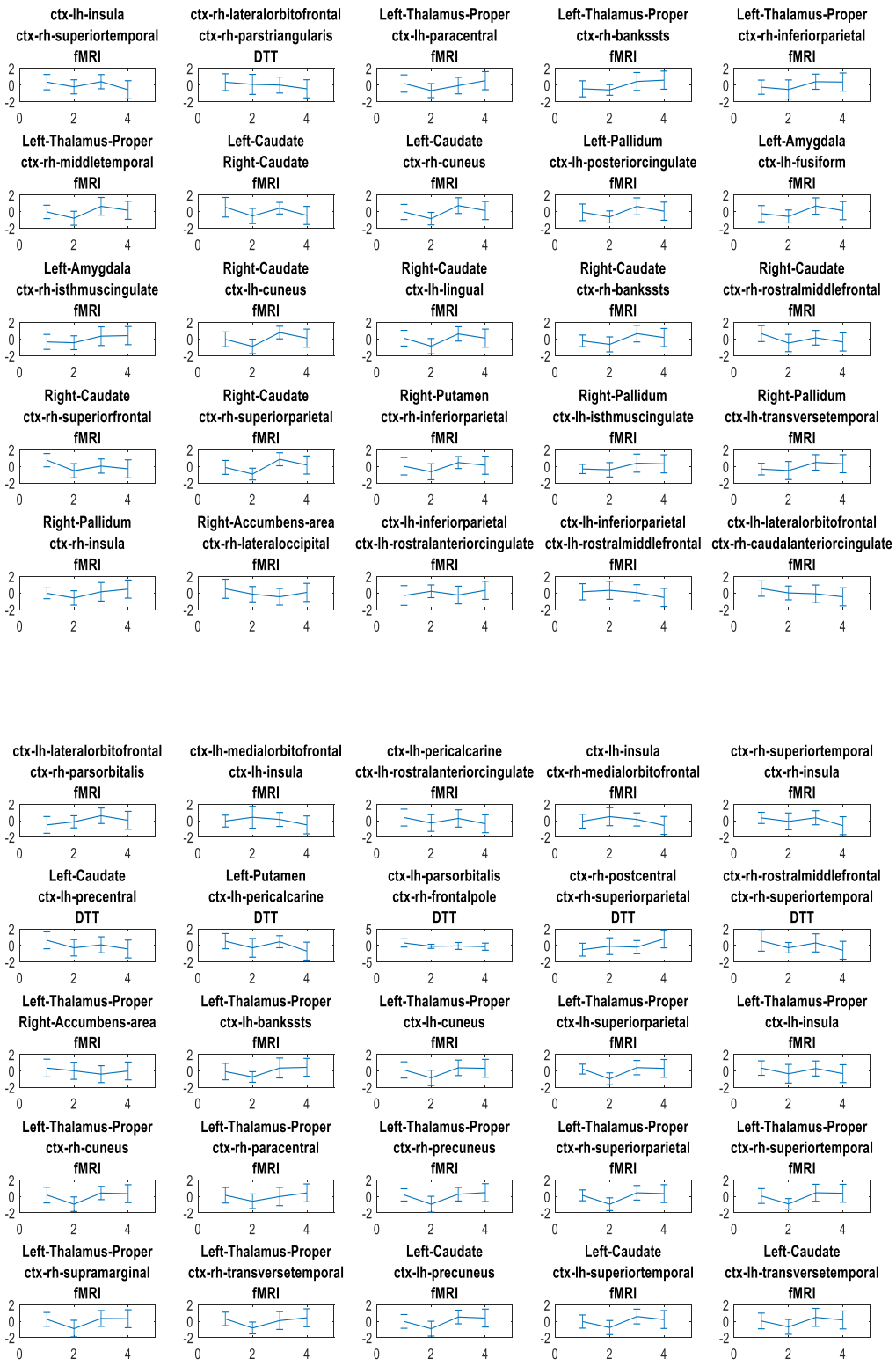

Figure 1.9.2 Plot of top 50 significant features in multimodal data selected by one to one method

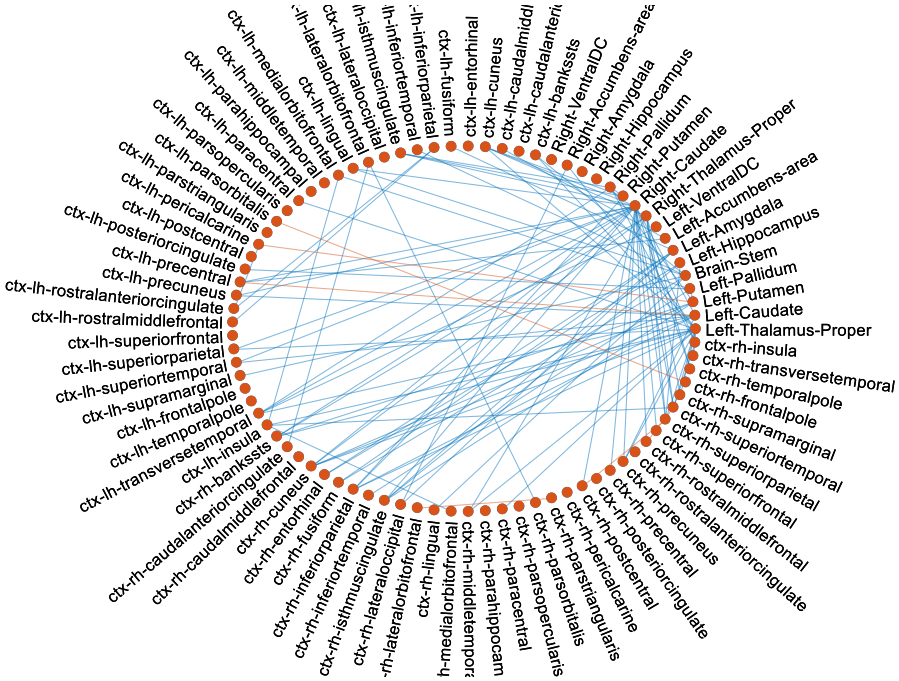

Figure 1.10.1 Graph of top 100 significant features in multimodal data selected by one to all model

Table 1.10 Regions with most frequency in top 100 significant features in multimodal data

Multimodal one to all (fMRI features are in blue, DTT features are in red)	
Region	Frequency in Top 100 edges
'Left-Thalamus-Proper'	19
'Right-Thalamus-Proper'	12
'Left-Caudate'	10
'Brain-Stem'	10
'Left-Amygdala'	7
'Left-Accumbens-area'	7
'Left-Putamen'	6
'Left-Pallidum'	6
'Left-Hippocampus'	5
'ctx-rh-cuneus'	5
'ctx-rh-middletemporal'	5
'Left-VentralDC'	4
'ctx-rh-inferiorparietal'	4
'ctx-rh-lateraloccipital'	4
'Right-Caudate'	3
'Right-VentralDC'	3
'ctx-lh-cuneus'	3
'ctx-lh-lateraloccipital'	3
'ctx-lh-precentral'	3
'ctx-lh-superiorparietal'	3

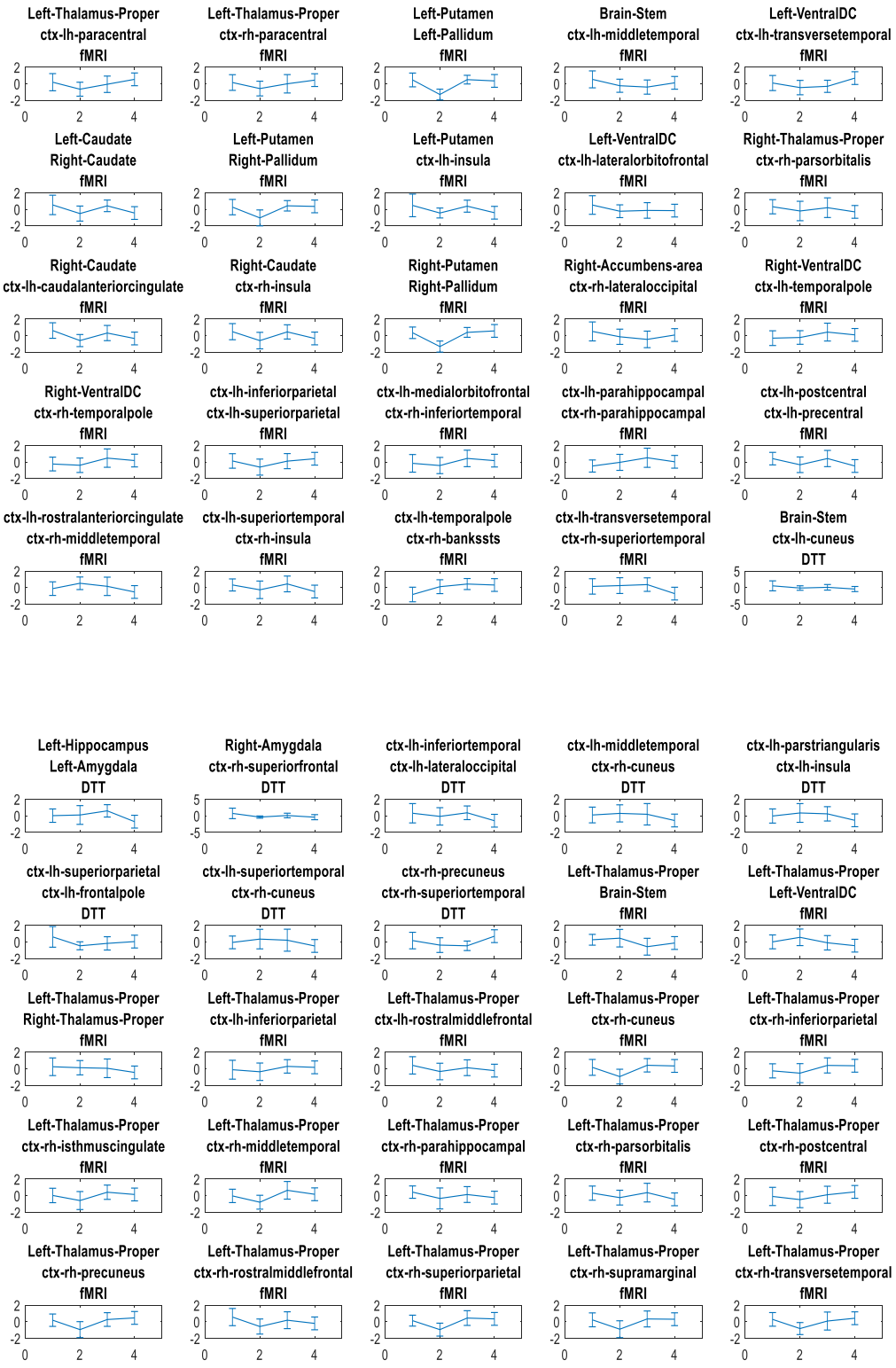

Figure 1.10.2 Plot of top 50 significant features in multimodal data selected by one to all method

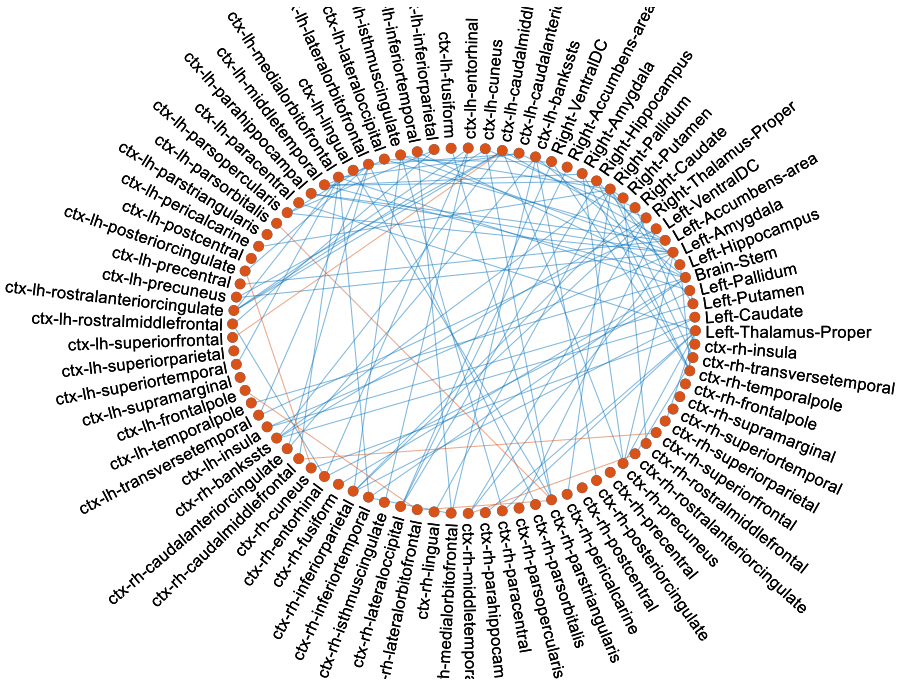

Figure 1.11.1 Graph of top 100 significant features in multimodal data selected by ordinal model

Table 1.11 Regions with most frequency in top 100 significant features in multimodal data

Multimodal ordinal (fMRI features are in blue, DTT features are in red)	
Region	Frequency in Top 100 edges
'Right-Pallidum'	8
'Left-Amygdala'	7
'Left-Thalamus-Proper'	6
'Brain-Stem'	6
'ctx-lh-caudalmiddlefrontal'	6
'ctx-lh-lateralorbitofrontal'	6
'ctx-lh-medialorbitofrontal'	6
'ctx-rh-parstriangularis'	5
'Left-Accumbens-area'	4
'Right-Thalamus-Proper'	4
'Right-Hippocampus'	4
'ctx-lh-bankssts'	4
'ctx-lh-isthmuscingulate'	4
'ctx-lh-middletemporal'	4
'ctx-lh-rostralanteriorcingulate'	4
'ctx-rh-bankssts'	4
'ctx-rh-inferiorparietal'	4
'ctx-rh-precuneus'	4
'ctx-rh-temporalpole'	4
'Left-Putamen'	3

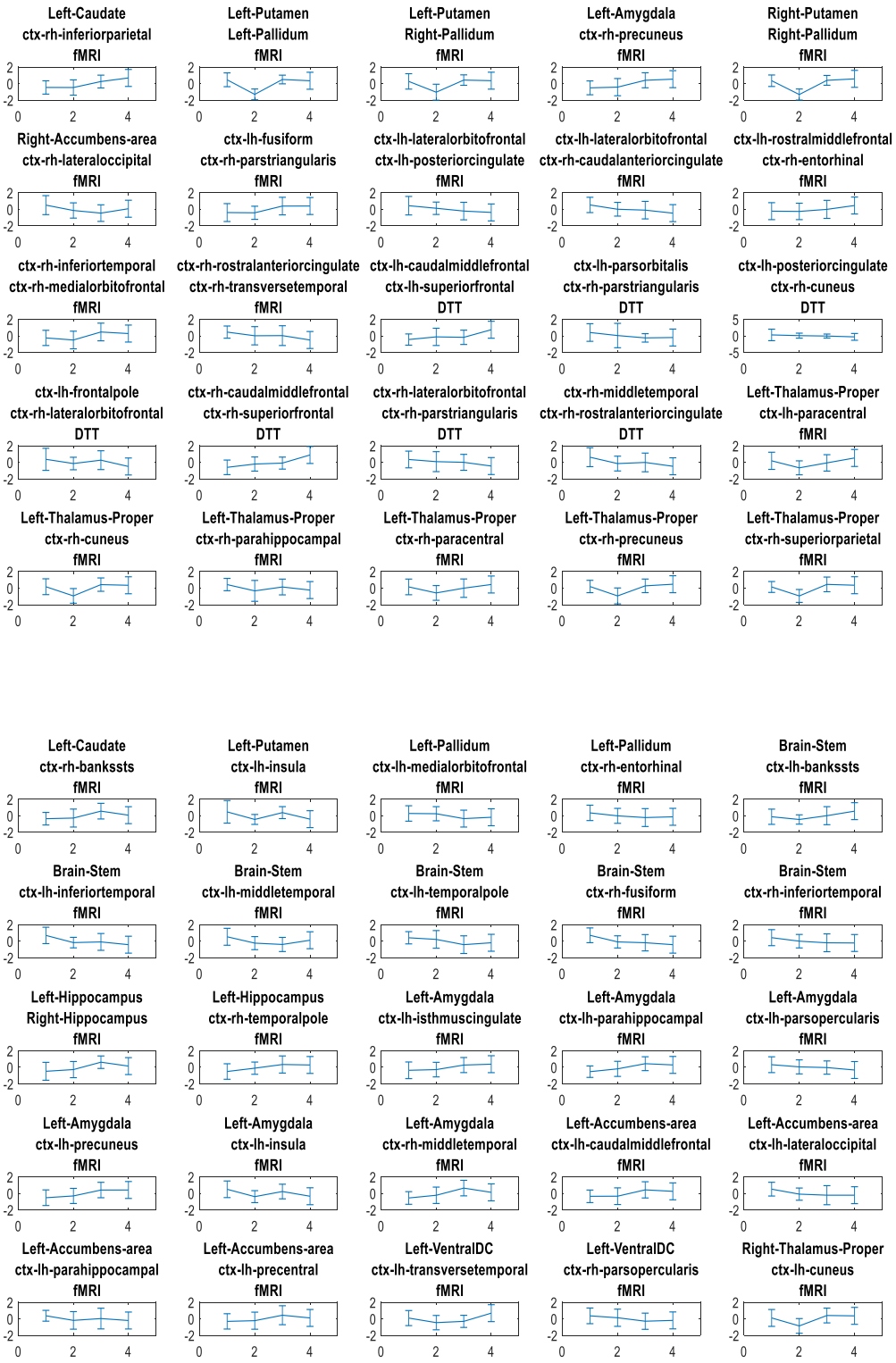

Figure 1.11.2 Plot of top 50 significant features in multimodal data selected by ordinal method

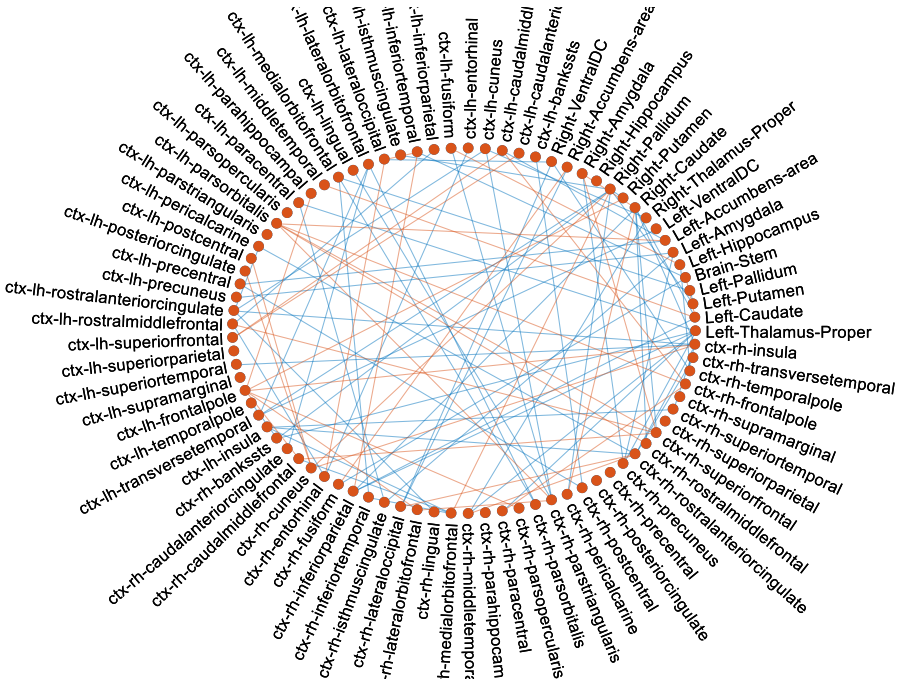

Figure 1.12.1 Graph of top 100 significant features in multimodal data selected by all-subset model

Table 1.12 Regions with most frequency in top 100 significant features in multimodal data

Multimodal all-subset ordinal (fMRI features are in blue, DTT features are in red)	
Region	Frequency in Top 100 edges
'Right-Pallidum'	8
'Right-Caudate'	7
'ctx-rh-insula'	7
'Left-Thalamus-Proper'	6
'ctx-lh-frontalpole'	6
'ctx-rh-rostralanteriorcingulate'	6
'ctx-lh-parsorbitalis'	5
'ctx-rh-parstriangularis'	5
'ctx-rh-superiorfrontal'	5
'Left-Caudate'	4
'ctx-lh-rostralanteriorcingulate'	4
'ctx-rh-cuneus'	4
'ctx-rh-inferiorparietal'	4
'ctx-rh-medialorbitofrontal'	4
'Brain-Stem'	3
'Left-Amygdala'	3
'Left-Accumbens-area'	3
'Right-Amygdala'	3
'Right-VentralDC'	3
'ctx-lh-caudalmiddlefrontal'	3

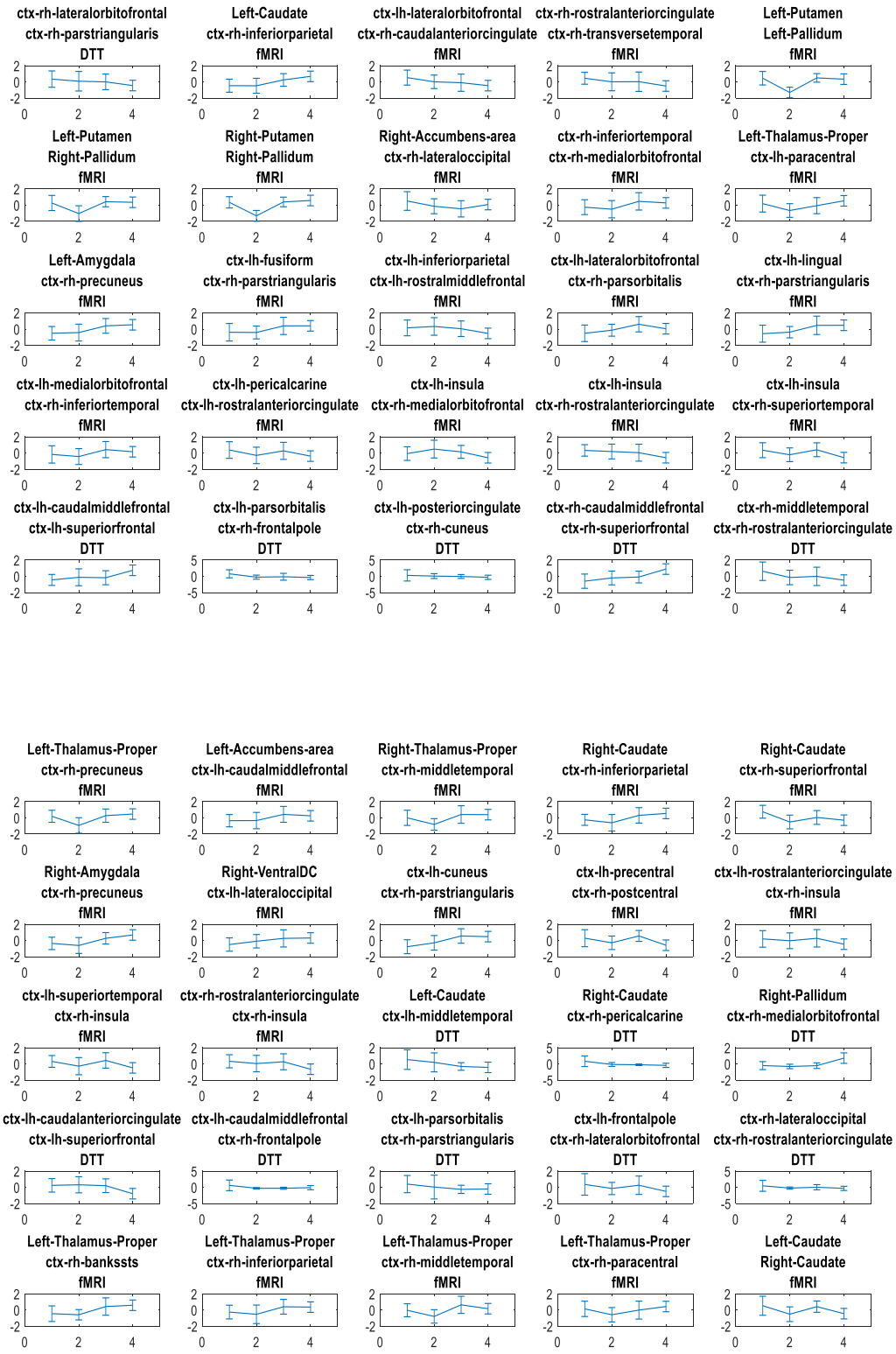

Figure 1.12.2 Plot of top 50 significant features in multimodal data selected by all-subset method