A COMPUTATIONAL MODEL OF THE BRAIN FOR DECODING MENTAL STATES FROM FMRI IMAGES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

SARPER ALKAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COGNITIVE SCIENCES

OCTOBER 2019

Approval of the thesis:

A COMPUTATIONAL MODEL OF THE BRAIN FOR DECODING MENTAL STATES FROM FMRI IMAGES

submitted by **SARPER ALKAN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in the Department of Cognitive Sciences, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek-Bozşahin
Dean, Graduate School of **Informatics**

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Sciences**

Prof. Dr. Fatoş Tünay Yarman-Vural
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Cem Bozşahin
Department of Cognitive Sciences, METU

Prof. Dr. Fatoş Tünay Yarman-Vural
Computer Engineering Department, METU

Prof. Dr. Aydın Alatan
Electrical & Electronics Engineering Department, METU

Assoc. Prof. Dr. Tolga Çukur
Electrical & Electronics Engineering Department, Bilkent Uni.

Assist. Prof. Dr. Gülşah Tümüklü-Özyer
Computer Engineering Department, Atatürk Uni.

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: SARPER ALKAN

Signature :

ABSTRACT

A COMPUTATIONAL MODEL OF THE BRAIN FOR DECODING MENTAL STATES FROM FMRI IMAGES

Alkan, Sarper
Ph.D., Department of Cognitive Sciences
Supervisor : Prof. Dr. Fatoş Tünay Yarman-Vural

October 2019, 128 pages

Brain decoding from brain images obtained using functional magnetic resonance imaging (fMRI) techniques is an important task for the identification of mental states and illnesses as well as for the development of brain machine interfaces. The brain decoding methods that use multi-voxel pattern analysis that rely on the selection of voxels (volumetric pixels) that have relevant activity with respect to the experimental tasks or stimuli of the fMRI experiments are the most commonly used methods. While MVPA based on voxel selection is proven to be an effective approach, we argue that an alternative approach exists, which resembles the processing hieararchy of the human brain for the processing and the representation of the mental states.

In this study, we propose a hierarchical brain model for brain decoding. The hierarchical model we propose first clusters a brain image into sets of voxels where the voxels that have a highly correlated activity with each other fall into the same set, which we call supervoxels. Using the supervoxels, we aim to capture the nervous activity from specialized brain regions, which are assumed to process a distinct aspect of a given stimulus or mental task such as processing color, texture, or shape of a given visual object. Then, we combine the brain activity represented by each supervoxel using a method that we call Brain Region

Ensembles (BRE) in order to decode mental states from fMRI images. Our analyses on multiple fMRI datasets show that the BRE is much better suited to the classification of mental states from fMRI images than classical voxel selection methodology. Additionally, we show that BRE can be used for the specification of brain regions that are relevant to the experimental tasks or stimuli when the aim is to identify the regions that have discriminative activity with respect to two different mental states.

Keywords: MVPA, fMRI, Brain Decoding, Clustering, Classifier Ensembles

ÖZ

FMRI GÖRÜNTÜLERİNDEN ZİHİNSEL DURUMLARIN ÇÖZÜMLENMESİ İÇİN HESAPLAMALI BİR BEYİN MODELİ

Alkan, Sarper
Doktora, Bilişsel Bilimler Bölümü
Tez Yöneticisi : Prof. Dr. Fatoş Tünay Yarman-Vural

Ekim 2019 , 128 sayfa

İşlevsel manyetik rezonans görüntüleme (iMRG) yöntemi kullanılarak elde edilen beyin görüntülerinden beyin çözümlemesi, zihinsel hastalıkların teşhisi, zihinsel durumların belirlenmesi ve beyin makine arayüzlerinin geliştirilmesi için önem arzetmektedir. İMRG görüntüleme deneylerindeki deneysel görev ya da uyaranlarla ilişkili aktivite gösteren voksellerin (hacimli piksel) seçimine dayanan çoklu-voksel örüntü çözümlemesi (ÇÖVÇ) yöntemleri bu iş için en çok kullanılan yöntemlerdir. Her ne kadar, voksel seçimine dayanan ÇÖVÇ, etkinliği gösterilmiş bir yaklaşım olsa da, buna beynin işlem aşamalarına ve gösterimlerine benzer yapıda çalışan bir alternatif bir yaklaşımın olduğunu savunuyoruz.

Bu çalışmada, zihinsel durumların çözümlenmesi için aşamalı bir beyin modeli önderiyoruz. Bu önderiğimiz aşamalı model, ilk olarak beyin görüntüsünü, birbirine işlevsel olarak yüksek benzerlik gösteren voksellerden oluşan, süpervoksel olarak adlandırdığımız voksel gruplarına bölütlüyor. Bu süpervokselleri kullanarak, verilen uyaranlar ya da zihinsel görevlerin işlenmesinde görev alan özelleşmiş beyin bölgelerindeki (görsel bir nesne için renk, doku veya şekli işleyen bölgeler gibi) sinirsel aktivitelerinin elde edilmesini amaçlıyoruz. Sonrasında her bir süpervoksel ile elde edilen beyin aktivitelerini, zihinsel durumun çözümlenmesi için, Beyin Bölgesi Toplulukları (BBT) adını verdiğimiz yöntemle birleştiriyoruz. Birden fazla iMRG veri kümesindeki analizlerimiz gösteriyor ki, BBT, zihinsel durumların sınıflandırılması işlevine klasik voksel seçimi yöntemlerinden daha uyumludur. Ayrıca, BBT'nin iki farklı zihinsel durumun için farklı aktivite gösteren ve deneysel uyaranlar ya da görevlerle ilişkili beyin bölgelerinin belirlenmesinde kullanılabileceğini gösteriyoruz.

Anahtar Kelimeler: ÇVÖÇ, iMRG, Beyin Çözümleme, Kümeleme, Sınıflandırıcı Toplulukları

To my parents

Azmi Alkan, Hatice Alkan,

and my lovely wife

Banu

ACKNOWLEDGMENTS

With my deepest gratitute, I would like to thank to my supervisor Fatoş T. Yarman-Vural for her endless support and her guidance through the ups and downs of this journey. She has been a role model for me not only in the academical matters, but in all aspects of life. It was a great honor and pleasure for me to work with her.

I would like to thank to my thesis committee members Prof. Dr. Aydın Alatan, Prof. Dr. Cem Bozşahin, Assoc. Prof. Dr. Tolga Çukur, and Assist. Prof. Dr. Gülşah Tümüklü Özyer for their constructive criticism and their support in all stages of this thesis.

I would like to thank to my colleagues in the Image Processing and Pattern Recognition Laboratory of METU, Dr. Mete Özay, Dr. Çağlar Şenaras for their friendship, their helpful comments and the encouragement they provided. I would also thank to Dr. Serdar Çiftçi, Dr. Orhan Fırat, Dr. Itır Önal, Hazal Moğultay, Barış Nasır, Arman Afrasiyabi, Burak Velioğlu, Emre Aksan, and Güneş Sucu for their friendship, for providing a great working environment, and for their cooperation.

I would like to thank to my friends and colleagues in the Informatics Institute, Dr. Fatih Nar, Dr. Rahime Belen Sağlam, Ali Sağlam, Dr. Zeynep Başgöze, Dr. Özkan Bayraktar, and Pelin Bayraktar for their support and friendship. I would like to thank Dr. Berna Bakır Batu, for her friendship, and for randomly dropping by and getting me through a proof up until midnight.

I would like to thank to my dear friends Taşkın Sözen, Serdar Aldanmazlar, Devrim Dikol, Kemal Levent Yaman, Berna Bakır Batu, Rahime Belen Sağlam for the good times they provided, who helped me to stay sane during this journey. I would like to express my special gratitutes to Oğuz Uçarcı and Evrim Turan for being there for me all the time, and having me in the midlle of the night to write this thesis in their place. I would also like to thank to my dear departed friend Serkan Orcan for all the good times we had together. Rest well my friend.

I owe special thanks to my dear friends Umut Batu and Selim Nar for their unending support, encouragement, and most importantly for their great friendship for the half of my life.

I would like to express my gratitute to my parents Azmi Alkan and Hatice Alkan, my dear sisters Deniz Alkan and Emel Alkan for their continous and unending support. Also, I would like to thank to my family, Belgin İnce Yetik, Semih Yetik, Halil İnce, Gizem İnce Erol, Berrin İnce, and especially İbrahim Özkan for their support.

Last but definitely not the least, I would like to thank to my love and my life partner, Banu for supporting me all these years and helping me to finish this thesis. It was her encouragement, her will, and her sacrifices that enabled me see this through. She brought us two lovely kids Tuna and Bora who have been the joy of our life, while our daughter being underway. Thank you Banu for all you have done. Together, we will make all of our dreams come true.

I thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) for supporting this study by the project (no 114E045) titled "Multi-layered cognitive learning model".

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS xi

LIST OF TABLES xix

LIST OF FIGURES xx

LIST OF ALGORITHMS xxviii

LIST OF ABBREVIATIONS xxix

NOMENCLATURE xxx

CHAPTERS

1 INTRODUCTION 1

1.1 What is Brain Decoding? 3

1.2 Why Brain Decoding? 3

1.3 Conceptual and Methodological Problems with Brain Decoding 4

1.3.1 Problem of Overfitting the Classifiers 5

1.3.2 Interpretation of Classification Results for Brain Decoding 6

1.3.2.1 Statistical Significance of Classification Results for Brain Decoding . . 6

1.3.2.2 Non-existance of Significant Brain Patterns 7

1.3.2.3 Identification of Cognitive Function vs. Representation in the Brain . . 7

1.3.3 Distributed Representations in the Brain . . . 8

1.3.4 Experimental Design 8

1.4 Our Contribution 9

1.5 Organization of the Thesis 10

2 LITERATURE SURVEY FOR THE BRAIN DECODING TECHNIQUES BASED ON FMRI 13

2.1 Part 1: Functional Magnetic Resonance Imaging Techniques 13

2.1.1 Parameters of fMRI Data Acquisistion 14

2.1.2 Representation of fMRI data 15

2.1.2.1 Data Representation for Event Related Experiments 16

2.1.2.2 Data Representation for Experiments with Block Design 16

2.2 Experimental Setup and the Data collection 17

2.2.1 Objects Dataset 17

2.2.2 Emotion Dataset 18

2.2.3 Pre-processing of the Emotion and Objects Datasets 19

2.2.4 Tower of London (TOL) Dataset 19

2.2.5 Train, Validation, and Test Sets 20

2.3 Part 2: Multi Voxel Pattern Analysis (MVPA) 21

2.4 Pattern Classifiers for MVPA 21

2.4.1 The Factors that Affect Brain Decoding Using Pattern Classifiers 23

2.4.1.1 Within-Class Variance in fMRI Data 23

2.4.1.2 Between-Class Variance in the fMRI Data 24

2.4.1.3 Distributions of Training, Validation, and Test Sets 24

2.4.1.4 Size of the Dataset for Sufficient Statistics 24

2.4.1.5 Type of the Pattern Classifiers Frequently used in Brain Decoding . . 25

2.4.1.6 Construction of Feature Vectors . . 26

2.4.2 Voxel Selection Strategies for MVPA 26

2.4.2.1 ROI Selection 27

2.4.2.2 Voxel Selection with ANOVA . . . 28

2.4.2.3 Voxel Selection with Mutual Information (MI) 28

2.4.2.4 Voxel Selection with Support Vector Machine (SVM) Weights 29

2.4.3 Dimensionality Reduction Techniques for MVPA by Feature Transformations 29

2.4.4 Mesh Network Representation 30

2.4.5 Time-series analysis 30

2.4.6 Methods that use Limited Voxel Spaces for Brain Decoding 31

2.4.6.1 Searchlight Methods 31

2.4.6.2 Clustering Methods for Brain Parcellation 31

2.5 Chapter Summary 32

3 A NEW BRAIN DECODING TECHNIQUE: BRAIN REGION ENSEMBLES (BRE) 35

3.1 Stage 1: Supervoxels and Brain Parcellation 36

3.1.1 Data preparation 37

3.1.2 Supervoxels 38

3.1.3 Homogenity Metric 39

3.1.4 Spatially Constrained Normalized Cuts Clustering (N-Cuts) for Brain Parcellation 40

3.1.5 K-Means Clustering for Brain Parcellation . . 41

3.2 Stage 2: Fuzzy Stacked Generalization for Supervoxels . 42

3.2.1 Logistic Regression 43

3.2.1.1 Binary classification 44

3.2.1.2 Multi-class Classification 45

3.2.1.3 Classifier outputs 46

3.2.2 Base layer classifiers 47

3.2.3 Meta classifier 48

3.3 Stage 3: Subsets of Supervoxels for Meta Classifiers . . . 49

3.3.1 Subsets Generated with Specific Parameters . . 50

3.3.2 Random Subsets of Supervoxels 51

3.4 Stage 4: Brain Region Ensembles for Brain Decoding . . 52

3.4.1 Formal definiton of a Brain Region Ensemble . 52

3.4.2 Random Subsets of Supervoxels for Brain Region Ensembles (RSS-BRE) 54

3.4.2.1 RSS-BRE with all Supervoxels . . . 55

3.4.2.2 RSS-BRE within Brain Partitions that are Tied to Specific Clustering Parameters 56

3.4.3 BRE within Brain Partitions that are Tied to Specific Clustering Parameters 56

3.4.4 Selection of the Best Performing Supervoxel with respect to Classification Accuracy 57

3.5 Diversity Measures for Classifier Ensembles 57

3.6 Region Specification with BRE 58

3.7 Chapter Summary 60

4 VALIDATION, VERIFICATION AND EMPIRICAL ANALYSIS OF BRAIN REGION ENSEMBLES (BRE) METHOD 63

4.1 Classifier Diversity Analysis 63

4.1.1 Diversity of Base Layer Classifiers that Use Supervoxels 64

4.1.1.1 Base Layer Diversity for Objects Dataset 66

4.1.1.2 Base Layer Diversity for Emotion Dataset with 2 Classes 66

4.1.1.3 Base Layer Diversity for Emotion Dataset with 4 Classes 67

4.1.1.4 Base Layer Diversity for TOL Dataset 67

4.1.2 Diversity comparison of Brain Region Ensembles that use Random Subsets of Supervoxels (RSS-BRE) with respect to Random Subspace Ensembles of Voxels (RSE) 68

4.1.2.1 Diversity Comparison of RSS-BRE vs. RSE for the Objects Dataset . . 70

4.1.2.2 Diversity Comparison of RSS-BRE vs. RSE for the Emotion Dataset: 2 Classes 73

4.1.2.3 Emotion Dataset: 4 Classes 73

4.1.2.4 TOL Dataset 80

4.2 Decoding Mental States by BRE 80

4.2.1 Data Preparation for Clustering and Classification 80

4.2.2 Clustering Parameters and Methods for Generating Supervoxels 82

4.2.3 Classification Parameters and Methods for the Validation of BRE 83

4.2.3.1 Classification Strategies Based on BRE 83

4.2.3.2 Classification Strategies Based on Voxel Selection 84

4.2.4 A Comparative Analysis of BRE and Baseline Methods on Objects Dataset 84

4.2.4.1 BRE Performances Obtained by Different Sets of Supervoxels 84

4.2.4.2 BRE Performances Obtained by Random Subsets of Supervoxels 87

4.2.4.3 Comparison of BRE with Voxel Selection 87

4.2.5 A Comparative Performance Analysis of BRE and Baseline Methods using Emotion 2 Classes Dataset 87

4.2.5.1 BRE Performances Obtained by Different Sets of Supervoxels 87

4.2.5.2 BRE Performances Obtained by Random Subsets of Supervoxels 89

4.2.5.3 Comparison of BRE with Voxel Selection 90

4.2.6 A Comparative Performance Analysis of BRE and Baseline Methods using Emotion 4 Classes Dataset 90

4.2.6.1 BRE Performances Obtained by Different Sets of Supervoxels 90

4.2.6.2 BRE Performances Obtained by Random Subsets of Supervoxels 93

4.2.6.3 Comparison of BRE with Voxel Selection 93

4.2.7 A Comparative Performance Analysis of BRE and Baseline Methods using Tower of London (TOL) Dataset 93

4.2.7.1 BRE Performances Obtained by Different Sets of Supervoxels 93

4.2.7.2 BRE Performances Obtained by Random Subsets of Supervoxels 98

4.2.7.3 Comparison of BRE with Voxel Selection 98

4.3 Selection of Brain Regions that Contribute to the Task of Classification of Mental States 98

4.3.1 Comparison of the Regions Specified using Base Layer Classifiers with the Existing Neuroscience Literature with respect to the FMRI Datasets used in this Study 99

4.3.2 Comparison of the Regions Specified using Base Layer Classifiers with the Regions Specified with ANOVA with respect to the FMRI Datasets used in this Study 100

4.4 Chapter Summary 101

5 CONCLUSION AND FUTURE DIRECTIONS 113

5.1 Future Directions 115

REFERENCES 117

CURRICULUM VITAE 127

LIST OF TABLES

TABLES

Table 4.1 The average accuracies for 10 validation runs are presented via per subject basis for the Objects dataset. The maximum accuracy acquired for each subject is marked with bold text. 85

Table 4.2 The average accuracies for 10 validation runs are presented via per subject basis for the Emotion dataset with two classes (emotional - non emotional). The maximum accuracy acquired for each subject is marked with bold text. 88

Table 4.3 The average accuracies for 10 validation runs are presented via per subject basis for the emotion dataset with four classes (furniture - kitchen appliances - fear - disgust). The maximum accuracy acquired for each subject is marked with bold text. 91

Table 4.4 Average accuracy results over 6 cross-validation runs are presented individually for subjects 1-6 in the TOL dataset. 94

Table 4.5 Average accuracy results over 6 cross-validation runs are presented individually for subjects 7-12 in the TOL dataset. 95

Table 4.6 Average accuracy results over 6 cross-validation runs are presented individually for subjects 13-18 in the TOL dataset. 96

Table 4.7 Accuracy results averaged over all subjects and 6 cross-validation runs are presented for TOL dataset 97

LIST OF FIGURES

FIGURES

Figure 2.1 Slice orientations and basic positional terminology for fMRI images, adopted from [97]. 14

Figure 2.2 BOLD signals and data sampling from fMRI experiments of block design and event-related design are depicted. The horizontal lines represent the time. The samples marked with a box are used in the construction of the design matrices for the respective types of experiments. The color of the stimulus boxes signify stimuli from a specific class where the samples are associated with. The figure is adopted from [64]. 15

Figure 2.3 Schematic representation of the experimental process for the Objects dataset. Images that belong to each object category were presented in a random sequential order. After each image presentation the participants' task were to indicate if the image belonged to the category of the previously presented image. In this example, the images that are indicated with red arrows belong to the category of the images that precede them. Image from Onal-Ertuğrul [101] . 18

Figure 2.4 The aim of a TOL puzzle is to reach the goal state by poping one ball at a time from the top of any one rack and placing it on any of the other racks that has an empty space. Images are from [71]. . . 19

Figure 3.1 Schematic layout of the suggested Brain Region Ensembles. In Stage 1, highly correlated voxels are grouped into clusters by using either a clustering algorithm (K-means or N-cuts), or a functional brain atlas (AAL). In the second stage, for each supervoxel, a base layer classifier is trained by using one-leave-out cross validation in the training set. Using these classifiers, class posterior probabilities are acquired for training, validation and test samples. In the third stage, subsets of supervoxels are formed. In the fourth stage, meta classifiers that are based on subsets of supervoxels are formed and then used for brain decoding. In this diagram, N_c stands for number of supervoxels, and N_ψ stands for number of supervoxel subsets. 37

Figure 4.1 Q-Statistic (column on the left, lower is better) and Disagreement Measure (column on the right, higher is better) for subspaces specified with N-Cuts, K-Means and AAL. The first row (a,b) shows the diversity values for the Objects dataset while the second row (c,d) is for the Emotion 2-Class experiment (e,f), the third one is for the Emotion 4-Class experiment, and the last one is for the TOL dataset (g,h). The number of supervoxels in the brain partitions that are generated by clustering algorithms are given at the bottom of every graph (100 to 1700). 65

Figure 4.2 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Objects dataset. The top row of figures (a,b) represents the results for subject 1, while the bottom row (c,d) represents the results for subject 2. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 71

Figure 4.3 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Objects dataset. The top row of figures (a,b) represents the results for subject 3, while the bottom row (c,d) represents the results for subject 4. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 72

Figure 4.4 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 2-Classes dataset. The top row of figures (a,b) represents the results for subject 1, while the bottom row (c,d) represents the results for subject 2. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 74

Figure 4.5 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 2-Classes dataset. The top row of figures (a,b) represents the results for subject 3, while the bottom row (c,d) represents the results for subject 4. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 75

Figure 4.6 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 2-Classes dataset. The results are given for Subject 5. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 76

Figure 4.7 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 4 Classes dataset. The results are given for the Subjects 1 (a,b) and 2 (c,d). The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 77

Figure 4.8 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 4 Classes dataset. The results are given for the Subjects 3 (a,b) and 4 (c,d). The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 78

Figure 4.9 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 4 Classes dataset. The results are given for the Subject 5. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 79

Figure 4.10 Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for TOL dataset. The results are given for the average of all Subject. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700). 81

Figure 4.11 The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and all 4 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for Objects dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS-BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter. 86

Figure 4.12 The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and all 5 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for Emotion 2 Class dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS - BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter. 89

Figure 4.13 The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and all 4 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for Emotion 4 Class dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS - BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter. 92

Figure 4.14 The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and 18 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for TOL dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS - BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter. 97

Figure 4.15 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 2 (b) of the Objects dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner. 102

Figure 4.16 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 3 (a) and 4 (b) of the Objects dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner. 103

Figure 4.17 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 1 (b) of the Emotion 2 Class dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner. 104

Figure 4.18 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 1 (b) of the Emotion 2 Class dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner. 105

Figure 4.19 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for all subjects from TOL dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner that are averaged across all 18 Subjects. 106

Figure 4.20 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 2 (b) of the Objects dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$. . . . 107

Figure 4.21 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 3 (a) and 4 (b) of the Objects dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$. . . . 108

Figure 4.22 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 2 (b) of the Emotion 2 Class dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$. 109

Figure 4.23 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 4 (a) and 5 (b) of the Emotion 2 Class dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$. 110

Figure 4.24 Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for all Subjects of TOL dataset, averaged over the Subjects. The effective voxels are selected using ANOVA with $p \leq 0.01$. 111

LIST OF ALGORITHMS

ALGORITHMS

Algorithm 3.1 Base layer classification 47

Algorithm 3.2 Meta classification 49

Algorithm 3.3 A Brain Region Ensemble 53

Algorithm 3.4 Random Subsets of Supervoxels for Brain Region Ensembles 54

Algorithm 3.5 Selection of Optimal Clustering Parameter by RSS-BRE . 56

Algorithm 3.6 Selection of Optimal Clustering Parameter by BRE . . . 57

Algorithm 3.7 Best Performing Supervoxel 57

LIST OF ABBREVIATIONS

AAL	Automated Anatomical Labeling
ANOVA	Analysis of Variance
EEG	Electroencenography
ERP	Event Related Potential
fMRI	Functional Magnetic Resonance Imaging
FSG	Fuzzy Stacked Generalization
kNN	K-Nearest Neighbor
MI	Mutual Information
MNI	Montreal Neurological Institute
MRI	Magnetic Resonance Imaging
MVPA	Multi-Voxel Pattern Analysis
N-Cuts	Normalized Cuts
PLS	Partial Least-Squares
BRE	Brain Region Ensemble
ROI	Region of Interest
RSE	Random Subspace Ensembles
RSS-BRE	Random Subsets of Supervoxels for Brain Region Ensembles
SRM	Shared Response Modeling
SVM	Support Vector Machine

NOMENCLATURE

$B_{m,c}$	Bernoulli random variable for cluster c and sample m
c	A supervoxel
C	A set of supervoxels
C^{θ}	A set of supervoxels that are generated by using the clustering parameter θ
C^{ψ}	A subspace of supervoxels
C^{Ψ}	A set of supervoxels that is used to generate a set of subspaces $C^{\psi} \in C^{\Psi}$
$\delta(\cdot)$	Kronecker delta function
$\mathbb{L}$	Set of class labels
$\mathbb{L}^N$	Set of vectors of class labels $\mathbb{L}$ of size N
N_c	Number of clusters
N_l	Number of (unique) class labels
N_s	Number of data samples
N_{te}	Number of test samples
N_{tr}	Number of training samples
N_{val}	Number of validation samples
N_v	Number of voxels
N_{θ}	Number of clustering parameters
$\mathbb{R}$	Set of real numbers
ϕ	A regularization parameter
Φ	A set of regularization parameters
ϕ_{best}	The best regularization parameter within a set of regularization parameters
ψ	An index for a subset of supervoxels
Ψ	A superset of supervoxels
θ	A clustering parameter
Θ	A set of clustering parameters
$\mathbb{R}^{N_1 \times N_2}$	Set of matrices of real numbers of size $N_1 \times N_2$

$Q_{i,j}$	Q-statistic between ith and jth classifiers
$D_{i,j}$	Disagreement measure between ith and jth classifiers
v	A vector of voxel intensity values for a specific voxel for all experimental samples
$\mathbf{x}$	A row vector of voxel intensity values that correspond to an experimental sample
$\mathbf{X}$	A set of fMRI samples in matrix form (the design matrix), where columns correspond to individual voxels and rows correspond to samples
x_{ij}	An entry of the design matrix where i corresponds to the sample index while j corresponds to the voxel index
$\mathbf{X}_{tr}^{c}$	A set of samples from the training set for the cluster c
$\mathbf{X}_{val}^{c}$	A set of samples from the training set for the cluster c
$\mathbf{X}_{te}^{c}$	A set of samples from the training set for the cluster c
y	Class label for a sample
$\mathbf{Y}$	An array of class labels
$\mathbf{Y}^{tr}$	An array of class labels for the training samples.
$\mathbf{Y}^{tr}$	An array of class labels for the validation samples.
$\mathbf{Y}^{te}$	An array of class labels for the test samples.
$\widetilde{\mathbf{Y}}_{tr}^{c}$	An array of class posteriori probabilities of the supervoxel c for the training samples
$\widetilde{\mathbf{Y}}_{val}^{c}$	An array of class posteriori probabilities of the supervoxel c for the validation samples
$\widetilde{\mathbf{Y}}_{te}^{c}$	An array of class posteriori probabilities of the supervoxel c for the test samples
$\mathbf{X}_{tr}^{\psi}$	An array of class posteriori probabilities concatenated for the supervoxel subset ψ for the training samples
$\mathbf{X}_{val}^{\psi}$	An array of class posteriori probabilities concatenated for the supervoxel subset ψ for the validation samples
$\mathbf{X}_{te}^{\psi}$	An array of class posteriori probabilities concatenated for the supervoxel subset ψ for the test samples
$\widehat{\mathbf{Y}}_{val}^{\psi}$	An set of predicted class labels for the subset ψ of supervoxels for the validation samples
$\widehat{\mathbf{Y}}_{te}^{\psi}$	An set of predicted class labels for the subset ψ of supervoxels for the test samples
$\widehat{\mathbf{Y}}_{te}^{\Psi}$	An set of predicted class labels for the superset Ψ of supervoxels

Z_c	A binominal random variable that correspond to the probability of having the number of successful guesses more than or equal to a certain threshold for the base layer classifier trained using supervoxel c as the input
$\{item^{parameter}\}$	A set of *items* each of which are specified by a *parameter*

CHAPTER 1

INTRODUCTION

Human brain has been a source of inspiration and a subject of curiosity for the researchers over the centuries. We seek to understand how brain takes control of the body, process the sensory input and comes up with meaning from the experiences. Also we try to simulate it, the way it processes information at the neural level, and within its deep and intricate networks. Lately, we started to scratch beneath the surface of the mystery of the brain. On one side, with the technological advances, we started to be able to build deep neural networks which display the slivers of power of this magnificent processor. On the other side, advances in brain scanning techniques helped us to decipher the functionalities of the brain by looking at the measured activities. In this thesis we deal with the latter issue. We seek to understand how human brain works when a person is engaged with a particular task or presented with a specific stimulus.

In order to decipher the information encoded within the brain, the related activities need to be measured in some way. Also, to be able to make sense of the measured brain activity, the activity itself needs to be about some specific brain states. The process of capturing brain activity that is about specific stimuli or specific mental tasks is called functional brain imaging, where the function is specified by the stimuli or the mental tasks.

Functional brain imaging techniques consists of a variety of approaches depending on the way that they capture the brain activity. Brain imaging techniques like electroencenography (EEG), which records electrical signals through electrodes placed on scalp, can be time locked to measure brain activity during the presentation of a specific stimulus. Event related potentials (ERP), which are the EEG signals recorded within a time-frame after the onset of a stimulus presentation, are used for that purpose. Brain images can also be captured by measuring the differences in blood oxygenation levels within the brain through magnetic resonance imaging (MRI). This procedure allows us to capture the three dimensional images of human brain that are composed of volumetric pixels (voxels) of several millimeter cubes. Such images can be captured at every 1-3 seconds depending on the properties of the image capturing equipment and image resolution. Similar to ERP, this procedure can be time locked with the presentation of a specific stimulus, or they can be captured while the patricipant is performing a mental task (such as, multiplication of two numbers, playing a game, or recalling an item from a list of objects which are viewed beforehand), which results in a series of images that are associated with a function, which is called functional MRI, or fMRI. This study is focused on the analysis and modeling of fMRI images. In the rest of the thesis, the term *brain image* refers to an image captured by an fMRI acquisition device under a pre-defined set of stimuli.

The aforementioned functional aspects of the image capturing processes (presentation of a specific set of stimuli or assignment of a mental task) are specified by an experimental procedure. The primary aim of such experiments is to bring the brain state of the participants of the experiment about a mental state which is then captured by the brain imaging apparatus. Depending on the context, the phrase, *mental state* can have a variety of ontological references. The specification of a mental state can be as broad as being awake or asleep. Alternatively, mental state can refer to a state that is vaguely defined such as distracted, curious, or doubtful. Also, it can refer to an emotional state such as fearful, or angry. Finally, it can refer to the occurrence of a distinct, well-defined process such as viewing a pattern of oriented lines [50], a particular object [66], reading a particular word [66], recalling an item from a memorized list [67], engaged with gambling [24], playing a game [71]. In the context of this dissertation, in conjunction with literature on the computational analysis of brain patterns, we will use the term mental state in the last two of the aforementioned senses of the phrase, where a mental state can be an emotional state, or a well defined mental process (such as the examples given above) that is specified by the conditions of the fMRI experiments.

The aim of this thesis is to analyze the patterns of brain activity which are captured by fMRI methods that are correlated with distinct mental states, caused by the experimental tasks or stimuli. Our analysis involves the classification of fMRI images according to the mental states and the specification of the regions of brain volumes that significantly contribute to the classification task. In other words, we seek to decode brain activity from the functional brain images. In this study, *brain decoding* is done by using a novel framework that involves activity dependent segmentation of brain images captured by fMRI and building classifier ensembles that are based on the segmented regions.

The distinguishing aspect of this study is that, we present a computational model of the human brain, which we use to decode the mental states from fMRI images. The computational model that we present is *Brain Region Ensembles* (BRE). In this computational model, we use two observations regarding the mental representations in the human brain. First, the human brain processes information coming from the senses using specialized brain regions (such as color, texture, or shape processing within the visual cortex) that is distributed across the brain. Second, the brain combines the activity of the specialized regions to come up with a coherent mental state regarding the stimuli (such as a visual object category) or the mental task that is currently presented. With BRE, we try to model this computational pattern. In BRE, we try to capture the activity of specialized regions in the brain by functionally homogenous voxel groups which we call *supervoxels*. Then, we combine activity of the supervoxels using classifier ensembles in order to decode the overall mental state. While we do not claim that BRE is exactly how the brain processes a given stimuli or a mental task, we claim that BRE captures two essential components of the brain processes that are: distributed mental representations in specialized brain regions, and a way to combine these distributed mental representations.

In this chapter, we provide an introduction to the problem of brain decoding based on fMRI data, and methodological issues associated with it. Then, that we describe our contribution to the field of brain decoding using fMRI data. Lastly, we present the organization of the thesis.

1.1 What is Brain Decoding?

Brain decoding is the process of analyzing patterns of brain activity in brain images in order to gain insights on the workings of the brain [72]. The early methodologies of brain decoding were limited to identification of the voxels, therefore brain regions, that are relatively active compared to other regions during the presentation of a stimulus. These active brain regions were assumed to be contribute to underlying cognitive process makes an implicit representation of the process. Nowadays, this simple approach is mostly discarded. Most of the current approaches now utilize pattern classifiers and other computational methods that use groups of voxels and utilize the intricate relations between them in order to decode brain patterns [72, 19, 60]. These type of analyses are called multi-voxel pattern analysis (MVPA), which we discuss in detail in the next chapter.

1.2 Why Brain Decoding?

Brain decoding, in its extreme forms, has been the focus of science fiction literature, movies, and TV-shows. For that reason alone, the topic can be compelling to researchers. While somewhat less amazing when compared to science fiction, the real applications of brain decoding are still impressive. In this section, we briefly discuss the importance of brain decoding.

Perhaps the most obvious and most important use of brain decoding is the ability to gain insights about a living and functioning human brain in a systematic manner. With brain decoding, brain regions that take part in processing of the experimental tasks or stimuli can be identified [50, 90]. Also, the correlations between individual brain regions while processing the stimuli can be decoded [101, 80]. Brain decoding, can, also be used to deduce thoughts and experiences from the brain images [87]. Moreover the nature of the processing (such as distributed or localized) in the brain for a given set of stimuli can be identified [66, 4]. Even the effects of the cognitive mechanisms, such as visual attention on the cortical representations, can be decoded from brain images [22]. There has been ongoing research about the brain machine interfaces to control of the machinery with brain signals including prosthetic arms [85], or the machines that turn brain signals to voice commands for people who are under complete paralysis [46]. Examples can be added to these, where in the Chapter 2, we present a detailed background on the applications of brain decoding to various application areas.

In addition to the insights gained about the nature of the brain processes, illnesses such as Alzheimer's disease, and mild cognitive impairment can be diagnosed with less than 1% error rate [14]. Also, the differantial diagnosis mental illnesses that show similar symptoms, such as schizophrenia, bipolar disorder, and schizoaffective disorder were made possible by using brain decoding methods [26].

Since the seminal article by Haxby et al. [43], when we look at the progress in the last 20 years, we expect brain decoding to find more application areas as well as providing valuable information on the working processes and conditions of the brain.

1.3 Conceptual and Methodological Problems with Brain Decoding

Despite the advances in neuroscience and cognitive sciences, brain decoding applications have some conceptual and methodological problems. One of the main problems with brain decoding is about the use of pattern classifiers, where the size of the dataset obtained from fMRI experiments are not statistically sufficient for the application of modern deep learning methods.

Another problem regarding brain decoding methodologies is the difficulty of interpretation of the results, obtained as the output of the brain decoding algorithm. With brain decoding, brain regions that are relevant to the experimental tasks can be identified. However, the regions that are found to be non-relevant, might have some relevant activity that can only be observed a sub-voxel resolution. Moreover, when using pattern classifiers to identify a region to be relevant for the processing regarding an experimental task, it is not always evident whether the particular region is only concerned with the representation of the task stimulus, or the region is directly involved with the functional processing of the task. For instance, visual objects can be decoded from the early visual areas (or even from retina) if a brain imaging system with high enough resolution can capture the images from them. However, that does not mean visual object recognition is performed at retina, or at early visual areas.

The third problem is the compexity of the neural representation of a cognitive process in the brain. It can be appealing for various reasons to exclusively mark a single brain region for the performance of a certain mental task or for processing a certain type of stimuli. For instance, searchlight analysis only considers voxel groups that contain voxels which are spatially close to each other [32]. However, we argue that, neural representations that are distributed across voxels, and even across multiple brain regions are much more likely than the representations that are contained in a single brain region for the mental tasks or stimuli used in the fMRI experiments. For instance in the fMRI experiments that are used in this study, presentation of visual objects [66], participants playing games [71], or participants viewing emotionally stimulating images all create neural representations that are distributed across neural regions. Thus, while the above argument holds, brain decoding algorithms that can make use of the information distributed across brain regions offer a better representation to discover the patterns in the voxel activity.

One last problem with the brain decoding methodologies is about their relations with the parameters and the methods of the fMRI experiments. A brain decoding methodology that is fit for one type of fMRI experiment can be unsuitable for another. Choosing the right methodology for the data at hand is important. In the following subsections, we take a detailed look at the above mentioned issues with brain decoding.

1.3.1 Problem of Overfitting the Classifiers

The primary challenge for any brain decoding system that uses pattern classifiers is posed by the nature of the brain imaging process. Due to the experimental constraints, only a limited number of samples can be obtained from each participant within the time frame that is safe for them. Furthermore, the resolution of the brain images are in the order of 20 to 200 thousand voxels (volumetric pixels) for each brain volume. The high number of voxels is both a blessing and a curse. While a high spatial resolution creates opportunities for a detailed investigation of anatomical regions corresponding to a brain state, it creates a high dimensional feature space of voxels for a pattern classifier. Considering the fact that the number of samples are relatively low compared to the dimensionality of the feature space, designing a classifier becomes a very difficult problem. Moreover, fMRI images are contaminated by the noise that is caused by the uncertainties regarding the image acquisition process and the cognitive disposition of the subject during the experimental process. High number of voxels, coupled with the low number of samples cause the data represented in the voxel feature space to be sparse. Both the noise and the sparsity of the feature space must be handled to improve test performance of classifiers and validity of the classification. For the nearest neighbor classifiers such as k-nearest neighbor (kNN), a high dimensional feature space makes all data points almost equidistant to a query sample, which makes the decision unreliable. For the linear classifiers, the high dimensionality of the feature space reduces the generalization performance through the dependency of the decision hyperplane to the training set of a small sample size. The classifiers, due to their opportunistic nature, are prone to memorize spurious patterns within high dimensional voxel spaces given a low number of training samples. As a result, the classifier overfits to the accidental patterns in the training data.

The problem of overfitting is more prominent for the complex, non-linear classifiers that require a high number of parameters to be trained. As the number of training parameters increase, the decision boundary within the feature space becomes convoluted. Moreover, when the number of training samples are limited, the classifier can easily be biased by the noise or by the outlier samples. In particular, this problem prevents the application of the modern classification methodologies such as deep learning to fMRI analysis, which thrives on high number of training samples in order to train complex networks. In addition to that, network architectures that are useful for the analysis of regular (2D) images, such as convolutional neural networks, are not as effective for fMRI analysis: The use of convolutional layers in a deep neural network reduces the number of total parameters in the network by using the assumption that a set of basic patterns (such as oriented edges) are repeated across the whole image. For the regular images, this assumption holds because an object can appear anywhere in the image. Also, the basic patterns can occur for a wide array of objects and shapes in a regular image. However, none of these are valid assumptions for fMRI images. Even for the primary visual cortex, the resolution of fMRI images are not high enough for the voxels to form repeated basic patterns, where Kamitani and Tong [50] can barely detect the sub-voxel activity that correspond to the oriented gratings that are shown to the participants. In the rest of the brain, we can not think of any reason to expect such repeated patterns of voxel activity to occur.

While there are more than one way to deal with this problem, the problem of overfitting is still a major obstacle for fMRI analysis [60]. In the brain decoding literature, the solution to this problem is found through the elimination of the voxels that are not relevant to the experimental tasks. The most commonly used approaches to decode mental states from fMRI images use voxel selection [19, 18, 56, 34, 2], region of interest (ROI) selection [36], dimensionality reduction [59], or searchlight analysis [50, 32] as ways to overcome the problem of overfitting for pattern classification. In the following chapter, we present the available solutions to this problem in the MVPA literature.

1.3.2 Interpretation of Classification Results for Brain Decoding

The use of pattern classifiers for brain decoding has become an established practice in the literature. For brain decoding, however, the results should be interpreted with caution in order to prevent the misjudgements regarding the attibution of cognitive functions to brain regions. In this section, we discuss the various pitfalls concerning the interpretation of the mental state classification results.

1.3.2.1 Statistical Significance of Classification Results for Brain Decoding

Brain decoding methods that rely on pattern classifiers use the classification accuracy, which is higher than chance level, as an indication for the involvement of a brain region with an experimental task [42, 60]. Alternatively, pattern classifiers can be used to identify brain regions that are selectively active for two distinct classes of experimental tasks or stimuli [4]. However, due to the low number of experimental samples (20-100 samples per class) in fMRI experiments, the statistical significance of the classifier accuracies is questionable. For instance, just higher than 50% accuracy obtained from a two class classifier might not be considered significant without sufficient number of test samples. In order to remedy this problem, statistical significance of the results should be carefully investigated.

1.3.2.2 Non-existance of Significant Brain Patterns

In brain decoding, a common problem is to explore the brain regions that are relevant and/or irrelevant to a particular experimental task. Especially after testing for statistical significance, the result of mental state classification might deem many brain regions to be irrelevant with respect to the experimental tasks or stimuli. However, non-existance of significant brain patterns does not entitle the brain regions to be irrelevant with respect to the given experimental tasks [50]. The reason for that is the coarseness of the voxel representation. Each voxel within a brain image represents the collective activity of thousands of neurons within a time frame of 2-3 seconds. From this perspective, some neural activities can be hidden by relatively more dominant neural activities within the same voxel.

1.3.2.3 Identification of Cognitive Function vs. Representation in the Brain

It is a common knowledge that human brain processes information in a hierarchical manner. When we look at the base of processing hierarchy, we observe the sensory neurons transducing the inputs from the outside world into neural representations at the first level. Then, the inputs received by the sensory neurons processed gradually in the neural hierarchy. From this perspective, we can see that the group of neurons that take inputs from the sensors in the retina have all the information necessary for representing a perceived visual object. However, we cannot attribute the function of object recognition to these neurons alone. This leads us to one of the overlooked pitfalls of brain decoding, which is *retina decoding* [49, 19]. The pitfall can be understood by imagining a brain decoding system that has the access to the retinal image of a perceived object as an input, through an advanced brain imaging procedure. Then, the contents of the image can be decoded by a sufficiently complex non-linear classifier (for instance, a deep neural network). However, that does not entitle the retina for performing a complex computation to decode the object. In other words, presence of complex representations within brain regions does not ensure presence of equally complex computations within the same regions. Thus, the brain decoding applications should avoid complex calculations to attribute functional properties to brain regions.

1.3.3 Distributed Representations in the Brain

When we consider the neural representations of cognitive functions in the brain, we observe that even very specific and simple stimulus; such as, specifically oriented lines elicit neural activity that are distributed across groups of neurons [48]. While one specific neuron might give the highest response to a stimulus, we cannot exclude the activities, (or inactivities) of other neurons from the total representation. Following this line of thought with voxels might sound problematic at first, since each voxel aggregates the activity of thousands of neurons. However, the mental states to be analyzed by using fMRI experiments are usually more complex than the participants observing oriented lines. Moreover, even a simple stimulus, such as, oriented line gratings can be detected by classification of activations of voxel groups [50]. Thus, an activity pattern that is distributed across multiple brain regions can be expected from fMRI experiments that present some higher order mental tasks to the participants such as processing semantic object categories [66], memory retrieval [67, 8], emotional processing [40], gambling [24], or playing a game [71]. For instance, semantic representation of objects is known to be distributed across multiple brain regions including sensory and motor cortices [63, 91, 52]. Furthermore, visual object representation is, also, distributed across occipital and inferior temporal cortices [25]. Also, emotional processing of fear and disgust inducing stimuli is known to be implicated with middle frontal gyrus, fusiform gyrus and insula, and amygdala [88].

When the nature of distributed representations in the brain is considered, it can be hypothesized that different aspects of a given mental task or stimulus would be processed in different brain regions. For instance, semantic categories of concrete objects are represented in both sensory and motor cortices as well as cerebellum and primary visual cortex [66, 91], probably because we recognize them through our interactions with them. We could be combining the sound of a bouncing ball with its visual shape and motor cortex representation of the way we handle the ball in order to form the semantic representation of a ball. In other words, it is likely to be the case that diverse aspects of the same semantic category are represented at different regions in the brain. If this is the case, then an ensemble of pattern classifiers, each of which are dedicated to model voxel groups of a different brain region can be successful in decoding the mental states, where it is known that a diverse set of pattern classifiers combined in a classifier ensemble can perform better than any single classifier [57]. While there are methods that use classifier ensembles for brain decoding, they do not utilize diverse representations that are distributed across brain regions. Instead of that, they use classifier ensembles on a set of voxels which are selected by means of common voxel selection strategies [56, 55].

1.3.4 Experimental Design

In a successful fMRI experiment, which is expected to measure activities of a set of mental states, brain images captured during the experimental procedure must be matched exclusively to specific mental processes under investigation. In order to conform with this criterion, two basic strategies for experimental design is utilized: Block design and event-related design [6].

In an experiment with block design, various types of stimuli, or types of cognitive tasks are presented sequentially to a participant during fMRI recordings. This strategy does not involve a resting state period and proceeds with switching between the types of different stimuli, or tasks. The aim of such experiments is to capture the brain state during each stimulus block for a relatively long time period (up to several minutes). Experiments with event-related design on the other hand, involves brief presentation of the task stimulus followed by a resting period which allows the investigation of the transient stimulus onset and offset periods in the analysis [94].

The type of experimental design can promote or discourage the use of a particular classification strategy. In this thesis, we construct a brain decoding methodology that suits both of the above experimental design procedures.

1.4 Our Contribution

The contribution of this thesis can be listed as follows:

1. We propose a new computational model of the human brain that we use for brain decoding, which is consistent with the state-of-the-art neuroscientific findings. For this model, our primary goal is to capture distributed representations of the mental states in the brain in terms of functionally homogenous voxel groups that we call supervoxels, and then combine those representations with classifier ensembles in order to decode mental states from fMRI images. Based on this goal, we hypothesise that, a computational model that can capture these distributed representations and utilize them, would achieve a better performance for decoding the mental states than available computational models that rely on selecting the relevant voxels or anatomical regions. Thus, just like the brain that processes incoming information in specialized brain regions and combine their activity to achieve a coherent mental state, our model combines the voxel activity from the functionally homogenous voxel groups in order to decode the present mental state.

2. In order to capture the distributed representations of mental states in brain, we use brain parcellation. For brain parcellation, we utilize unsupervised clustering techniques on series of fMRI images to find groups of voxels that are functionally correlated. These homogeneous groups of voxels that we call *supervoxels* constitute the basis for brain decoding, and they also contribute to the identification of the brain regions that are relevant in the processing of the experimental tasks. The supervoxels offer a solution to problem of overfitting, where the voxel space within each individual supervoxel is much smaller when compared to voxels from the whole brain for the classification purposes. Furthermore, we propose that the supervoxels can create a diverse basis of features that is desirable for ensemble learning.

3. We propose a new ensemble learning method named *Brain Region Ensembles*

(BRE) for brain decoding. This method uses supervoxels to encode distinct aspects of the mental processes within the brain in terms of the activity of functionally homogenous groups of voxels (supervoxels). The encodings are then fused within multiple random subsets of voxel clusters by using meta classifiers of fuzzy-stacked generalization algorithm. The classification results from each random subset are then combined by majority voting. We show that BRE achieves higher performance on mental state classification than the state of the art methodologies that use voxel selection.

4. We compare the effectiveness of supervoxels that are formed by different methodologies for the classification of mental states. The voxels are clustered by the methods that either use the voxels functional relationships (using K-means clustering), or their functional relationships while constrained by their spatial proximity (using spatially constrained normalized cuts clustering). Also, we analyze the supervoxels that are formed by the collection of voxels in anatomically labeled brain regions (AAL regions). We show that supervoxels formed by the clustering methods yield better classification results than the ones formed with AAL regions.

5. We propose a method to identify the voxel clusters that are effective in the classification of the mental states. For that purpose, we use a statistical significance measure that selects the discriminative supervoxels. We show that the distribution of the voxels that belong to those discriminative supervoxels within anatomical brain regions are consistent with their functional properties specified in the cognitive neuroscience literature. Furthermore, with our analysis we were able to discover variations for the representation of visual objects in the brain of an individual subject that challenge our understanding of visual pathways.

1.5 Organization of the Thesis

The thesis is organized in the following chapters:

In Chapter 2, we first discuss the nature fMRI experiments and the data collection procedures for the datasets that we use in the thesis. Then, we present an overview of brain decoding methods in the literature in order to better place this thesis in the domain of Cognitive Sciences. A brief introduction to the progress of fMRI analysis methods is followed by the presentation of current state of the art in the context of the question: How can we make an effective use of the data provided by fMRI experiments?

In Chapter 3, we first present the fMRI datasets that we use in the thesis. Then, we provide our method BRE which makes use of the information encoded in clusters of functionally correlated voxels (supervoxels) for cognitive state classification. Additionally, we show how supervoxels can be used to effectively identify brain regions that encode differential information across cognitive tasks. Furthermore we present methods to analyze the working principles of BRE. For this purpose, we use classifier diversity analysis.

In Chapter 4, we present results of our brain decoding methodology on datasets we introduce in the Chapter 3. We start with the analysis of BRE with respect to cluster diversity measures. After that we provide the results of the mental state classification experiments that uses BRE in comparison with state of the art methods of brain decoding. Lastly, we present the results of our region identification procedure by using supervoxels in relation with the cognitive neuroscience literature.

In Chapter 5, we present our concluding remarks and future directions for the research.

CHAPTER 2

LITERATURE SURVEY FOR THE BRAIN DECODING TECHNIQUES BASED ON FMRI

In this chapter, we present the nature of fMRI data, the datasets that we use in this study, and also the brain decoding approaches available in the literature.

The chapter is divided in two parts. In the first part, we describe the techniques for the acquisition of fMRI data, the data samples, and their relations with the subjects. Also, we present the formal representations for the fMRI data that we use throught this thesis. In the second part, we discuss the brain decoding methods in the literature with a focus on classifier based methods. In that part we also present a basic classification process in order to set up the notations which we follow afterwards.

2.1 Part 1: Functional Magnetic Resonance Imaging Techniques

The fMRI data is a collection of brain images that are captured by using an MRI device while the experimental participant performs mental tasks which are specified by an experimental procedure designed for the analysis of the predefined stimuli or mental tasks. The characteristics of fMRI data are specified by the nature of the experimental tasks, and the technical specifications of the MRI device as well as the settings of the MRI device during the experimental procedure.

In this section, we first present the basics of fMRI data acquisition process. Then we describe the formal notations of fMRI data samples, which we use in this thesis. Following that, we present the datasets that are used in this study.

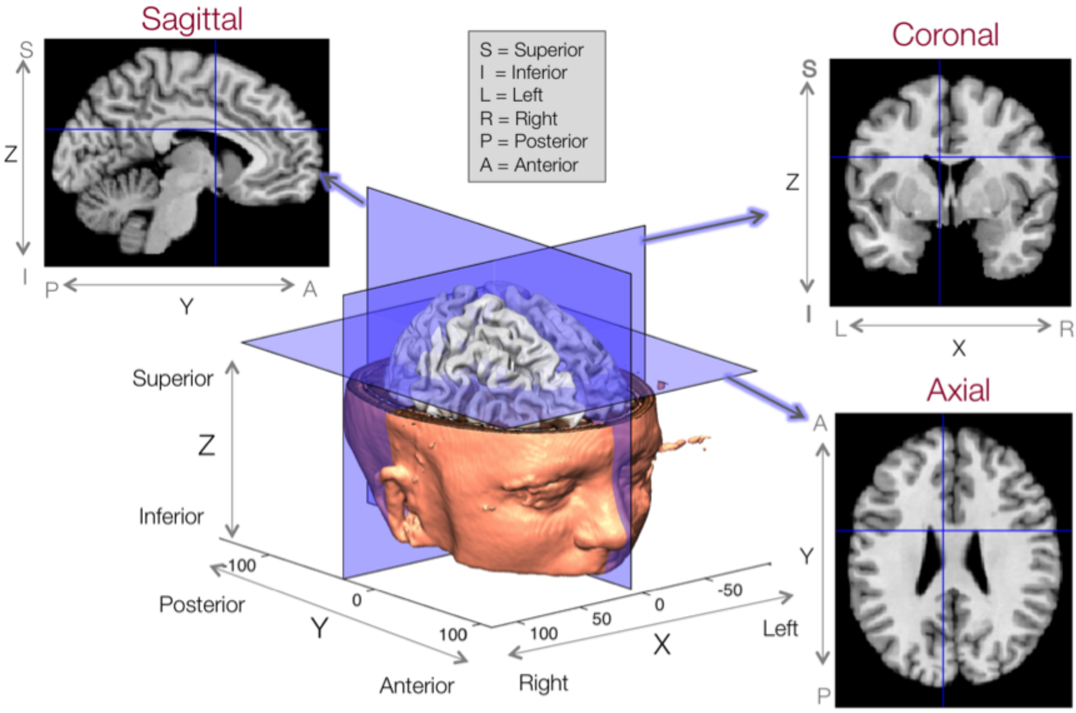


Figure 2.1: Slice orientations and basic positional terminology for fMRI images, adopted from [97].

2.1.1 Parameters of fMRI Data Acquisistion

For functional MRI scans, an MRI device measures the blood oxygenation level dependent (BOLD) contrast, which is a result of the increased blood flow within the active regions of the brain. The process itself is quite complicated and is out of the scope of this study.

While the acquisition of the fMRI images is a fairly complicated process, the basic temporal parameters that we are interested in are as follows: The repetition time in milliseconds (TR) specifies the time between the consecutive pulses of the MRI device, and echo time in milliseconds (TE) which specifies the time between the pulse and the echo of the pulse from the tissue, which is then recorded as an MRI slice.

The basic spatial parameters are the slice orientation (axial, saggital, or coronal) with respect to the head orientation (Figure 2.1), inter-slice gap in millimeters, and slice thickness in millimeters. The slice orientation determines the orientation in which the slices are recorded in MRI. The spatial parameters specify the volume of the resultant voxels, while the temporal parameters specify the time frame of a single fMRI image.

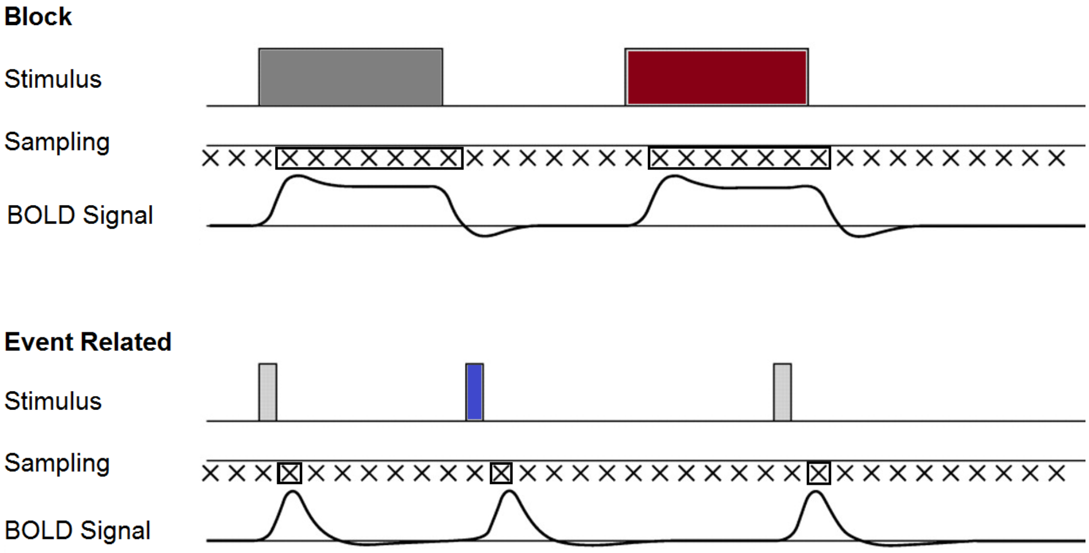


Figure 2.2: BOLD signals and data sampling from fMRI experiments of block design and event-related design are depicted. The horizontal lines represent the time. The samples marked with a box are used in the construction of the design matrices for the respective types of experiments. The color of the stimulus boxes signify stimuli from a specific class where the samples are associated with. The figure is adopted from [64].

2.1.2 Representation of fMRI data

The fMRI data, depending on the type of the experimental design (event-related or block design) comes in consecutive *brain volumes*, where each volume is a three dimensional collection of voxels that capture one time-frame of brain activity. In this thesis, we test our brain decoding methodology on each single subject. Thus, the data representation methods that are presented in the following sub-sections refer to the collection and use of the data from a single subject.

In experiments that we used in this study, the data samples are collected at consecutive and seperate stimulus presentations for each participant. A collection of these stimulus presentations are called *epochs* where specific sets of stimuli are presented to the participant. Individual experimental epochs might be performed at different experimental sessions or at different days. Thus, in order to take the changes across sessions (alignment of the head, the attentiveness of the participant) into account while testing the brain decoding algorithms, it is recommended to not mix the samples recorded from distinct epochs when forming train, validation and test sets. At the end of this section we present how we formed these sets according to the individual properties of the fMRI datasets we use.

2.1.2.1 Data Representation for Event Related Experiments

In the fMRI experiments with event related design, each stimulus onset is captured by a series of brain volumes with gradually changing voxel intensities from the rest level just after the stimulus onset, and then returning to the rest level after a short period of time (10-12 seconds). The stimulus is usually presented for a short period of time to allow the brain to return to the resting state. As a result, a total of 5-6 brain volumes per presented stimulus is recorded. At this point, the use of the volumes obtained by the experiments can differ depending on the brain decoding applications. Some brain decoding applications well-utilize all of the brain volumes through the onset and offset of a stimulus [29, 73], while other applications prefer to use only the peak voxel intensities for each stimulus presentation [5, 66].

In our preliminary testing steps of our model we did not observe any particular benefit of using all of the voxel volumes, or average of the volumes for a stimulus presentation. Thus, we use peak intensities for the data from the event-related experiments. As a result, a data sample, the peak voxel intensitiy values that are obtained from a particular stimulus presentation during an event related experiment can be represented as a single brain volume of voxel intensity values (Figure 2.2), which can be flattened into a row vector of $\mathbf{x} = \{x_1, x_2, ..., x_{N_v}\}$, where N_v denotes the number of voxels in the brain volume. Conversely, the peak voxel intensity values for a single voxel for the whole experiment is represented by $\boldsymbol{\upsilon} = \{\upsilon_1, \upsilon_2, ..., \upsilon_{N_s}\}$, where N_s represents the number of stimulus presentations (or data samples). Then, the data samples that are obtained from the whole experiment is denoted by a design matrix $\mathbf{X} \in \mathbb{R}^{N_v \times N_s}$, where each row ($\mathbf{x}$) is a vector containing all voxel intensity values from a data sample while each column ($\boldsymbol{\upsilon}$) is a vector containing the voxel intensity values of a single voxel across all data samples. For classification purposes, each data sample $\mathbf{x}$ can then be associated with the class of stimulus $y \in \mathbb{L}$ that the sample belongs to. Here, $\mathbb{L} = \{0, 1, ..., N_l - 1\}$ is the set of all class labels each of which are signified by a different integer, and N_l is the number of all class labels. In order to show the correspondence with the set of all samples $\mathbf{X}$, class labels of all samples are denoted by $\mathbf{Y} \in \mathbb{L}^{N_s}$.

2.1.2.2 Data Representation for Experiments with Block Design

In the experiments with block design, the stimuli, or the mental tasks are presented in long time frames where MRI device captures a large number of voxel volumes during each time frame. In each time frame a single mental task, or multiple stimuli of the same class is presented to the participants, thus, forming the volumes that belong to a single class.

For this type of fMRI experiments, such as the Tower of London (TOL) experiments which we present in the following sections, the design matrix $\mathbf{X} \in \mathbb{R}^{N_v \times N_s}$ consists of all the brain volumes captured by the imaging device during the fMRI experiment. Here, N_v is the number of voxels in a single brain volume, and N_s is the number of brain volumes captured during the presentation of mental tasks or stimuli that are specified with any class label $y \in \mathbf{Y}$. In this case a single sample $\mathbf{x} = \{x_1, x_2, ..., x_{N_v}\}$ is the vectorized form of any brain volume captured during the presentation of a mental task or stimulus that has a label (Figure 2.2). Class labels of all samples $\mathbf{Y} \in \mathbb{L}^{N_s}$ is then formed by the assigning all the samples within the same block with the a particular class label $y \in \mathbf{Y}$. Here, $\mathbb{L} = \{0, 1, ..., N_l - 1\}$ is the set of all class labels each of which are signified by a different integer, and N_l is the number of all class labels.

2.2 Experimental Setup and the Data collection

In this study, we use three fMRI datasets in order to validate the suggested computational model. The first one is a two class dataset where the subjects were presented two categories of visual objects which we call the Objects Dataset. The second dataset is a four class dataset where the subjects were presented with emotion arousing images (fear, disgust) and emotionally neutral images (furniture, kitchen appliances) which we call the Emotion Dataset. The data in this dataset can be treated as having two classes as well (emotional, non-emotonal). These two datasets are available in the website: http://neuro.ceng.metu.edu.tr. The last dataset we use is the Tower of London (TOL) dataset where participants were presented with Tower of London puzzles and were asked to solve them.

2.2.1 Objects Dataset

This dataset was collected by the members of METU Imagelab at Bilkent University UMRAM (Ulusal Manyetik Rezonans Araştırma Merkezi). The goal of the experiment is to create a benchmark fMRI dataset for brain decoding with two visual categories. This dataset consist of fMRI images of subjects performing a one-back repetition task with two visual objects. In this task, the subjects indicated whether or not the currently displayed image matches the category (bird, flower) of the previously displayed image. Each trial was 12, 14 or 16 seconds long where the image was presented for 4 seconds, followed by 8, 10 or 12 seconds of rest period. The images were captured with a Siemens 3T Magnetom TRIO MRI system, and using an EPI sequence, where TR $=$ 2000 msec, TE $=$ 30 msec, flip angle $= 90^o$, 34 interleaved axial slices with inter-slice gap $=$ 0.2 mm, and 3x3x3 mms of voxel volumes. In each trial 6 scans were obtained.

The experiment consists of 6 epochs in total. In each epoch 36 images were displayed to the patients, 18 from each object category. The data was collected from 5 subjects. However, the data of one epoch from the 3rd subject were corrupted. Thus, we omitted the subject altogether in order to have comparable results across the subjects.

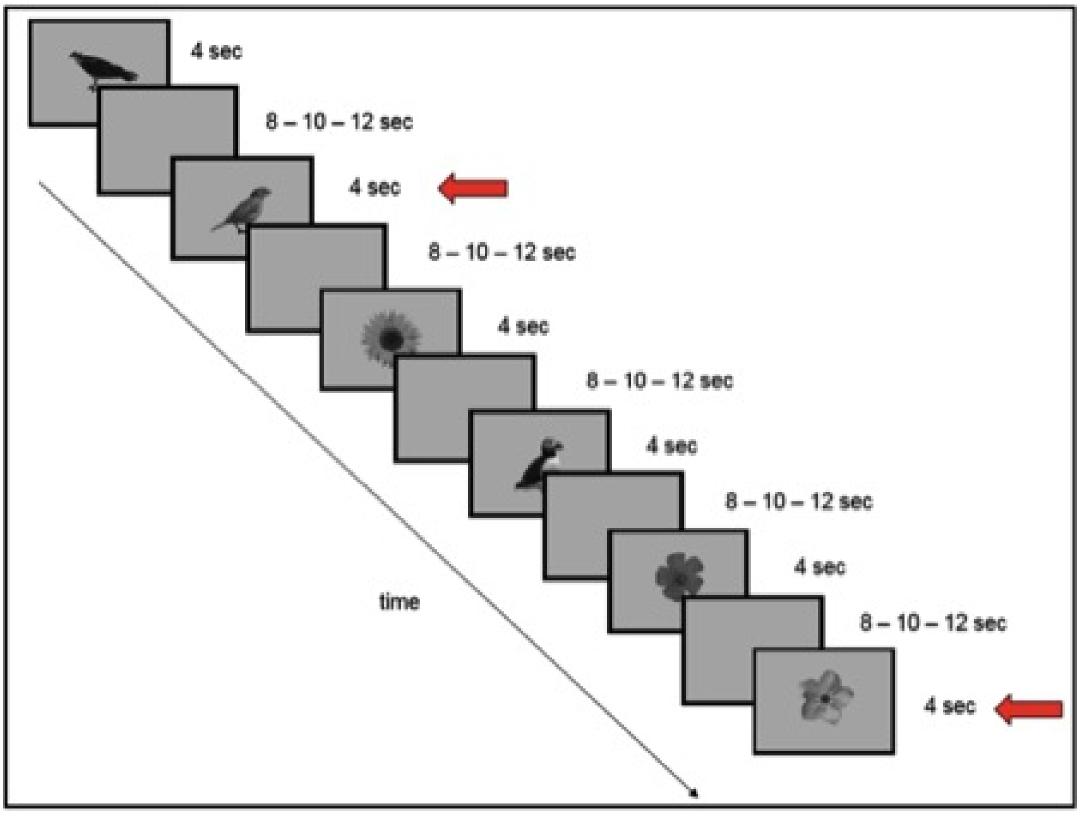


Figure 2.3: Schematic representation of the experimental process for the Objects dataset. Images that belong to each object category were presented in a random sequential order. After each image presentation the participants' task were to indicate if the image belonged to the category of the previously presented image. In this example, the images that are indicated with red arrows belong to the category of the images that precede them. Image from Onal-Ertuğrul [101]

.

2.2.2 Emotion Dataset

This dataset was collected by the members of METU Imagelab and members of Koç University Department of Psychology at Bilkent University UMRAM (Ulusal Manyetik Rezonans Araştırma Merkezi). For this dataset, the subjects were asked to decide if a query image was in the set of 5 images that were shown previously. Depending on the category (fear inducing, disgust inducing, kitchen appliances, furniture) of the presented images the experiment aims to find if there is a difference in the participants' recall rates between emotion inducing, and emotionally neutral images under the hypothesis that emotion inducing images are more likely to be recalled successfully.

Each trial started with a 12 seconds fixation period. Then the subjects were shown 5 consecutive cue images from the same category where each image was displayed for 1200 msecs. After the last image, there was a 12 secs period where the subjects were asked to solve mathematical problem that involved addition or

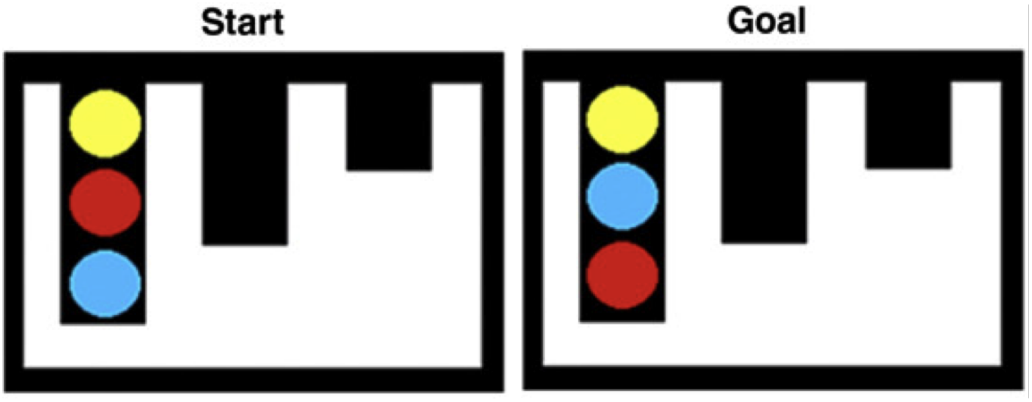


Figure 2.4: The aim of a TOL puzzle is to reach the goal state by poping one ball at a time from the top of any one rack and placing it on any of the other racks that has an empty space. Images are from [71].

subtraction of two digit integers. Following that, a query image was presented to the subjects for 2 seconds, which was in the same category with the cue images. The subjects' task were to decide if the query image was the member of the last set of cue images.

For each subject there are 6 epochs of experiments, each of which contains 70 trials, details of which are specified above. The images were captured with a Siemens 3T Magnetom TRIO MRI system, and using a gradient EPI sequence, where TR = 2000 msec, TE = 30 msec, flip angle = 90^o, 34 interleaved axial slices with inter-slice gap = 0.3 mm, and 3x3x3 mms of voxel volumes. In each trial 6 scans were obtained during the display of cue images.

2.2.3 Pre-processing of the Emotion and Objects Datasets

The images captured in the Objects and the Emotion datasets are pre-processed using the SPM8 toolbox [76] as follows: Firstly, slice acquisition timing was corrected across slices. Secondly, the images were re-aligned to the first slice to correct head movements. Thirdly, all images were normalized to a standard template given in SPM2. Then, the images were smoothed by using a 6 mm full-width half maximum isotropic Gaussian kernel. Lastly, for each voxel, the average of three highest values of 6 images which were captured at each trial were averaged and assigned as the voxel intensity value for that trial. Also, AAL regions [92] were segmented by registering the voxels to MNI space [89] by using MARSBAR toolbox [10].

2.2.4 Tower of London (TOL) Dataset

The last dataset that we utilize is the Tower of London (TOL) dataset [71]. This dataset is collected at the Indiana University. The aim of their study was to investigate the complex problem solving task of humans to the cortical structures that are related to problem solving. In particular, they seek to observe the neural responses with respect to the goal hierarchy and the number of moves to achieve the goal. In order to achieve that, the brain regions participants that are differentially active during either the planning or the execution stages while solving a series of Tower of London puzzles (see Figure 2.4) were identified. In order to solve the puzzle, the participants were expected to move the top ball from any of the three racks and put on the top of any other rack. By performing that process iteratively, the participants were expected to reach a goal state (Figure 2.4). The participants of the experiments were 22 graduate students with ages between 19 and 38. We had the access to the data of 18 of the participants in our studies which we present here.

The experimental procedure is as follows: At first, the participants were subjected to a training period where the only participants that can solve the problems within seven moves during the training period were admitted to the fMRI scans. In the scanning procedure the participants were presented with the problem for 5 seconds while the word "plan" was shown in the screen, during which the participants were instructed not to execute the problem solving. After that period, the patricipants could continue planning or begun execution of the solution where the word "execute" is presented on the screen. The execution period took 12 seconds. Then, a rest period followed where the participants were fixated on a cross on the screen. Each 590 second run contained 6 5-move problems and 12 6-move problems, where only the 6-move problems were analyzed.

The images were gathered in $2 \times 2 \times 2$ millimeters spatial resolution per voxel, and smoothed with 8 mm wide Gaussian filter at half-maximum. Then, a high-pass filter at 1/128 Hz cut-off frequency was applied to eliminate linear drifts. The images were, then, registered to MNI space [71].

2.2.5 Train, Validation, and Test Sets

In this study, for all purposes of parameter tuning, the experimental data is partitioned into three sets: Training, validation, and test. Since we are performing single subject analysis for the Objects, Emotion, and TOL datasets, we divide the dataset of each subject into three parts. For the Objects and Emotion datasets, 4 experimental epochs form the training set, one epoch is for validation and one for testing. For the TOL dataset, since there are only 4 epochs, we use 2 epochs for training and one each for validation and test. For all datasets, except TOL, the experimental epochs (Objects, and Emotion) are permuted randomly 10 times to form individual cross validation runs. Since there are only 4 epochs for each subject in the TOL dataset, 10 fold cross-validation meant re-use of the same training data multiple times while only switching the test data with validation data. Thus, we run 6 fold cross-validation in order to ensure a different training set for each fold. For all classification algorithms, parameter tuning is performed for validation set. The test set results for the chosen parameters are then reported as the end results.

2.3 Part 2: Multi Voxel Pattern Analysis (MVPA)

Currently, mainstream brain decoding applications use multi-voxel pattern analysis (MVPA) as opposed to univariate analysis of voxel intensity values. MVPA employs methods from computer science and machine learning on groups of voxels in order to gain insight on the functional properties of their respective brain regions, or to classify mental states of the participants that are induced by the experimental tasks.

In the early years of the 21st century, fMRI images were mainly analyzed by using univariate voxel analysis where the activity of each individual voxel is contrasted across the cases when the subject is engaged with an experimental task and when the subject is at resting state. The voxels whose activity are significantly different than their resting state activity (measured via a t-test), were considered to be correlated with the cognitive task. However, these univariate methods are now obsolete, replaced by methods called multi voxel pattern analysis (MVPA), which makes use of the activity of groups of voxels by using modern computational tools such as pattern classifiers.

The main strength of MVPA methods lies in the observation that, each voxel is representing a collection of neurons, for which, the activity of a single neuron is uninformative if the whole collection of neurons is not considered. In a similar vein, groups of voxels are deemed to be more informative than a single voxel when they are used to understand the content and characteristics of the brain networks that they represent. Moreover, using groups of voxels can allow us to decode some sub-voxel level information from the images [50]. However, there is a serious drawback of MVPA methods. As more and more voxels are included in a pattern analysis scheme, the activity and the information content of individual voxels can become less and less significant. As a result, informative voxels can be overwhelmed by the noise of non-informative ones. For instance, in the extreme, using voxels from the whole brain for cognitive state classification fails due to the problem of overfitting as mentioned in Chapter 1.3.1.

Depending on the research questions of the fMRI experiments, the way to process the data varies across brain decoding applications. Still, one of the most widely used tools in brain decoding is the pattern classifiers. These tools are quite successful for the identification of the cognitive tasks and the active brain regions that are responsible for processing the underlying tasks [4]. They are also used to study and diagnose neural diseases [26], and even thoughts and experiences of the participants of the participants can be guessed [87]. In the next sub-section, we present a detailed background on the brain decoding applications that use pattern classifiers.

2.4 Pattern Classifiers for MVPA

Pattern classification is one of the primary methods of MVPA. For this strategy, activities of voxels, (or some features derived from voxel activities) are represented in a feature space. Then, a classifier that is trained on these features are used to discern different cognitive states. The general procedure for training classifiers and the basic notation for inputs and outputs are presented as follows:

Let $\mathbf{x} \in \mathbb{R}^{N_f}$ be an N_f dimensional feature vector that represents a brain volume (or an aggregation of multiple brain volumes) with a label $y \in \mathbb{L}$, where a brain volume is the set of voxels captured at each scan of the MRI device. The features can be the voxel activity values for each brain volume, or a transformation of them such as peak, or mean value of individual voxels across multiple volumes. Also, in almost all MVPA applications the features are a subset, or a transformation of the voxel activity values recorded at each brain volume which we discuss in the following sub-sections. Here, $\mathbb{L}$ is the set of all labels each of which corresponds to a mental state specified by the fMRI experiment.

A classifier, then, learns a mapping from a set of training samples ($\mathbf{x}_{tr}$) that are represented by a matrix $\mathbf{X}_{tr} \in \mathbb{R}^{N_{tr} \times N_f}$ to a set of training class labels $\mathbf{Y}_{tr} \in \mathbb{L}^{N_{tr}}$. The training set of feature vectors is in matrix form where rows correspond to experimental trials (samples), and each column correspond to a different feature dimension, or the activity of a single voxel. The set of class labels are represented in vector form, and N_{tr} being the number of training samples. The way to form training data matrix can differ between datasets. In the previous section, we have described how to select training datasets. Using the training dataset, a classifier model can then be formed by using the training features and class labels:

$$model = train(\mathbf{X}_{tr}, \mathbf{Y}_{tr}), \tag{2.1}$$

where *train* is a procedure for training a pattern classifier. The training procedure usually involves using a separate set of validation samples $\mathbf{X}_{val} \in \mathbb{R}^{N_{val} \times N_f}$ and class labels $\mathbf{Y}_{val}$ in order to optimize classification parameters during the training procedure.

Given the trained classifier, for a test sample which is an fMRI image that correspond to an unknown mental state, the feature vector $\mathbf{x}_{te}$ for that sample can be formed by using the procedure that is applied to the training samples. Using the trained *model*, the class label for the test sample, $\hat{y_{te}}$ can then be predicted by

$$\hat{y}_{te} = predict(model, \boldsymbol{x}_{te}). \tag{2.2}$$

The *predict* procedure applies the mapping from the feature vectors to class labels that correspond to mental states that is learned by the *model* to the test samples. Here, the predicted class label $\hat{y}_{te}$ may or may not be equal to the actual class label y_{te} for that sample. The details of the methodologies that we use in this thesis regarding *train* and *predict* procedures are explained in Chapter 3.

In the next subsection, we present the factors that affect the success of pattern classification strategies for brain decoding.

2.4.1 The Factors that Affect Brain Decoding Using Pattern Classifiers

The success of the pattern classifiers in brain decoding depends on many factors: (**1**) the similarity in the voxel representations with mental states with the same label (i.e. within-class variance); (**2**) the difference of the voxel representations of mental states with different labels (i.e. between-class variance); (**3**) whether or not the training, validation, and the test samples come from the same distribution; (**4**) the total number of training, and validation samples (i.e. the size of the dataset for sufficient statistics); (**5**) the type of the pattern classifier; (**6**); and lastly, the nature of the procedure that is used to form the feature vectors. Let us examine how these problems are handled in the current brain decoding literature in the following subsections.

2.4.1.1 Within-Class Variance in fMRI Data

One of the prominent factors which determines the classifier performance for brain decoding is the variance of the fMRI data that represent the mental states with the same label for the same subject. Within-class variance can be caused by many factors including design of the experiment, the uncertainties introduced by the image acquisition process and participant's focus during the experiments. The factors can cause systematic deviations or random noise.

As an exapmle for the systematic deviations, the tiredness of the participant can cause loss of attention, thus results in the later sections of an experimental trial to be less relevant to the presented experimental tasks or stimuli resulting in a large within-class variance. We have observed such a case, when we examined the results of a two class dataset (Objects dataset), where the accuracy of our classifiers decayed as we analyzed the later part of each subject's experimental session. In order to account for such an effect, we use random sampling among experimental trials for the formation of training, validation and test sets.

Random factors, on the other hand, are more problematic. The issue with the training data with random noise is that the dataset can have outlier samples. An outlying sample in the training set can then cause the classifier to be biased, which is the result of overfitting during the training procedure. A biased classifier has propensity to fail, when finding the correct decision boundary that separates the examples from different categories. This issue can be partly remedied by the application of regularization during the training session. Regularization reduces the convolutions in the decision boundary of the classifier, and helps the classifier to be less prone to the outlier examples. Regularization is widely used in MVPA applications especially useful when the dimensionality of the feature space is high. One way to apply regularization is to reduce the number of model parameters of the classifier (see Chapter 3 for an example). Another way is to regularize the feature space, where total number of significant features are reduced by sparsification of the feature space [65, 20].

2.4.1.2 Between-Class Variance in the fMRI Data

Another factor that determines the success of pattern classifiers is how distinct the samples from different classes. For an fMRI experiment, the distincness of the samples are determined by the distinctness of the mental states under consideration and the specific stimuli to elicit those mental states. In addition to that, an fMRI sample is a brain volume, a volume of voxels that correspond to a mental state specified by the fMRI experiment. Given any degree of specialization within brain regions with respect to the mental tasks, some (if not the most) of the voxels activities would not be relevant to the mental state specified by the fMRI experimental setup. In a feature space composed of the voxel activity values, activities of the voxels which are irrelevant to the mental states can overwhelm the activities of the relevant voxels thus reducing the between-class variance. Limiting the voxels to be used for brain decoding to the voxels that are highly correlated to the experimental tasks is a way to keep between-class variance high [56, 18].

2.4.1.3 Distributions of Training, Validation, and Test Sets

For a certain pattern classifier to have comparable performance in training, validation, and test sets, it is required to have the samples from all sets to come from the same distribution. When this fact is kept in mind, it seems to be reasonable to mix the samples from all epochs from an fMRI experiment when forming training, validation and test sets. The reason for that is an experiment can be performed by the same participant in multiple sessions sometimes spanning across days. However, such an approach would undermine the performance of the classifier when the novel test samples are gathered much later than the training of the classifier in a real scenario. Thus, it is suggested to not shuffle the samples across experimental epochs [95].

2.4.1.4 Size of the Dataset for Sufficient Statistics

As the number of training samples increases for a classification problem, more and more complex classifiers can be utilized to solve the problem, where experimental noise becomes less of an issue. For instance, deep neural networks are able to learn highly complex functional relationships if the necessary amount of training data is available. The required number of data samples can be in the order of tens of thousands, depending on the complexity of the problem, and more is always the better [39]. However, in this age of big datasets and deep neural networks, the study of MVPA is limited by the number of fMRI sessions a participant can take. Standard fMRI experiments are run with 300-400 trials per subject, which is hardly enough for justifying any complex (deep) classifiers [71, 67].

In order to address the above problem, an approach that aggregates and registers data from multiple subjects have been utilized, which is called shared response modeling (SRM). Using this model, the responses from individual voxels are aggregated in time domain to register data from across subjects [17, 44]. As a result, the data can be gathered from multiple subjects with application of the same experimental procedure. The drawback of this method is the data being coarsely represented in the time domain, which might not be suitable for some temporally sensitive experimental stimuli [19].

The Human Connectome Project (HCP) dataset is formed for such purposes, where each voxel is precisely aligned to allow across-subject studies [8]. Such large datasets can be analyzed by using data intensive deep learning methodologies such as convolutional autoencoders [16]. Also, while not necessarily named as SRM's, wavelet transform in the time domain is frequently perfomed to aggregate the temporal aspect of data in the across-subject datasets[80, 29]. These methods however, are not suited for event-related experiments were each stimulus presented in very small time-frames.

2.4.1.5 Type of the Pattern Classifiers Frequently used in Brain Decoding

There is a wide array of possible classifiers used in MVPA applications including decision trees, linear classifiers, and classifier ensembles. Let us go over the ones that are most commonly used in MVPA applications while we discuss their pros and cons.

Decision tree classifiers form a tree where at each branching, the training data is split in two which optimize a cost function. The cost function can be an information theoretic measure such as minimization of class-entropy, or it can be as simple as selecting a feature and setting a value that splits the samples best where each of the following branches contains more samples from a specific category than the other. The tree grows until at each final node there are samples from only one category, or a limit for tree depth is reached [83].

One problem with decision tree classifiers is that they can create overly complex trees which are prone to overfitting. In order to deal with this problem, tree depths can be limited, or decison tree ensembles can be used [82]. Such ensembles (decision forests) are much more robust to overfitting, especially for the small datasets. Thus they are widely used in MVPA applications such as classifiying alcohol dependendence in patients using their brain images [100], classifying resting state MRI images in order to detect Alzheimer disease [84], or subtyping cognitive profiles for autism spectrum disorders [35].

Linear classifiers such as logistic regression and support vector machines (SVM) are the type of classifers that partition the feature space and separate the classes by using a decision hyperplane. Due to their simplicity, this type of classifiers are more robust to overfitting than decision trees. A linear classifier only learns as many parameters as the size of the feature vectors, and the parameters can be subjected to regularization in order to make the classifier more robust to overfitting. These properties make them one of the best classifiers to be used in MVPA applications where training samples are limited in number while the data is noisy and feature dimensionality is high. Kuncheva et al. have shown that support vector machines are one of the best standalone classifiers for brain decoding [55].

While the linear classifiers are robust to overfitting, their simplicity might prevent them from learning intricate relationships between the features and the data labels. This problem can be addressed by using an ensemble of classifiers each of which operates on a part of the dataset (bagging) [34], or a part of the feature space such as stacked generalization, boosting, and random sub-spaces [69, 5, 68, 55, 53]. The final decision using the classifier ensemble is then made by using a voting procedure, or training an additional classifier by using the outputs of the classifier ensemble.

2.4.1.6 Construction of Feature Vectors

The feature vectors are the final representations of the data to be fed to the pattern classifiers, which are transformed from the voxel intensity values across the experimental samples. In order to form the feature vectors, some irrelevant voxels can be discarded while some other voxel activities can be combined within a low dimensional space by using proper transformations. In these final representations, the source of the problem of overfitting (like noise and high feature dimensionality) can be eliminated to a certain extent, while making the samples that belong to different categories seperable for the classifiers using voxel transformation or elimination strategies. In the following subsections we discuss these strategies.

2.4.2 Voxel Selection Strategies for MVPA

As we have mentioned in the previous sections, fMRI data are problematic for decoding mental states because of four major factors. The first one is the high feature dimensinality of the raw voxel space (20 to 180 thousand voxels per image), and the second one is the limited number of samples for each cognitive state. Due to these factors, end-to-end methods (the methods that take the raw, or slightly pre-processed data to achieve the final classification) such as convolutional deep learning models are not easily applicable to brain decoding applications, except for learning feature representations when they are applied to a voxel space with already reduced dimensionality [37].

Thirdly, overfitting is a major problem for brain decoding. Reducing the dimensionality of the feature space, while preserving the relevant features is a direct solution to that problem. Thus, some form of feature transformation/selection is required for a successful brain decoding strategy.

Finally, one of the major drawback of voxel selection strategies is the lack of a robust measure which eliminates redundant and irrelevant voxels with respect to the underlying mental process. Whether the selected set of voxels is formed by using a ROI or a voxel selection algorithm, there is always a possibility to leave out the voxels that are relevant to the mental tasks in fMRI experiments. In some cases, their activity range might be so small, that would make them irrelevant with respect to the voxel selection criteria (for instance ANOVA), while they might become relevant if they are combined with a complementing set of voxels. For instance, a set of voxels that encode color can be combined with a set of voxels that encode texture in order to make a decision on the category of an observed object.

2.4.2.1 ROI Selection

Region of interest (ROI) selection is a method to isolate the regions that are effective for the underlying mental task to measure a predefined set of mental processes. ROI selection strategy can be powerful when the experimental tasks are correlated with easily localized brain regions. From our preliminary experiments we have concluded that classifiers that are use feature vectors which are the raw voxel intensity values from relevant regions (such as occipital cortex for Objects dataset) can outperform that use the voxel intensity values from the whole brain, in which case the classification accuracy is equal to the chance level for selecting the class labels randomly. Further improvements to this approach are made by using the information regarding the time-course relations among the voxels that are in a spatial [74, 28] or functional [36] (calculated using the correlation of voxel time-courses over the experimental epochs) neighborhood. A further improvement for this approach is brought by including the full time-course of voxel intensity values instead of the peak values at each epoch of an fMRI experiment [73].

A problem with the ROI selection is that even the simplest tasks can create widely distributed brain patterns due to the nature of information processing in the brain. In turn, it can lead to inferior classification performance due to the omitted information. For example, cognitive processing of semantic categories, which is a main research topic, is known to be distributed across various sensory and motor cortical areas as well as medial temporal lobe and occipital and parietal cortices [43, 66]. Also, emotional processing was known to be focused in amygdala [1] and prefrontal cortex [58]. However, more recent fMRI studies suggest emotional processing is also spread through the limbic system [51], medial prefrontal, and anterior cingulate cortices [31]. Another problem is that the prior knowledge regarding the brain regions that contribute to the processing of the experimental task might not be available at all.

The most widely used strategy for MVPA is to reduce the dimensionality of the feature space by using voxel selection/elimination. Feature regularization methods that we mentioned in the previous subsection is a similar approach where the voxels not necessarily selected, but their activity values are set low that they unable to affect the classifier outcome [65, 20]. However, the most common methods for voxel selection are: analysis of variance (ANOVA), mutual information (MI) and the method that use SVM weights [56].

2.4.2.2 Voxel Selection with ANOVA

ANOVA is a methodology to determine the likelihood of two sets of data to be generated by the same stochastic processes. This method is applied to each individual voxel in order to determine if the voxel responds significantly different with respect to one experimental condition when compared to another. Usually this method is used for univariate analysis of voxels (without MVPA) in order to determine the voxels that are more active, when the experimental condition is presented when compared to the resting state. For MVPA, the voxels that are more active with respect to their resting state activity, and a number of most active voxels can be selected. Then, the voxels can be used for pattern classification [2, 62, 55].

2.4.2.3 Voxel Selection with Mutual Information (MI)

Chou et al. [18] introduces a method for voxel selection by using mutual information (MI). In the original article Chou et al. [18] used beta-map values instead of the voxel intensities for their analysis.

For the feature selection, suppose that we have the training data matrix $\mathbf{X}_{tr}$ of size $N_{tr} \times N_v$, and class labels $\mathbf{Y}_{tr}$ of size N_{tr}, where N_{tr} is the number of training data samples, and N_v is the initial number of voxels. The feature matrix is formed by normalization of the input features by calculating the z-scores for each column of the original input. The elements of $\mathbf{X}_{tr}$ is denoted by x_{ij} and the elements of $\mathbf{Y}_{tr}$ is denoted by y_i, where $y_i \in \mathbb{L} = \{0, 1, ..., N_l - 1\}$, and N_l is the number of class labels.

Mutual information $MI(\mathbf{Y}, \boldsymbol{v}_j)$ across all class labels each of wich corresponding to mental states, and the voxel $\boldsymbol{v}_j$, which denotes jth column of $\mathbf{X}_{tr}$, is calculated by using the formulation proposed by [18]

$$MI(\mathbf{Y}, \boldsymbol{v}_j) = \sum_{y=0}^{N_l-1} \int_{\boldsymbol{v}_j} p(y, x_j) \left(\frac{p(y, x_j)}{p(y)p(x_j)} \right) dx_j, \tag{2.3}$$

where $p(y, x_j)$ denotes $p(\mathbf{Y}_{tr} = y, \boldsymbol{v}_j = x_j)$. The distribution $p(y, x_j)$ is estimated by using the chain rule $p(y, x_j) = p(y)p(x_j|y)$, and the Parzen-Rosenblatt window approximation $\hat{p}(x_j|y)$ of $p(x_j|y)$:

$$\hat{p}(x_j|y) = \sum_{i=1}^{N_{tr}} \delta_{y,y_i} \Gamma \left(\frac{x_j - x_{ij}}{\sigma} \right) / \left(\sigma \sum_{i'=1}^{N_{tr}} \delta_{y,y_i} \right), \tag{2.4}$$

where $\delta_{y,y_i} = 1$ if $y = y_i$, 0 otherwise. The kernel function $\Gamma(\cdot)$ is Gaussian with σ as the standard deviation. The distribution $p(x_j)$ is calculated by marginalization: $p(x_j) = \sum_{y=0}^{N_l-1} p(y)p(x_j|y)$, and we assure $p(y)$ is a uniform distribution with $1/N_l$.

After computing the mutual information values for each voxel, they are sorted and a pre-specified number voxel indexes with the highest corresponding mutual information values are selected.

2.4.2.4 Voxel Selection with Support Vector Machine (SVM) Weights

The voxel selection algorithm with SVM weights is presented in [55]. The procedure is as follows: First, two class SVM classifiers are trained by using the training data for each of N_l classes. In this setup, for each classifier, the remaining $N_l - 1$ classes are all labeled as the second class, where $\mathbb{L}$ being the set of all class labels and N_l is the number of class labels. Then, for each classifier, the SVM weights are calculated and their absolute values are sorted into separate lists. Lastly, each list is visited one by one. With each visit, the voxel index that is not selected previously, and associated to the top weight in the list is collected. This procedure is continued until a pre-specified number of voxel indexes are selected.

2.4.3 Dimensionality Reduction Techniques for MVPA by Feature Transformations

While the voxel selection methods are used for simplifying the feature space, feature transformations are used in conjunction with these methods to better represent the data in the feature space. In this section we will discuss the feature transformation methods that are used for MVPA.

Principal component analysis (PCA) and independent component analysis (ICA) are widely used feature transformations for MVPA. By the use of such linear transformations, the initial voxel space can be mapped to a lower dimensional feature space, which alleviates the problem of overfitting for the pattern classifiers.

Principal component analysis use singular value decomposition to find the eigenvalues and the eigenvectors of the training matrix $X_{tr} \in \mathbb{R}^{N_{tr} \times N_v}$. Then, the eigenvalues are sorted. A number of eigenvalues that correspond to the largest eigenvalues are then used to construct a linear mapping from the training matrix to a lower dimensional matrix $X'_{tr} \in \mathbb{R}^{N_{tr} \times N_e}$, whose feature dimensionality is equal to the number of eigenvectors that are used in the transformation (N_e). Since it is a linear transformation, PCA can be considered as another way of voxel elimination, where this time, linear compositions of voxels are selected. This method is usually combined with ICA in order to pinpoint the voxels that are involved with the fMRI tasks.

On the other hand, independent component analysis models the signal (the voxel intensity values) as a composition of the signals coming from a number of (N_e) sources. The sources then can be separated as long as they have non-Gaussian signals, and if they are statistically independent. ICA has been used to further refine the set of voxels that are already been selected by using a ROI or another voxel selection algorithm such as PCA [11]. These methods are especially useful for combining multiple fMRI datasets (i.e. forming SRMs [19]) on the same functional task and extract useful information from the combined sets [59].

2.4.4 Mesh Network Representation

While the use of voxel intensity values as features for pattern classifiers has provided promising results for brain decoding, methods that use mesh network representation has shown that information encoded by a network of voxels can be extracted and utilized for brain decoding [74].

Meshes within the local networks of voxels have been proposed to capture the information latent in the brain networks [74]. A mesh around a single voxel is formed by a number voxels that are in proximity with that central voxel either spatially [74], or functionally [36]. In this model, the value (or the time-course [73]) of each central voxel is represented by using a weighted linear superposition of the surrounding voxels. The connection weights are estimated by using Levinson-Durbin Recursion [93]. The estimated connection weights are then used as a representation of the central voxel. When this procedure is performed on all voxels of interest, an alternative representation of voxels that are composed of connection (mesh) weights are obtained.

It has been shown that using mesh weights can yield more accurate results for pattern classification than the values of the voxel intensities [74, 36, 73]. Also, the mesh weights themselves can be used to analyze the task dependent connectivity patterns in the brain networks [29]. Furthermore, mesh network representations can be encoded by Fisher-vectors, or bag of visual words for decoding mental tasks and mental states from fMRI images [30]. One drawback of this method is the high feature dimensionality in the final representation, which limits the number of voxels that can be used with this method.

2.4.5 Time-series analysis

The datasets like Human Connectome Project (HCP) provides voxel intensity values of task related voxel activity over a period of time. Therefore, the sequences of images that are gathered during an experimental task can be processed by methods that use time-series analysis such as wavelets or short-time Fourier transforms. Such methods are especially useful for brain connectivity analysis. For instance, Richiardi et al. use discrete wavelet transform to represent averaged voxel activity in AAL regions in different time-scales [80]. Then they form connectivity graphs by using the correlations between each region at each time-scale. The connectivity graphs can then be used to form feature vectors for classification purposes [80]. Ertuğrul and Yarman-Vural on the other hand, formed mesh networks by using the wavelet transformed AAL region averages on HCP dataset [29].

While these methods are useful to analyze and decode mental states across the subjects, they are not applicable to event-related experiments, where each stimulus is presented in very short time durations. In such cases time series of a trial is limited to few samples and there is not much to be gained in a time-series analysis, except for the use of their direct correlation for determination of functional proximity of individual voxels [73].

2.4.6 Methods that use Limited Voxel Spaces for Brain Decoding

Voxel selection is not the only way to define a feature space to the input of a classifier that is robust to overfitting. Training multiple classifiers on sub-sets of voxels is an alternative way to reduce the dimensionality. Searchlight and voxel clustering methods are two popular approaches to reduce feature dimensionality, as we discuss in the following subsections.

2.4.6.1 Searchlight Methods

Searchlight methods move a searchlight window that over all voxels. At each voxel, a number of voxels in the spatial proximity of the center voxel, including itself are used to train a classifier [32, 45]. This way, the feature dimensionality is reduced and the classifier becomes less prone to overfitting. Also, the regions that is sensitive to the experimental tasks can be determined by selecting the voxels where the searchlight classifier performs better than the chance level.

Searchlight methods offer a way to select out groups of voxels that are significant for the experimental task. For instance, Kamitani and Tong used a $3 \times 3 \times 3$ searchlight (1 voxel in each spatial direction around the central voxel including the diagonals) in order to detect voxels that are effective in distinguishing oriented line gratings in the visual cortex [50]. However, the exact voxels that are effective in a particular cognitive task not certain with this methodology. Any voxel that is significant for the task in the $3 \times 3 \times 3$ window can make the window sensitive for a classifier, while the others might not be significant at all. Furthermore, this method only considers spatial neighborhoods while a group of voxels can be significant together if are used for classification while they are seperated [32]. Also, the size of the searchlight window determines the effectiveness of the method, while there is not any apriori way to know which size should be used.

2.4.6.2 Clustering Methods for Brain Parcellation

Clustering is a technique that is used to determine "similar" sub-groups of data-points within a dataset. Clustering methods are defined with respect to a similarity metric, which define the relationship between individual data points, and a grouping algorithm that puts the similar data points in the same group. Typically, clustering methods employ a parameter to determine the size, or the number of data clusters. In order to determine correct clustering parameter that forms the ideal set of clusters, it is necessary to experiment with different clustering parameters and use a validation measure to test the results.

In the literature of fMRI analysis, clustering is mostly used on resting state fMRI data for functional and automated brain parcellation. For those studies, validation is done by comparing the clustering results to the functional brain atlases such as AAL [9, 21]. Clustering is also used for grouping the region responses after a time series analysis in order to find out similarly responding brain regions.[80].

Aksan et al. [3] suggested functional Markov Random Fields model to cluster voxel intensity values to find representative sub-regions in the brain that can be used for classification purposes. Their study provides a better representation with respect to the classification accuracy of the classifiers that use the voxel intensity values of the voxels within individual clusters as feature vectors than state-of-the-art clustering methods, such as K-Means, however their method is computationally more expensive than them. On the other hand, Moğultay formed a two layered cognitive architecture in order to represent fMRI data. The first layer contains mesh representation between individual voxels while second layer has mesh representations between supervoxels which are formed by N-Cuts clustering [70]. The resultant representations were then used as the inputs for a classifier ensemble. Other than the studies that are the precursors of this thesis [5, 68] where we have employed a primitive form of BRE, the usage of voxel clusters for the classification of fMRI images is limited to the above mentioned studies, to the best of our knowledge.

2.5 Chapter Summary

In this chapter, we presented the state of the art of brain decoding methods. We, first, introduced multi-voxel pattern analysis (MVPA) for brain decoding. Then we described various methods of MVPA that are based on the pattern classifiers.

We explained pattern classification strategies for MVPA. We, first, introduced the pattern classification paradigm in general. Then, we presented the factors that affect the success of pattern classification for MVPA. We pointed out the difficulties that stem from the nature of the fMRI data for classification strategies, where the most prominent of them was the problem of overfitting.

We presented the way that the problem of overfitting is handled in the current state-of-the-art brain decoding literature by using voxel selection, feature transformation, and by limiting voxel spaces (by using searchlight, or clustering). We observed that voxel selection or transformation strategies can leave out voxels relevant to the experimental cognitive tasks while searchlight methods may not be very accurate with the region predictions. Following these observations, we propose a clustering based brain parcellation model for brain decoding:

In this study, we use clustering for the non-resting state fMRI data in order to come up with a functional segmentation that is correlated with the tasks in the experiments. The clustering approach allows us to run classifiers on small subsets of voxels that are also functionally correlated, which helps us to deal with the problem of overfitting. Also, training a classifier per voxel cluster provides an effective basis for a classifier ensemble as shown in our preliminary study [5], which improves our chances to come up with a better classifier model in the end.

CHAPTER 3

A NEW BRAIN DECODING TECHNIQUE: BRAIN REGION ENSEMBLES (BRE)

In this chapter, we present a new computational model for brain decoding called Brain Region Ensembles (BRE). The primary purpose of BRE is to model the distributed representations of mental states within the brain for brain decoding.

Functionally homogenous regions of brain has been identified by dedicated experiments and cellular studies, which are then mapped onto brain atlases [92]. However, the contribution of each region for the mental representations varies for different mental states. Also, mental representations of stimuli or tasks that are presented in fMRI experiments are known to be distributed across multiple brain regions. For instance, mental representation of a concrete object is distributed across occipital, temporal, parietal, sensory, and motor cortices [66, 4], while an emotional state such as fear is represented at middle frontal gyrus, fusiform gyrus, insula, and amygdala [88]. Even within the same brain region (such as occipital cortex) sub-regions can encode different aspects of the given stimulus such as color, shape, and texture [54, 13]. We suggest that, as long as the mental representations are spread across multiple voxels, activity patterns from an fMRI experiment can be used to identify functionally homogenous voxel groups that capture distinct aspects of distributed mental representations. These homogenous voxel groups, which we call *supervoxels*, serve as the basic building blocks of BRE.

In BRE, supervoxels are not just functionally homogenous voxel groups. We assume that each supervoxel represents a distinct aspect of the mental state that is under consideration. For instance, when we consider the semantic representation of a concrete object such as a ball, we propose that each brain region encodes a specific aspect of the semantic representation. The visual cortex may encode the visual shape of a ball while the motor cortex encodes the articulation of the hands when grabbing a ball. Similarly, auditory cortex could encode the sound of a ball jumping and the sensory cortex might be encoding the feel of a ball when held. We propose that a combination of these distinct representations constitute the overall representation of a mental state. Consequently, if these distinct representations can be identified in an fMRI image, as supervoxels, their combined activity can be used for the recognition of the underlying mental state. That is the reason for the use of supervoxels for the recognition of mental states in BRE. We propose that, due to their diversity, these homogenous voxel groups can form the basis of an ensemble classification strategy that is more effective than the usual voxel selection methodologies. For the classifier ensembles, BRE makes use of fuzzy stacked generalization (FSG) and random subspace (RS) ensembles which we explain in this chapter. In our methodology, on one hand, FSG enables fusion of the information contained in the supervoxels that are distributed across the brain. On the other hand, RS provides a robust way to combine a multitude of FSG ensembles, each of which is based on a different set of supervoxels.

Supervoxels are not only useful as the basic building blocks for ensemble learning in BRE, but also, they provide a way to determine brain regions that are effective in the processing and the representation of the stimuli/tasks provided in fMRI experiments. Unlike voxel selection methods such as ANOVA, supervoxels provides access to multi-voxel activity for brain region specification. Furthermore, in contrast to searchlight methods, homogenous voxel groups are less likely to include irrelevant voxels when compared to the voxels included in a searchlight for brain region specification.

An additional strength of BRE with respect to voxel selection is that BRE makes use of the voxel activity from the whole brain instead of a selected set of voxels. This aspect of BRE is more likely to capture the latent information within the voxel groups which would be eliminated by the voxel selection processes.

In the first four sections of this chapter, we explain the stages that we presented in the Figure 3.1. In Stage 1, we describe how to form supervoxels using the brain parcellation methods, including K-Means clustering, spectral N-Cuts clustering and AAL brain atlas. In Stage 2, we present FSG algorithm and its application to supervoxels, thus forming the set of base layer classifiers. Following that, in Stage 3, we explain how we combine supervoxels in order to form subspaces of supervoxels, where each subspace is used to train a meta classifier for FSG. In Section 4, we describe how to form Brain Region Ensembles by the utilization of meta classifiers on subspaces of supervoxels.

After going over our methodology, we present the tools that measure classifier diversity that we use to compare BRE with the state-of-the-art MVPA methods that use random subspace ensembles with voxel selection. Lastly, we describe a method to identify discriminative supervoxels.

3.1 Stage 1: Supervoxels and Brain Parcellation

We begin this section with the steps we take for data preparation. Following that, we provide the definition of supervoxels. Then, we follow with the description of the homogenity metric we use, and we explain the logic behind the choice of the similarity metric. Lastly, we present the methods that we use to form

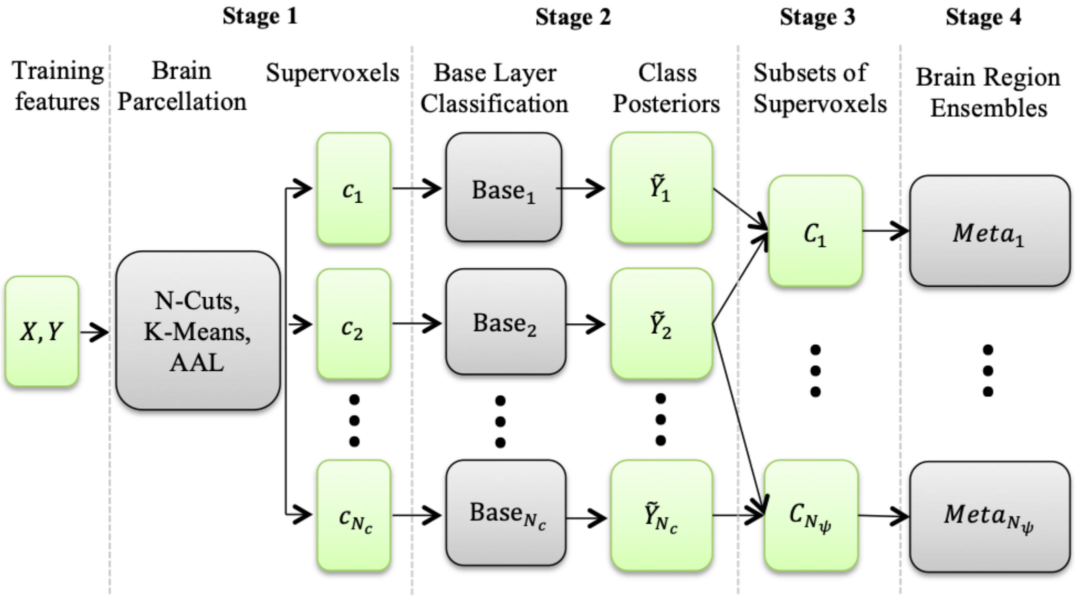


Figure 3.1: Schematic layout of the suggested Brain Region Ensembles. In Stage 1, highly correlated voxels are grouped into clusters by using either a clustering algorithm (K-means or N-cuts), or a functional brain atlas (AAL). In the second stage, for each supervoxel, a base layer classifier is trained by using one-leave-out cross validation in the training set. Using these classifiers, class posterior probabilities are acquired for training, validation and test samples. In the third stage, subsets of supervoxels are formed. In the fourth stage, meta classifiers that are based on subsets of supervoxels are formed and then used for brain decoding. In this diagram, N_c stands for number of supervoxels, and N_ψ stands for number of supervoxel subsets.

voxel clusters.

3.1.1 Data preparation

In BRE, we aim to capture the activity of the brain regions that are specialized for encoding/processing different aspects of a given stimulus, or a mental state (such as color, texture, or shape of a visual object). In order to achieve this goal we partition the brain volume of an fMRI image in functionally homogenous regions that we call supervoxels, where each supervoxel c is a set of voxel indices. In order to enforce functional similarity, we require the voxels in the supervoxel satisfy a homogenity predicate. The formal definition of a set of supervoxels is given below.

Let the ith sample $\mathbf{x}_i = [x_{i,1} \quad x_{i,2} \quad \dots \quad x_{i,N_v}]$ be a row vector of voxel intensity values measured from all voxels during a single cognitive task for a single subject, where N_v is the number of voxels in an fMRI image for a particular subject. Each sample is the vectorized form of a brain volume that is captured at a time instance that correspond to a peak value after a stimulus presentation from an event related experiment (Objects, or Emotion datasets), or at any time instance that is in either "plan", or "execute" phases for the images in TOL dataset. Each sample $\mathbf{x}$ is then labeled with a class label $y \in \mathbb{L}$ where $\mathbb{L} = \{0, 1, \ldots, N_l - 1\}$, N_l being the number of class labels. The vector of class labels from all samples is denoted by $Y \in \mathbb{L}^{N_s}$

The design matrix $\mathbf{X} \in \mathbb{R}^{N_s \times N_v}$ is formed by using all samples $\{\mathbf{x}_i\}_{i=1}^{N_s}$ for a particular subject, where N_s is the number of all samples, and N_v is the number of voxels. Here, each sample $\mathbf{x}$ corresponds to a row of the design matrix $\mathbf{X}$. Conversely, each column of the design matrix, which is a vector of voxel intensity values $\boldsymbol{v}_j \in \mathbb{R}^{N_s}$ for all samples, where the jth voxel is denoted by $\boldsymbol{v}_j = [x_{1,j} \quad x_{2,j} \quad \ldots \quad x_{N_s,j}]^T$. Here, ^{T} is the transpose operator. We use the convention where $\mathbf{x}$ specify the samples (which is a row of the design matrix) and $\boldsymbol{v}$ specify the voxels (which are columns of the design matrix) for the sake of clarity in the notation that we use in the following sections.

Given a design matrix $\mathbf{X}$, we form training ($\mathbf{X}_{tr} \in \mathbb{R}^{N_{tr} \times N_v}$), validation ($\mathbf{X}_{val} \in \mathbb{R}^{N_{val} \times N_v}$), and test ($\mathbf{X}_{te} \in \mathbb{R}^{N_{te} \times N_v}$) matrices for the suggested brain decoding methododology, where $\mathbf{X}_{tr} \cup \mathbf{X}_{val} \cup \mathbf{X}_{te} = \mathbf{X}$, and $N_{tr} + N_{val} + N_{te} = N_s$. Here, N_{tr}, N_{val}, and N_{te} correspond to number of training, validation, and test samples while N_v correspond to the number of voxels. The set of labels correspond to these data matrices are denoted as $\mathbf{Y}_{tr} \in \mathbb{L}^{N_{tr}}$, $\mathbf{Y}_{val} \in \mathbb{L}^{N_{val}}$, and $\mathbf{Y}_{te} \in \mathbb{L}^{N_{te}}$. Train, validation, and test sets are split within subjects for Objects, Emotion, and TOL datasets as specified in Chapter 2.

3.1.2 Supervoxels

Let $\boldsymbol{v}_j \in \mathbb{R}^{N_s}$ be a vector of voxel intensity values of jth voxel across all samples from an fMRI experiment for a subject, where each $\boldsymbol{v}_j$ correspond to the jth column of the design matrix $\mathbf{X}$, and N_s is the number of samples. Furthermore let c be a supervoxel which is a set of voxel indices. Given a similarity predicate P, voxels indexed by j and j', where $j, j' \in J = \{1, 2, ..., N_v\}$, belong to the same supervoxel c such that $j \in c \wedge j' \in c$ if and only if:

$$P(\boldsymbol{v}_j, \boldsymbol{v}_{j'}) = TRUE \tag{3.1}$$

The similarity predicate for the supervoxels that are formed using AAL regions is that voxels indexed by j and j' being in the same brain region that is marked by AAL. Similarity predicates for supervoxels formed by the clustering algorithms use homogenity metrics and the algorithms themselves which are described in the following subsections.

For the set of all voxel indices $J = \{j\}_{j=1}^{N_v}$, a clustering algorithm forms a set of supervoxels C^θ using the clustering parameter $\theta \in \Theta = \{\theta_1, \theta_2, ..., \theta_{N_\theta}\}$, where θ is a pre-set number of clusters for K-Means and N-Cut clustering and Θ is a set of clustering parmeters θ. since we do not know which clustering parameter θ provide the most suitable brain parcellation for our purposes, we analyze the set of supervoxels generated by each parameter C^{θ} individually, as well as we analyze the set of all supervoxels with all clustering parameters $C = \bigcup_{\theta} C^{\theta}$ Thus, a set of supervoxels $C^{\theta} = \{c_1^{\theta}, c_2^{\theta}, ..., c_{N_c^{\theta}}^{\theta}\}$ is formed for each clustering parameter θ, where $c^{\theta} \subset J$, and for each $c^{\theta} \in C^{\theta}$. Here, N_c^{θ} is the number of supervoxels in the set of supervoxels C^{θ}.

For a clustering parameter θ, each supervoxel $c^{\theta} \in C^{\theta}$ is disjoint:

$$\bigcup_{c^{\theta} \neq c^{\theta\prime}}^{c^{\theta}, c^{\theta\prime} \in C^{\theta}} (c^{\theta} \cap c^{\theta\prime}) = \emptyset. \tag{3.2}$$

For each supervoxel c, training ($\mathbf{X}_{tr}^{c} \in \mathbb{R}^{N_{tr} \times N_v^c}$), validation ($\mathbf{X}_{val}^{c} \in \mathbb{R}^{N_{val} \times N_v^c}$), and test ($\mathbf{X}_{te}^{c} \in \mathbb{R}^{N_{te} \times N_v^c}$) matrices are formed using the voxel intensity values for the voxels indices $j \in c$ by horizontal concatenation of corresponding columns $\{\boldsymbol{v}_{tr,j}\}_{j \in c}$ from training ($\mathbf{X}_{tr}$), $\{\boldsymbol{v}_{val,j}\}_{j \in c}$ from validation ($\mathbf{X}_{val}$), and $\{\boldsymbol{v}_{te,j}\}_{j \in c}$ from test ($\mathbf{X}_{te}$) matrices. Here, N_v^c signifies the number of voxels in the supervoxel c while N_{tr}, N_{val}, N_{te} are the the number of samples in training, validation and test sets respectively.

In this study, various clustering techniques as well as AAL regions are used to form supervoxels. Since there are no a priori ways to know which clustering algorithm works best for our brain decoding methodology, we explored two different algorithms that have some essential differences in their approaches: Spatially constrained normalized cuts clustering (N-Cuts) [21], and K-Means clustering. Also, we use brain regions specified by AAL (Automated Anatomical Labelling) as supervoxels.

The essential difference between N-Cuts and K-Means algorithms is that K-Means algorithm only uses functional correlation between the voxels to form supervoxels, while N-Cuts algorithm proposed by Craddock et al. has a spatial constraint. The constraint is satisfied when every voxel within a particular supervoxel is in spatial proximity with at least one other voxel of the supervoxel. On the other hand, supervoxels specified by AAL serve as a baseline to compare the effectiveness of the other clustering algorithms.

3.1.3 Homogenity Metric

The critical design issue of all clustering algorithms is the selection of a similarity metric. For this purpose, we use Pearson correlation coefficient. Pearson correlation coefficient is a commonly used metric for fMRI image analysis [21]. The reason for heavy use of this metric for fMRI image analysis lies in the working principles of the human brain. Given a stimulus, or an experimental task, various regions of the human brain register the stimulus or engages in the processing of the task at different rates. For example, a visual stimulus progresses within the brain from Thalamus to early visual areas to Temporal and Parietal lobes in a sequential manner, and using multiple pathways in parallel. Therefore, regardless of the signal intensity, as long as the shape of the waveform of the activation of voxels are similar, the similarity conveys crucial information about both the topology of the voxel network and the function of the voxels. Since Pearson correlation measures linear correlation between two variables, the correlation coefficient between the voxel intensity values of two voxels for a series of scans can convey the information regarding how similar these two voxels behave with regards to experimental conditions.

For the purpose of clustering, we first obtain voxel intensity values $\boldsymbol{v}_{tr,j}$ across all the training samples for each voxel index $j \in J = \{1, 2, ..., N_v\}$, where N_v is the number of voxels in a brain volume. If we consider two voxel indices j and j', such that $\boldsymbol{v}_{tr,j}$ and $\boldsymbol{v}_{tr,j'} \in \mathbb{R}^{N_{tr}}$ are the jth and j'th columns of $\mathbf{X}_{tr}$, where they are the voxel intensity values of all samples for jth and j'th voxels in vector form, while N_{tr} being the number of training samples. The exact correspondents of training samples are explained in the previous subsection. Then, Pearson correlation $\rho_{j,j'}$ between these voxels activations is defined as:

$$\rho_{j,j'} = \frac{cov(\boldsymbol{v}_{tr,j}, \boldsymbol{v}_{tr,j'})}{\sqrt{var(\boldsymbol{v}_{tr,j}) \cdot var(\boldsymbol{v}_{tr,j'})}}, \tag{3.3}$$

where *cov* stands for covariance and *var* stands for variance between the random variables $\boldsymbol{v}_{tr,j}$, and $\boldsymbol{v}_{tr,j'}$.

3.1.4 Spatially Constrained Normalized Cuts Clustering (N-Cuts) for Brain Parcellation

The first clustering algorithm that we employ for brain parcellation is the spectral N-Cuts clustering algorithm suggested by [21]. This algorithm employs spatially constrained normalized cuts for clustering. In order to perform the clustering, first, a connectivity graph G is formed by using the voxel to voxel similarity matrix of all voxels $S = \{\rho_{j,j'}\} \in \mathbb{R}^{N_v \times N_v}$ by using (Equation 3.3) as the similarity metric. Then, the graph is further constrained by limiting the connectivity of each voxel to its 26 direct neighbors in the 3D voxel grid. That is to make sure make sure the voxels in a parcel are always spatially enclosed within a single region. Lastly, graph-cut is performed by using N-Cuts algorithm to obtain correlated and spatially connected voxel groups (sub-graphs of G). The clustering parameter for this algorithm specifies the number of sub-graphs. The N-Cuts algorithm stops when the specified number of seperate sub-graphs is reached.

Due to the spatial proximity constraint enforced above, any voxel in a cluster must be spatially connected to at least one other voxel in the cluster. This is the primary difference of this algorithm compared to K-Means clustering.

N-Cuts algorithm can be summarised as follows. Given a graph G, a graph cut algorithm that cuts the graph in two regions A and B can be formalized by the minimization of the sum of connection weights ($\rho_{j,j'}$) of the voxels (j, and j') that connects these two regions:

$$Cut(A, B) = \sum_{\boldsymbol{v}_j \in A, \boldsymbol{v}_{j'} \in B} \rho_{j,j'}. \tag{3.4}$$

However, a procedure that is applied repetitively to the graph that uses the above cost function ends up with isolated voxels in the end [21]. N-Cuts algorithm avoids this problem. Using this algorithm, for every cut region, the cut cost is normalized by using the sum of the connection weights of the voxels belong to that region between all other voxels in the graph. So the N-Cuts cost function is as follows:

$$J_{Cut}(A, B) = \frac{Cut(A, B)}{\sum_{v_j \in A, v_k \in G} \rho_{j,k}} + \frac{Cut(A, B)}{\sum_{v_{j'} \in B, v_k \in G} \rho_{j',k}}. \tag{3.5}$$

By minimizing the cost function (Equation 3.5), the graph G is cut so that a number of regions specified by the clustering parameter remains.

3.1.5 K-Means Clustering for Brain Parcellation

K-Means is a clustering algorithm that is used to iteratively determine the locations of a pre-specified number (where K comes from) of cluster means [61].

The algorithm starts with the initialization of $\theta = \{1, 2, ..., K\}$ cluster means $\{\bar{\boldsymbol{v}}_k\}_{k\in\theta}$ in random locations in the voxel space. The voxel space is formed by the vectors $\{\boldsymbol{v}_{j,tr}\}_{j\in J}$, of voxel intensity values from the samples that are in the training set, where , where $\boldsymbol{v}_{j,tr} \in \mathbb{R}^{N_{tr}}$ and $\bar{\boldsymbol{v}}_k \in \mathbb{R}^{N_{tr}}$. Here, $J = \{1, 2, ..., N_v\}$ is the set of all voxel indices, N_v is the number of voxels in a brain volume, N_{tr} is the number of training samples, while K is the number of clusters. One minus Pearson correlation $(1 - \rho_{j,k})$ (Equation 3.3) is used as the distance metric between the cluster means $\bar{\boldsymbol{v}}_k$ and the vectors of voxel intensity values $\boldsymbol{v}_{j,tr}$ that correspond to each voxel index j . The algorithm works in two phases. In the first phase, every voxel is assigned to the nearest cluster mean (thus forming supervoxels $c_k \subset J$), followed by the recalculation of the cluster means. After sufficient iterations, the second phase starts where each individual voxel is re-assigned, if doing so reduces the total distance of the voxels to the cluster means, and all the means are re-calculated for each re-assignment. This phase is repeated until the change in the sum of the distances (D_{Total}) of voxels to the nearest cluster means that they are assigned to are small enough, where:

$$D_{Total} = \sum_{k\in K} \sum_{j\in c_k} (1 - \rho_{j,k}) \tag{3.6}$$

In this study, the whole algorithm is run 500 times for different initializations of the means and the best result with minimum total distances are used in the further stages of BRE. We used the MATLAB implementation of this algorithm.

3.2 Stage 2: Fuzzy Stacked Generalization for Supervoxels

Once we partition the human brain into supervoxels, each of which consists of highly correlated voxels, the nexst step is to train a seperate classifier using the voxel activities within each supervoxel. This step enables us to investigate the role of each supervoxel which contribute to a pre-defined mental task or stimulus. While some voxels can be obselete, others may be crucial for the underlying mental task or stimulus. The suggested BRE technique use ensembles of the classifiers each of which are trained using a specific supervoxel. As mentioned previously, the aim here is to model the activity of specialized brain regions that represent or process a specific aspect of a given stimulus or mental task. We propose to capture the activity of specialized brain regions in terms of voxel activity in supervoxels. In order to capture the voxel activity and later on combine them in order to decode the mental state elicited by the mental task or stimulus, we use Fuzzy Stacked Generalization (FSG) algorithm.

Fuzzy Stacked Generalization (FSG) is an ensemble learning algorithm by Özay and Yarman-Vural [75]. In the original implementation, the algorithm is used to fuse the features spaces, each of which represents a specific attribute of the dataset where classification is performed. For instance, for an image dataset, visual attributes that are provided by MPEG-7 feature set are used such as, Color Structure, Color Layout, Edge Histogram, Region-Based Shape, Haar Filtering, Dominant Color etc. [27], where the aim is to map distinct aspects of a given image into different feature spaces. The main premise of FSG algorithm is to fuse those distinct attributes in a decision space where categories are easily seperated. When considering the human brain, we make the observation that the brain itself does such a mapping already by itself. Given a visual image, the spatial location of an object in the image is mapped at the parietal cortex [25], while shape change in an arbitrarily shaped object can be distinguished by the activities in lateral occipital cortex, and posterior intraparietal sulcus [12]. Likewise, texture change can be distinguished by posterior collateral sulcus, and color change can be distinguished by regions including lingual gyrus, fusiform gyrus, dorsolateral prefrontal cortex, and medial intraparietal sulcus [13, 12]. Thus, we suggest that, FSG can make use of the distinct representations in the brain that are mapped into different supervoxels, where each supervoxel is considered to form a distinct feature space in terms of the activities of the voxels contained within the supervoxel.

The algorithm uses a two layered architecture. In the base layer (first layer), there are multiple classifiers, each of which receive input from a single supervoxel in terms of activities of the voxels contained within. For this purpose, the activities of each voxel within the supervoxel is concatenated in vector form and fed to a base layer classifier. The output of a base layer classifier is the class posteriori probabilities of a given sample for each class label. Here, a class posteriori probability for a base layer classifier is defined to be the likelihood of a stimulus category given the activity values of the voxel within the supervoxel. In the meta layer, (second layer) all of the outputs of the base layer classifiers in terms of class posteriori probabilities are concatenated and fed to a meta classifier.

From the neuroscientific perspective, output of a base layer classifier in terms of class posteriori probabilities is a model for the activity of a brain region that contributes to a mental state. Depending on the specific stimulus or the mental task, activity of some of the brain regions can be decisive for the recognition of the stimulus or the processing of the mental task. For instance, when viewing a field of grass, the texture and the color of the field is critical for the recognition of the viewed stimulus as a field of grass. Likewise, for the example of the field of grass, representation of a definite shape is lacking. When we consider the brain regions specialized for color, texture, and shape processing, we claim that brain regions that process color and texture can highly contribute to the decision that the viewed stimulus is a field of grass, while the lack of activity in the brain regions that process shape also contributes to the final decision. From this perspective, the class posteriori probabilities of base layer classifiers model the activity of specialized brain regions with regards to their activity given a set of stimulus or mental tasks that are present in an fMRI experiment.

In the original implementation of FSG, a base layer classifier is trained for each feature type and a single meta layer classifier is trained for the final classification. In BRE however, we use the outputs of the base layer classifiers that are trained for a subset of supervoxels (of the set of all supervoxels) and a meta classfier is trained for that subset. Then, we combine the outputs of the meta classifiers for all such subsets of supervoxels in order to decide the class label of a given input. The details of the inputs and outputs as well as the operational characteristics layer classifiers are explained in the next couple of subsections (Subsections 3.2.1 and 3.2.2). The details of meta classifiers are explained in Subsection 3.2.3, after we define the notion of subsets of supervoxels.

3.2.1 Logistic Regression

In the original study of Özay and Yarman-Vural [75], FSG algorithm used k-nearest neighbor (KNN) classifiers for base layer classifiers in order to obtain class posteriori probabilities for each sample. However, we observed that a linear classifier is much better suited for the classification of fMRI images due to the sparsity of the feature spaces where KNN's performance is significantly inferior with respect to a linear classifier. Similar results have been obtained in other comparative studies [55]. Hence, we use logistic regression classifier (presented in the next subsection) for the purposes of base layer classification task. Logistic regression is particularly suitable for our purposes since, while being a linear classifier, it naturally outputs class posteriori probabilities for each class, and with softmax activation function, it can easily be adapted to multi-class classification. While a case can be made for SVM classifier, its performance is not significantly better than the logistic regression classifier [55, 18], and obtaining class posteriori probabilities is takes an additional step of processing [77] which we prefer to avoid due to its computational cost.In this subsection we describe how a logistic regression classifier works, and how it provides class posteriori probabilities.

Logistic regression is a linear classifier that can be visualized as a two-layer neural network. In the input layer, there are as many units as the input feature dimension. In the output layer there is a single output unit for binary classification, and a number of output units that is equal to number of classes for multi-class classification. All input units are connected to all output units.

Since a base layer classifier is trained for each supervoxel c, the input to the logistic regression classifer is the voxel intensity values from the set of voxel indices that belong to the supervoxel c. Thus, a training input to the algorithm is denoted as $\mathbf{x}^c_{tr} \in \mathbb{R}^{N^c_v}$, where N^c_v is the number of voxels in the supervoxel. Each training sample $\mathbf{x}^c_{tr}$ corresponds to a row of the training matrix $\mathbf{X}^c_{tr}$, while a test sample $\mathbf{x}^c_{te}$ correspond to a row of the test matrix $\mathbf{X}^c_{te}$ which are constructed for the supervoxel c (see Section 3.1.2). Here, N^c_v is the number of voxels within the supervoxel c.

3.2.1.1 Binary classification

For binary classification, let us denote a sample as $\mathbf{x}^c \in \mathbb{R}^{N^c_v}$, which is a row vector of activity values of voxels that are in the supervoxel c. A sample $\mathbf{x}^c$ can either be a training sample $\mathbf{x}^c = \mathbf{x}^c_{tr}$ that is a row of training matrix $\mathbf{X}^c_{tr}$, a validation sample $\mathbf{x}^c = \mathbf{x}^c_{val}$ that is a row of validation matrix $\mathbf{X}^c_{val}$, or a test sample $\mathbf{x}^c = \mathbf{x}^c_{te}$ that is a row of the test matrix $\mathbf{X}^c_{te}$. The activation of the output unit $\tilde{y}$ is calculated by using sigmoid (sgn) function as follows:

$$\tilde{y} = sgn(\mathbf{x}^c\mathbf{w}^c + b^c) = \frac{1}{1 + e^{-(\mathbf{x}^c\mathbf{w}^c + b^c)}}, \tag{3.7}$$

where $\tilde{y} \in (0 \leq \mathbb{R} \leq 1)$ is an estimate of the posterior probability $p(y = 1|\mathbf{x}^c)$, for the true class label $y \in \mathbb{L} = \{0, 1\}$. Here, $\mathbf{w}^c \in \mathbb{R}^{N^c_v}$ is weight vector and b^c is the bias parameter that allows decision boundaries that do not pass the origin of the feature space. Using $\tilde{y}$, the estimated class label $\hat{y} \in \mathbb{L} = \{0, 1\}$ is:

$$\hat{y} = \begin{cases} 1 & \text{if } \tilde{y} \geq 0.5 \\ 0 & \text{if } \tilde{y} < 0.5 \end{cases}. \tag{3.8}$$

Given training samples $\mathbf{x}^c_{tr}$ that are rows of $\mathbf{X}^c_{tr}$, and their respective labels $y_{tr} \in \mathbf{Y}_{tr}$, the connection weights and the bias term are trained by minimization of the following cost function:

$$J_{binary} = \phi\frac{1}{2}(\mathbf{w}^{cT}\mathbf{w}^c) - \frac{1}{N_{tr}} \sum_{y_{tr} \in \mathbf{Y}_{tr}}^{N_{tr}} (y_{tr} log(\tilde{y}_{tr}) + (1 - y_{tr}) log(1 - \tilde{y}_{tr})), \tag{3.9}$$

where ^{T} is the transpose operator.

The first part of the cost function (Equation 3.9) is for the regularization of the connection weights, where ϕ is the regularization parameter that becomes important if the classifier over-fits the training data, where increasing it increases the weight of the first term that effectively reduces the dimensionality of the feature space by making w sparse. The second part of the cost function is the cross-entropy cost [23]. The cost function has many desirable properties of a quadratic cost function such that it is non-negative, and the cost approaches to zero when $\tilde{y}$ approaches to y. Additionally, it prevents the slow-down in learning while using sigmoid activation when $|\mathbf{x}^c\mathbf{w}^c+b|$ becomes large. The cost function can be minimized by using gradient descent over the parameters $\mathbf{w}$ and b. Other optimization techniques for a logistic regression classifier can be found in [33].

3.2.1.2 Multi-class Classification

For multi-class classification, let us denote a sample as $\mathbf{x}^c \in \mathbb{R}^{N_v^c}$, which is a row vector of activity values of voxels that are in the supervoxel c. A sample $\mathbf{x}^c$ can either be a training sample $\mathbf{x}^c = \mathbf{x}^c_{tr}$ that is a row of training matrix $\mathbf{X}^c_{tr}$, a validation sample $\mathbf{x}^c = \mathbf{x}^c_{val}$ that is a row of validation matrix $\mathbf{X}^c_{val}$, or a test sample $\mathbf{x}^c = \mathbf{x}^c_{te}$ that is a row of the test matrix $\mathbf{X}^c_{te}$. For the classification of multiple classes, we use softmax at the output layer instead of sigmoid activation. In this case, there are a multitude of outputs where the number of them is equal to the number of classes (N_l). Given a sample $\mathbf{x}^c \in \mathbb{R}^{N_v^c}$, the activation of the outputs $\tilde{y}_l \in (0 \leq \mathbb{R} \leq 1)$, $l \in \{0, 1, ..., N_l - 1\}$, are calculated by using their respective weights $\mathbf{w}^c_l \in \mathbb{R}^{N^c}$ and biases $b^c_l \in \mathbb{R}$. Here, N_l denotes the number of class labels. If we define $z_l = \mathbf{x}^c\mathbf{w}^c_l + b^c_l$, then the softmax activation for each $\tilde{y}_l$ becomes:

$$\tilde{y}_l = \frac{e^{z_l}}{\sum_{i=0}^{N_l-1} e^{z_i}}, \tag{3.10}$$

where each $\tilde{y}_l$ is an estimate of the posteriori probability $p(y = l|\mathbf{x}^c)$, and $y \in \mathbb{L} = \{1, 2, ..., N_l - 1\}$ is the true class label for the given sample x^c. The predicted class label $\hat{y} \in \{0, 1, ..., N_l - 1\}$ then becomes:

$$\hat{y} = \underset{l}{\operatorname{argmax}} \{\tilde{y}_l\}_{l=0}^{N_l-1}. \tag{3.11}$$

Also, the output, that is the class posteriori probabilities $\tilde{\mathbf{y}} = \{\tilde{y}_l\}_{l=0}^{N_l-1}$ is an N_l dimensional vector.

Given the training features $\mathbf{x}^c_{tr}$ and their respective true class labels $y_{tr} \in \{0, 1, ..., N_l - 1\}$, the expected output $y_{tr,l}$ of the output unit l is defined as follows:

$$y_{tr,l} = \begin{cases} 1 & \text{if } y_{tr} = l \\ 0 & \text{else} \end{cases}. \tag{3.12}$$

Cost function with cross-entropy is defined as follows:

$$J_{multi} = \phi \frac{1}{2N_l} \sum_{l=0}^{N_l-1} (\mathbf{w}_l^{cT} \mathbf{w}_l^c) - \frac{1}{N_l N_{tr}} \sum_{m=1}^{N_{tr}} \sum_{l=0}^{N_l-1} y_{tr,l} log(\tilde{y}_{tr,l}), \tag{3.13}$$

where $\tilde{y}_{tr,l}$ is the posteriori probability $p(y_{tr} = l|\mathbf{x}_{tr}^c)$.

3.2.1.3 Classifier outputs

Given a matrix of voxel activity values $\mathbf{X}_{te}^c \in \mathbb{R}^{N_{te} \times N_v^c}$ of the test set for the supervoel c, the primary output of a logistic regression classifier is a matrix of class probability estimations $\widetilde{\mathbf{Y}}_{te}^c \in \mathbb{R}^{N_{te} \times N_l}$, and a vector of predicted class labels $\widehat{\mathbf{Y}}_{te}^c \in \mathbb{L}^{N_{te}}$. For the binary classification, a row of $\widetilde{\mathbf{Y}}_{te}^c$ consists of two values $\tilde{y}_{te}$ and $1 - \tilde{y}_{te}$ that correspond to the input sample $\mathbf{x}_{te}^c$ that is a row of $\mathbf{X}_{te}^c$. For the multi-class classification, a row of $\widetilde{\mathbf{Y}}_{te}^c$ consists of N_l values each of which correspond to the class posteriori probability $\tilde{y}_{te,l}$ for the class l given the sample $\mathbf{x}_{te}^c$, where N_l is the number of classes.

Another output of the logistic regression classifiers is their percentile accuracy (Acc). Specifically, for the test samples x_{te}^c that are rows of X_{te}^c, where, using the predicted class labels $\hat{y}_{te} \in \{0, 1, ..., N_l - 1\}$, and the true labels $y_{te} \in \{0, 1, ..., N_l - 1\}$, the percentile accuracy (Acc_{te}^c) of the classifier for N_{te} test samples is calculated as follows:

$$Acc_{te} = 100 \sum_{i=1}^{N_{te}} \frac{\delta(y_{te,i}, \hat{y}_{te,i})}{N_{te}}, \tag{3.14}$$

where $\delta(\cdot)$ is the Kronecker delta function which is defined as follows:

$$\delta(y_{te}, \hat{y}_{te}) = \begin{cases} 1 & \text{if } y_{te} = \hat{y}_{te} \\ 0 & \text{if } y_{te} \neq \hat{y}_{te} \end{cases}. \tag{3.15}$$

In a similar fashion, the percentile accuracy for the validation samples (Acc_{val}^c) can be calculated by using validation samples $\mathbf{x}_{val}^c$ which are the rows of the validation matrix $\mathbf{X}_{val}^c \in \mathbb{R}^{N_{val} \times N_v^c}$, using the above methodology.

To sum up, a logistic regression classifier is trained by using a training set of samples $\mathbf{X}_{tr}^c$, their class labels $\mathbf{Y}_{tr}^c$ and a regularization parameter ϕ and forms a classifier $model^c = (\mathbf{w}^c, b^c, \phi)$ for the supervoxel c. After the training, for the given set of samples ($\mathbf{X}_{tr}^c$, $\mathbf{X}_{val}^c$, or $\mathbf{X}_{te}^c$) the classifier $model^c$ can be used to predict the class labels ($\widehat{\mathbf{Y}}_{tr}^c$, $\widehat{\mathbf{Y}}_{val}^c$, and $\widehat{\mathbf{Y}}_{te}^c$), class posteriori probabilities ($\widetilde{\mathbf{Y}}_{tr}^c$, $\widetilde{\mathbf{Y}}_{val}^c$, and $\widetilde{\mathbf{Y}}_{te}^c$), and percentile accuracies if class labels are also provided for the validation and test sets.

3.2.2 Base layer classifiers

In order to train a logistic regression model for a supervoxel c, sets of feature matrices for training ($\mathbf{X}^c_{tr}$), and validation ($\mathbf{X}^c_{val}$), vectors of their respective class labels ($\mathbf{Y}_{tr}$, and $\mathbf{Y}_{val}$), and a set of regularization parameters $\phi \in \Theta$ are used. The best regularization parameter ϕ_{best} is selected by using the validation set. Using the model, the classifier then returns class posteriori probabilities ($\widetilde{\mathbf{Y}}^c_{val}$, and $\widetilde{\mathbf{Y}}^c_{te}$), estimated class labels ($\widehat{\mathbf{Y}}^c_{val}$, and $\widehat{\mathbf{Y}}^c_{te}$), and accuracies ($Acc^c_{val}$, Acc^c_{te}) for the training and test sets (Algorithm 3.1, lines 2-9).

The class posteriori probabilities ($\widetilde{\mathbf{Y}}^c_{tr}$), for the training set is then obtained by using the FSG algorithm (3.1, lines 10-16) where for each sample $\mathbf{x}_m$ that is a row of the training matrix $\mathbf{X}^c_{tr}$, a new classifier is trained. In this part of the algorithm, first, a sample of the training set $\mathbf{x}_i$ is set aside. Second, a logistic regression classifier is trained for the rest of the training samples and a classifier model $model_i$ is obtained. Third, for the sample that is set aside, class posteriori probabilities ($\tilde{\mathbf{y}}_i$) for that sample is estimated using the model ($model_i$). Lastly, a matrix of class posteriori probabilities ($\widetilde{\mathbf{Y}}^c_{tr}$) for the training samples are formed by assigning the class posteriori probabilities of each sample to the rows.

Algorithm 3.1 Base layer classification

```
1:  procedure BASE($\mathbf{X}^c_{tr}, \mathbf{X}^c_{val}, \mathbf{X}^c_{te}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te}$)
2:      for all $\phi \in \Phi$ do                                   ▷ $\phi$ is a regularization parameter
3:          $model \leftarrow LRtrain(\mathbf{X}^c_{tr}, \mathbf{Y}_{tr}, \phi)$
4:          $Acc^{\phi} \leftarrow LRpredict(model, \mathbf{X}^c_{val}, \mathbf{Y}_{val})$
5:      end for
6:      $\phi_{best} \leftarrow \underset{\phi}{\mathrm{argmax}}(\{Acc^{\phi}\})$
7:      $model = LRtrain(\mathbf{X}^c_{tr}, \mathbf{Y}_{tr}, \phi_{best})$
8:      $(\widehat{\mathbf{Y}}^c_{val}, \widetilde{\mathbf{Y}}^c_{val}, Acc^c_{val}) \leftarrow LRpredict(model, \mathbf{X}^c_{val}, \mathbf{Y}_{val})$
9:      $(\widehat{\mathbf{Y}}^c_{te}, \widetilde{\mathbf{Y}}^c_{te}, Acc^c_{te}) \leftarrow LRpredict(model, \mathbf{X}^c_{te}, \mathbf{Y}_{te})$
10:     for all $\mathbf{x}_i \in \mathbf{X}^c_{tr}$ do                      ▷ $i$ indexes a training sample
11:         $\mathbf{X}^c_{tr,i} \leftarrow \mathbf{X}^c_{tr}[i] = []$                      ▷ delete $i$'th row from $\mathbf{X}^c_{tr}$
12:         $\mathbf{Y}_{tr,i} \leftarrow \mathbf{Y}_{tr}[i] = []$                      ▷ delete $i$'th row from $\mathbf{Y}_{tr}$
13:         $model_i \leftarrow LRtrain(\mathbf{X}^c_{tr,i}, \mathbf{Y}_{tr,i}, \phi_{best})$
14:         $\tilde{\mathbf{y}}_i \leftarrow LRpredict(model_i, \mathbf{x}_i)$
15:         $\widetilde{\mathbf{Y}}^c_{tr}[i] \leftarrow \tilde{\mathbf{y}}_i$                             ▷ assign $i$'th row of $\widetilde{\mathbf{Y}}^c_{tr}$
16:     end for
17:     return $\widetilde{\mathbf{Y}}^c_{tr}, \widetilde{\mathbf{Y}}^c_{val}, \widetilde{\mathbf{Y}}^c_{te}, \widehat{\mathbf{Y}}^c_{val}, \widehat{\mathbf{Y}}^c_{te}, Acc^c_{val}, Acc^c_{te}$
18: end procedure
```

In the following section we describe a meta classifier, which is the component for BRE to fuse distinct representations of brain activity from supervoxels. Each meta classifier operates on a subspace of supervoxels and fuses the outputs of the base layer classifiers from each supervoxel.

3.2.3 Meta classifier

In the FSG architecture, the meta classifier is the one that is responsible for the fusion of contribution of each supervoxel. In the original article, there is a single meta classifier that fuses the posterior probabilities from all feature spaces [75]. However, in our framework we extend the idea of fusion. We utilize a multitude of meta classifiers each of which operates on a subset C^{ψ} of all possible supervoxels $C = \bigcup_{\theta \in \Theta} C^{\theta}$ for a given partitioning of a clustering algorithm. Here, θ denotes a clustering parameter in the set of all clustering parameters Θ, and C^{θ} denotes a set of supervoxels that are generated by the clustering parameter θ. The subset of supervoxels C^{ψ} can be a random subset of supervoxels (a number of supervoxels that are selected randomly within C), or it can be the set of all supervoxels that are formed by a specific clustering parameter θ, which makes $C^{\psi} = C^{\theta}$. We explain how and why we form these subsets of supervoxels in the following subsections.

Given a subset of supervoxels $C^{\psi} = \{c_1, c_2, ..., c_{N_v^{\psi}}\}$, and the base layer classifier outputs $\widetilde{\mathbf{Y}}^{c}_{tr}, \widetilde{\mathbf{Y}}^{c}_{val}, \widetilde{\mathbf{Y}}^{c}_{te}$ of each supervoxel $c \in C^{\psi}$, the training matrix $\mathbf{X}^{\psi}_{tr} \in \mathbb{R}^{N_{tr} \times N_l N_v^{\psi}}$ is formed by column-wise concatenation of the class posteriori probabilities $\widetilde{\mathbf{Y}}^{c}_{tr}$:

$$\mathbf{X}^{\psi}_{tr} = \begin{bmatrix} \widetilde{\mathbf{Y}}^{c_1}_{tr} & \widetilde{\mathbf{Y}}^{c_2}_{tr} & \dots & \widetilde{\mathbf{Y}}^{c_{N_v^{\psi}}}_{tr} \end{bmatrix} \tag{3.16}$$

Here, N_v^{ψ} denotes the number of supervoxels in C^{ψ}, N_{tr} is the number of training samples, and N_l is the number of classes. Similarly, validation $\mathbf{X}^{\psi}_{val}$, and testing $\mathbf{X}^{\psi}_{te}$ feature matrices are formed using their respective set of class posteriori probability outputs:

$$\mathbf{X}^{\psi}_{val} = \begin{bmatrix} \widetilde{\mathbf{Y}}^{c_1}_{val} & \widetilde{\mathbf{Y}}^{c_2}_{val} & \dots & \widetilde{\mathbf{Y}}^{c_{N_v^{\psi}}}_{val} \end{bmatrix} \tag{3.17}$$

$$\mathbf{X}^{\psi}_{te} = \begin{bmatrix} \widetilde{\mathbf{Y}}^{c_1}_{te} & \widetilde{\mathbf{Y}}^{c_2}_{te} & \dots & \widetilde{\mathbf{Y}}^{c_{N_v^{\psi}}}_{te} \end{bmatrix} \tag{3.18}$$

At this stage, a support-vector machine (SVM) with second order regularization (please refer to Section 3.2.1.1 for an example of this regularization technique) is used as the classifier in order to comply with voxel selection based MVPA algorithms. Since it is a commonly used algorithm, the implementation details of SVM is not given here. We used LIBSVM implementation of this algorithm [15].

The classifier outputs the predictions for validation $\widehat{\mathbf{Y}}^{\psi}_{val}$, and $\widehat{\mathbf{Y}}^{\psi}_{te}$ test sets as well as their respective percent accuracies Acc^{ψ}_{val}, Acc^{ψ}_{te} (Algorithm 3.2).

Algorithm 3.2 Meta classification

```
procedure META($\mathbf{X}^{\psi}_{tr}, \mathbf{X}^{\psi}_{val}, \mathbf{X}^{\psi}_{te}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te}$)
    for all $\phi \in \Phi$ do                                   ▷ $\phi$ is a regularization parameter
        $model = SVMtrain(\mathbf{X}^{\psi}_{tr}, \mathbf{Y}_{tr}, \phi)$
        $Acc^{\phi} \leftarrow SVMpredict(model, \mathbf{X}^{\psi}_{val}, \mathbf{Y}_{val})$
    end for
    $\phi_{best} \leftarrow \underset{\phi}{\text{argmax}}(\{Acc^{\phi}\})$
    $model = train(\mathbf{X}^{\psi}_{tr}, \mathbf{Y}_{tr}, \phi_{best})$
    $(\widehat{\mathbf{Y}}^{\psi}_{val}, Acc^{\psi}_{val}) \leftarrow SVMpredict(model, \mathbf{X}^{\psi}_{val}, \mathbf{Y}_{val})$
    $(\widehat{\mathbf{Y}}^{\psi}_{te}, Acc^{\psi}_{te}) \leftarrow SVMpredict(model, \mathbf{X}^{\psi}_{te}, \mathbf{Y}_{te})$
    return $\widehat{\mathbf{Y}}^{\psi}_{val}, \widehat{\mathbf{Y}}^{\psi}_{te}, Acc^{\psi}_{val}, Acc^{\psi}_{te}$
end procedure
```

3.3 Stage 3: Subsets of Supervoxels for Meta Classifiers

Mental representations in the brain have multiple aspects. For example, color, shape, or texture in the visual stimuli for the representation of a visual object, or the auditory and visual stimuli and the recollection of specific memories that act together to induce fear. In order to capture and model these aspects, as the first step of BRE, we partitioned the 3-dimensional brain volume into "homogenous" voxel groups with respect to a similarity predicate. Each homogenous region, called supervoxel, is assumed to participate a specific aspect of a mental representation. Therefore, it is assumed that each mental state is represented by an ensemble of supervoxels.

Based on the above assumption, we train a base layer classifier for each supervoxel to capture the degree of participation of that particular supervoxel to the set of mental states that are specified by the fMRI experiment. At this point, what matters is to find a way to find the supervoxels that contribute to the processing or representation of the mental states that are induced by the fMRI experiments. If we define the set of all supervoxels that are generated by every clustering parameter $C = \bigcup_{\theta \in \Theta} C^{\theta}$ as the result of the brain parcellation procedures, where C^{θ} is the set of supervoxels that are generated using the clustering parameter θ, the problem reduces to finding a specific subset of supervoxels $C^{\psi} \subseteq C$ that contribute to the processing or representation of the given mental tasks or stimuli. Given such a subset of supervoxels, a meta classifier can be trained using FSG in order to combine the mental representations encoded in terms of voxel activity within the supervoxels.

When confronted with the problem of finding the right composition of supervoxels, the first solution that comes to mind is to find the supervoxels, which are expected to contribute to underlying mental processes, through a measure for selecting supervoxels. For this purpose, we have used the classification accuracy results of the base layer classifiers for each supervoxel. However, when we form a subset of supervoxels by eliminating the supervoxels for which the base layer classifier accuracy is below a certain threshold, where we experimented with an array of such thresholds, did not yield to higher classification accuracies for the meta classifiers that are trained with them than the meta layer classifier that used all supervoxels.

In order to deal with the above stated problem, we first present a naiive method, where we use a subset of supervoxels that are generated with a brain partition specified by a clustering parameter for training a meta classifier. Second, we present the primary method that we use to build BRE, where we select random subsets of supervoxels from the set of all supervoxels to train meta classifiers.

3.3.1 Subsets Generated with Specific Parameters

The first approach we use to select a subset of supervoxels is similar to the original version of FSG, where they used the outputs of the base layer classifiers that cover all available features to train a single meta classifer. For that purpose, we combine the supervoxels that is the result of a brain partitioning process, through the use of a single θ parameter to train a meta classifier. In which case, we have $C^{\psi} = C^{\theta}$, and we train a single meta classifier that uses the base layer classifier outputs from every supervoxel that are generated with a specific clustering parameter. With this approach, we leave problem of finding the right combination of supervoxels to the meta classifier, while we have control over the number of partitions by selecting the clustering parameter θ. We can, then, find the right number of partitions by trying out a set of clustering parameters Θ and selecting the parameter $\theta_{best} \in \Theta$ that result in a meta classifier that have the highest accuracy in the validation set. Similarly, brain partitioning of AAL can be used for this purpose, where a meta classifier can be trained for the combination of supervoxels that are specified by anatomical regions specified by AAL labeling, which makes $C^{\psi} = C^{AAL}$, where C^{AAL} is the set of supervoxels formed by AAL.

One problem with this approach is about the dimensionality of the inputs with which the meta classifier is trained. As we mentioned in the previous section, we use the matrix of class posteriori probabilities $\mathbf{X}_{tr}^{\psi}$ for the training of a meta classifier. The dimensionality of the matrix is $N_{tr} \times N_l N_{\psi}$, where N_l is the number of class labels and N_{ψ} is the number of supervoxels in the subset C^{ψ}. When we select the subset of supervoxels to be the one generated from the whole brain with a clustering parameter θ, C^{ψ} becomes equal to C^{θ} and $N_{\psi} = N_{\theta}$. With that in mind, as the number of supervoxels that is tied to the θ parameter increases, the dimensionality of the input space $(N_l N_{\psi})$ increases, where problem of overfitting would become prominent. In order to deal with this problem, in our preliminary studies, we applied a stage of feature selection to the inputs of the meta classifiers. Note that, this approach is different than the supervoxel elimination procedure in the sense that, with this approach, output of a base layer classifier not necessarily eliminated totally, no matter what the accuracy rating of that classifier is. While the results of this approach were promising [5], we later abandoned it due to the computational complexity introduced by the feature selection stage.

In this study, we present the classification results using this approach, without the feature elimination stage, in order to investigate the effect of the clustering parameters and methods (thus, the size and the composition of the supervoxels) to the classification accuracy of a meta classifier.

3.3.2 Random Subsets of Supervoxels

As we have mentioned previously, composition of the subset of supervoxels $C^{\psi} \subseteq C$, where $C = \bigcup_{\theta \in \Theta} C^{\theta}$ being the set of all supervoxels formed by the partitioning processes that use every clustering parameter θ, is critical for decoding mental states using the mental representations captured by the base layer classifiers that act on the supervoxels. However, finding the right combination of supervoxels to form a subset for meta classification is not an easy task. Furthermore, depending on the mental states under consideration for each fMRI experiment, the composition of such subsets are expected to change. Hence, instead of trying to find such a subset, we propose a method that relies on random sampling of such subsets, where a meta classifier is trained for each subset and their results are aggregated. This method, which we use to build brain region ensembles is called random subsets of supervoxels (RSS).

The methodology for forming RSS is as follows. Given a set of supervoxels C^{Ψ}, N_{ψ} supervoxels are randomly selected within C^{Ψ} without replacement, where N_{ψ} is proportional to the cardinality of C^{Ψ}. This process is performed N_t times, each time forming a subset $C^{\psi} \in C^{\Psi}$.

The method that we propose here, is based on the random subspace ensembles [7] of pattern classifiers, which provides promising results with respect to classification accuracy on the datasets where number of sampes is low and to feature dimensionality is high [38, 86, 98]. Given those properties, the application of random subspace ensembles to the field of brain decoding is not new. Kuncheva et. al have shown that ensembles of SVM classifiers each of which use a random subspace of voxels achieve higher classification accuracies when compared to other ensemble learning methods [55] and single classifiers [56]. Also, random subspace ensembles are used in real time classification of fMRI data [78]. In these studies the authors form the random subspaces within a set of voxels that are selected using voxel selection algorithms such as voxel selection with ANOVA, or voxel selection with SVM (see Section 2.4.2).

In their study regarding the parameter selection of random subspaces for fMRI analysis, Kuncheva et. al postulates that accurate and diverse set of classifiers would make a good classifier ensemble [56]. Therefore, when we think in terms of supervoxels sampled by RSS, it is preferrable to each subset contain at least one important feature which is the output of a base layer classifier that uses a supervoxel that is critical in the processing of the set of stimuli introduced by the fMRI experiment. Furthermore, we would not want every one of such critical supervoxels to be in more than one particular subset, where they would result in redundant classifiers.

In order to come up with the ideal parameters for the number of subspaces and the number of voxels in each subspace with respect to the number of critical voxels, Kuncheva et. al [56] has run a simulation studies that reflect the properties of an actual fMRI dataset. While they did not come up with an ideal set of parameters, their results shown that the number of voxels in each subspace should be high, about to be the half of the number of voxels in the original voxel space. Also, they suggested a low number of subspaces to be generated, which is around 100 of such subspaces. In this study we follow these guidelines in the way we form RSS, where to form each subset of supervoxels C^{ψ}, half the number of supervoxels in the superset C^{Ψ} is selected, where the number of random subsets $N_t = 100$.

Please note that for each one of the subsets, a meta classifier of FSG is trained. In the next section, we describe how we use RSS for building various forms of Brain Region Ensembles.

3.4 Stage 4: Brain Region Ensembles for Brain Decoding

The primary goal of the suggested BRE model is to capture the information distributed across the brain to decode cognitive states from fMRI images. We claim to achieve this goal through the use of supervoxels, in other words, functionally homogenous voxel groups. With FSG algorithm, we build a base layer classifier for each supervoxel, which receives input in the form of the voxel activity values that belong to the supervoxel, and outputs the class posteriori probabilities for each input sample with respect to the stimulus categories that are present in the fMRI experiment. We claim that the class posteriori probabilities that are output from each base layer classifier captures a distinct aspect of a given stimulus or mental task such as color, shape, texture for a visual stimulus or the recollection of a specific memory for stimulus that is expected to elicit emotional responses. However, in this study, we do not try to determine the exact nature of the supervoxels themselves in terms of which aspect of the given stimulus that they represent. Rather than that, we fuse the representations generated by them in order to decode the underlying mental state.

The main strength of our model is that it fuses the information encoded in the supervoxels using meta classifiers. The resulting model do not operate under any assumptions regarding the location, size or the composition of the contributing supervoxels. Still, our model can decode the contributions of individual supervoxels, relative to the rest of the brain volume, to the classification tasks.

The formal definition of a brain region ensemble is given in the following subsection.

3.4.1 Formal definiton of a Brain Region Ensemble

Brain region ensembles (BRE) is a model for brain decoding that utilizes the information captured by a set of supervoxels. A BRE is the building block for such ensembles that operates on a specific subset of supervoxels. A set of supervoxels C can be formed by either through the partitioning of the brain volume that is composed of voxels, or through the use of a functional brain atlas, such as AAL, where each supervoxel is a collection of voxels. The methods that we use for the brain partitioning process is given in Section 3.1.

A brain region ensemble is formed for a given subset of supervoxels $C^{\psi} \subseteq C$. A BRE receives inputs in the form of the voxel activity values from each supervoxel $c \in C^{\psi}$ within the subset of supervoxels C^{ψ} it operates on. Using those voxel activity falues, first, a base layer FSG classifier is trained for each supervoxel $c \in C^{\psi}$ using Algorithm (3.1) described in Section 3.2.2. Then, a meta classifier is trained using the class posteriori probabilities that are output from the base layer classifiers using Algorithm 3.2 that is described in Section 3.2.3. As the output, a BRE provides estimated class labels for a set of test samples, using the trained meta classifier.

Formally speaking, the inputs to the base layer classifier for the supervoxel c are the feature matrices $\mathbf{X}^c_{tr}$, $\mathbf{X}^c_{val}$, $\mathbf{X}^c_{te}$, and the class labels for the training samples $\mathbf{Y}_{tr}$. The class labels of the validation samples $\mathbf{Y}_{val}$ are used for the parameter optimization for the base layer classifiers. The class labels of the test set $\mathbf{Y}_{te}$ are only used to obtain an accuracy value from the base layer classifier for the test samples. At the outputs of the base layer classifier, class posteriori probabilities $\widetilde{\mathbf{Y}}^c_{tr}$, $\widetilde{\mathbf{Y}}^c_{val}$, and $\widetilde{\mathbf{Y}}^c_{te}$ are obtained, where each of which stands for the respective posteriori probability outputs for the training, validation, and test samples using the supervoxel $c \in C^{\psi}$. This procedure is applied to all such supervoxels c.

After the class posteriori probabilities are collected for every supervoxel $c \in C^{\psi}$, they are concatenated across the supervoxels using the Equations 3.16, 3.17, and 3.18 , thus, forming the inputs to the meta classifier ($\mathbf{X}^{\psi}_{tr}$, $\mathbf{X}^{\psi}_{val}$, and $\mathbf{X}^{\psi}_{te}$) that is specifically trained for the subset of supervoxels C^{ψ}. The meta classfier outputs the predictions for the validation and test sets $\widehat{\mathbf{Y}}^{\psi}_{val}$, $\widehat{\mathbf{Y}}^{\psi}_{te}$, as well as the accuracy values for them $Acc^{\psi}_{val}, Acc^{\psi}_{te}$.

Algorithm 3.3 A Brain Region Ensemble

```
procedure BRE($\{\mathbf{X}^c_{tr}\}_{c\in C^{\psi}}, \{\mathbf{X}^c_{val}\}_{c\in C^{\psi}}, \{\mathbf{X}^c_{te}\}_{c\in C^{\psi}}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te}$)
    for all $c \in C^{\psi}$ do
        $\widetilde{\mathbf{Y}}^c_{tr}, \widetilde{\mathbf{Y}}^c_{val}, \widetilde{\mathbf{Y}}^c_{te} \leftarrow BASE(\mathbf{X}^c_{tr}, \mathbf{X}^c_{val}, \mathbf{X}^c_{te}, \mathbf{Y}_{tr}, \mathbf{Y}_{val})$
    end for
    $\mathbf{X}^{\psi}_{tr} \leftarrow concatenate(\{\widetilde{\mathbf{Y}}^c_{tr}\}_{c\in C^{\psi}})$
    $\mathbf{X}^{\psi}_{val} \leftarrow concatenate(\{\widetilde{\mathbf{Y}}^c_{val}\}_{c\in C^{\psi}})$
    $\mathbf{X}^{\psi}_{te} \leftarrow concatenate(\{\widetilde{\mathbf{Y}}^c_{te}\}_{c\in C^{\psi}})$
    $\widehat{\mathbf{Y}}^{\psi}_{val}, \widehat{\mathbf{Y}}^{\psi}_{te}, Acc^{\psi}_{val}, Acc^{\psi}_{te} \leftarrow META(\mathbf{X}^{\psi}_{tr}, \mathbf{X}^{\psi}_{val}, \mathbf{X}^{\psi}_{te}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te})$
    return $\widehat{\mathbf{Y}}^{\psi}_{val}, \widehat{\mathbf{Y}}^{\psi}_{te}, Acc^{\psi}_{val}, Acc^{\psi}_{te}$
end procedure
```

In the following sections we will progressively explore the possible classification strategies that can be used with a BRE. For each method, we describe how do we obtain accuracy ratings and the predictions for the test data. We present the accuracy ratings of each method in the Experiments section by using the labels provided here.

All of the following methods are compatible to the supervoxels formed by a single clustering algorithm (N-Cuts, or K-Means), while the ones that do not require multiple clustering levels also are also compatible to the brain regions specified by AAL.

3.4.2 Random Subsets of Supervoxels for Brain Region Ensembles (RSS-BRE)

Algorithm 3.4 Random Subsets of Supervoxels for Brain Region Ensembles

```
procedure RSS-BRE($\{\mathbf{X}^c_{tr}\}_{c\subseteq C^\Psi}, \{\mathbf{X}^c_{val}\}_{c\subseteq C^\Psi}, \{\mathbf{X}^c_{te}\}_{c\subseteq C^\Psi}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te}$)
    given $c \in \bigcup_{\psi=1}^{N_\psi}(C^\psi) \subseteq C^\Psi$
    for $\psi = 1$ to $N_\psi$ do
        $\widehat{\mathbf{Y}}^{\psi}_{te} \leftarrow BRE(\{\mathbf{X}^c_{tr}\}_{c\in C^\psi}, \{\mathbf{X}^c_{val}\}_{c\in C^\psi}, \{\mathbf{X}^c_{te}\}_{c\in C^\psi}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te})$
    end for
    for $i = 1$ to $N_{te}$ do                                  ▷ Loop over all test samples
        for $l = 0$ to $N_l - 1$ do          ▷ Initialize counters for each class label $l$
            $count[l] \leftarrow 0$
        end for
        for $\psi = 1$ to $N_\psi$ do
            for $l = 0$ do$N_l - 1$
                $count[l] \leftarrow count[l] + \delta(l, \widehat{\mathbf{Y}}^{\psi}_{te}[i])$          ▷ Use Kronecker delta
            end for
        end for
        $\widehat{\mathbf{Y}}^{\Psi}_{te}[i] \leftarrow \underset{l}{\mathrm{argmax}}(count)$
    end for
    $RSS - BRE \leftarrow CalculateAccuracy(\widehat{\mathbf{Y}}^{\Psi}_{te}, \mathbf{Y}_{te})$        ▷ See Section 3.2.1.3
    return $RSS - BRE$
end procedure
```

In the original article of Ozay and Yarman-Vural [75] on FSG, a single meta classifier was built in order to fuse the posteriori probabilities of each sample for all feature spaces, each of which represents a specific attribute of the given dataset. The premise for that approach is that all feature spaces contribute to the final classification task. However, in our problem, it is more likely to find a subset of supervoxels that is more effective in classification of a set of mental states that are elicited by the stimuli in the fMRI experiments, rather than all availalbe supervoxels. As we mentioned in the earlier sections, the main problem is to find the set of supervoxels is effective in the classification of the particular mental states that are under consideration for the fMRI experiments. To address this problem, we propose to use random subsets of supervoxels (RSS) to be used with BRE, where a BRE is formed for each random subset, and the the predictions are aggregated with majority voting.

The primary premise of RSS-BRE is that the combined outputs of the meta classifiers that are based on random subspaces of supervoxels can improve upon training a single meta classifier on specifically selected set of supervoxels such as using AAL regions, or using a particular clustering parameter. Also, random subspaces method has been tried and tested for fMRI data, albeit the subspaces formed by the earlier implementations used the voxel space of selected voxels [56, 55].

The secondary premise of RSS-BRE is that, using random subsets of supervoxels instead of using random subspaces of selected voxels expected to provide a better classifier ensemble. This premise is based on the hypothesis that the diversity of the classifier ensemble RSS-BRE should be higher than the random subspace ensembles (RSE) that uses the voxel space of selected voxels. The reason for this hypothesis is the fact that RSS-BRE does not only include voxels which would be eliminated by a voxel selection process, but also aims to capture the distinct aspects of the mental representations in the supervoxels.

RSS-BRE is formed by using a set of random subsets of supervoxels $\{C^{\psi}\}_{\psi=1}^{N_{\psi}}$. Each subset C^{ψ} is sampled randomly from within a superset of supervoxels C^{Ψ}, specification of which is provided in the next couple of subsections. For each random subset C^{ψ}, a brain region ensemble is built. Then, the predictions $\widehat{\mathbf{Y}}_{te}^{\psi}$ are obtained for each BRE. Lastly, a voting scheme is performed by using the set of predictions, $\{\widehat{\mathbf{Y}}_{te}^{\psi}\}_{\psi=1}^{N_{\psi}}$ from all BRE's in order to get the final prediction (Algorithm 3.4). The algorithm works as follows: First, a BRE is built for each random subspace C^{ψ} (Algorithm 3.4, line 3). Then, a majority voting is performed over the class labels $\widehat{\mathbf{Y}}_{te}^{\psi}$, which are estimated for the test set as the outputs of every BRE (Algorithm 3.4, lines 7-16). The class label with the most votes is then reported as the final estimate for each test sample $i \in \{1, 2, ..., N_{te}\}$, where N_{te} is the number of test samples.

3.4.2.1 RSS-BRE with all Supervoxels

The most general form of RSS-BRE works with the set of supervoxels $C^{\Psi} = C = \bigcup_{\theta \in \Theta}(C^{\theta})$. This method does not have any specific assumptions about the composition and the size of the contributing supervoxels given the set of clustering parameters $\theta \in \Theta$.

Within the set of supervoxels $C^{\Psi} = C$, random subsets of supervoxels $C^{\psi} \subseteq C^{\Psi}$ are sampled, where for each random subset C^{ψ}, a BRE is built. The classification accuracy of the output predictions $\widehat{\mathbf{Y}}_{te}^{\psi}$ of each BRE are then combined with the majority voting scheme of RSS-BRE (Algorithm 3.4) in order to get the final prediction of RSS-BRE ($\widehat{\mathbf{Y}}_{te}^{\Psi}$).

3.4.2.2 RSS-BRE within Brain Partitions that are Tied to Specific Clustering Parameters

RSS-BRE can also be built for the brain partitions that are tied to specific clustering parameters, which would allow us to determine the optimal number of the supervoxels for the analysis of given set of fMRI data. This approach uses a set of the supervoxels generated by a clustering method using a specific clustering parameter as the superset for RSS-BRE, where there is a superset $C^{\Psi_i} = C^{\theta_i}$ for each $\theta_i \in \Theta = \{\theta_1, \theta_2, ..., \theta_{N_\theta}\}$, where Θ is the set of all clustering parameters θ_i, and N_θ is the number of clustering parameters.

RSS-BRE for each brain partitioning (specified by a clustering parameter θ_i), samples the random subsets of supervoxels within using its corresponding superset C^{Ψ_i}, as specified by Algorithm 3.4, producing test set predictions $\widehat{\mathbf{Y}}_{te}^{\Psi_i}$ and accuracy $Acc_{te}^{\Psi_i}$. RSE algorithm is also run for the validation set and accuracy results $Acc_{val}^{\Psi_i}$ are obtained. Using the results from the supersets RSS-BRE built for the brain partitions that correspond to each clustering parameter θ, $(\{\widehat{Y}_{te}^{\Psi_i}\}_{i=1}^{N_\theta}$, $\{Acc_{te}^{\Psi_i}\}_{i=1}^{N_\theta}$, and $\{Acc_{val}^{\Psi_i}\}_{i=1}^{N_\theta})$ the best clustering parameter for RSS-BRE is determined by using the validation accuracies. The test accuracy for the brain partitions that are generated by the best superset of supervoxels, which in turn correspond to clustering parameter that is optimal for RSS-BRE (θ_{best}), is then reported (Algorithm 3.5).

Algorithm 3.5 Selection of Optimal Clustering Parameter by RSS-BRE

```
procedure RSS-BRE PARAM($\{Acc_{val}^{\Psi_i}\}_{i=1}^{N_\theta}, \{Acc_{te}^{\Psi_i}\}_{i=1}^{N_\theta}, \{\widehat{Y}_{te}^{\Psi_i}\}_{i=1}^{N_\theta}$)
    $\Psi_{best} \leftarrow \underset{\Psi_i}{\text{argmax}}(\{Acc_{val}^{\Psi_i}\}_{i=1}^{N_\theta})$
    $BestParamRSS - BRE \leftarrow Acc_{te}^{\Psi_{best}}$
    return $BestParamRSS - BRE$
end procedure
```

3.4.3 BRE within Brain Partitions that are Tied to Specific Clustering Parameters

In this study, in addition to methods that are based on RSS, we have investigated the effectiveness of the supervoxels that are produced by a brain partitioning that is tied to a specific clustering parameter for building a single BRE. Using this method allows us to judge the effectiveness of RSS in terms of classification accuracy over an array of superset (C^Ψ) sizes each of which are tied to a clustering parameter. In order to find that optimal clustering parameter, and to compare the classification accuracy of the BRE that uses the corresponding subset of supervoxels, we propose the Algorithm 3.6. In the Algorithm 3.6, first, a BRE is built for each subset C^θ (lines 3-7). Then, BRE, which uses the subset of supervoxels and corresponding clustering parameter (θ_{best}), with the greatest validation accuracy is selected. Finally, test accuracy is reported for the best performing BRE.

Algorithm 3.6 Selection of Optimal Clustering Parameter by BRE

```
procedure BRE PARAM($\{\mathbf{X}^c_{tr}\}_{c \in C}, \{\mathbf{X}^c_{val}\}_{c \in C}, \{\mathbf{X}^c_{te}\}_{c \in C}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te}, \Theta$)
    given $\theta \in \Theta$ and $c \in C = \bigcup_{\theta \in \Theta}(C^\theta)$
    for all $\theta \in \Theta$ do
        $Acc^\theta_{val}, Acc^\theta_{te}, \widehat{\mathbf{Y}}^\theta_{te} \leftarrow BRE($
        $\{\mathbf{X}^c_{tr}\}_{c \in C^\theta}, \{\mathbf{X}^c_{val}\}_{c \in C^\theta}, \{\mathbf{X}^c_{te}\}_{c \in C^\theta}, \mathbf{Y}_{tr}, \mathbf{Y}_{val}, \mathbf{Y}_{te}$
        $)$
    end for
    $\theta_{best} \leftarrow \underset{\theta}{\text{argmax}}(\{Acc^\theta_{val}\}_{\theta \in \Theta})$
    $BestParamBRE \leftarrow Acc^{\theta_{best}}_{te}$
    return $BestParamBRE$
end procedure
```

3.4.4 Selection of the Best Performing Supervoxel with respect to Classification Accuracy

Algorithm 3.7 Best Performing Supervoxel

```
procedure BESTSUPERVOXEL($\{Acc^c_{val}\}_{c \in C}, \{Acc^c_{te}\}_{c \in C}, \{\widehat{Y}^c_{te}\}_{c \in C}$)
    $c_{best} \leftarrow \underset{c}{\text{argmax}}(\{Acc^c_{val}\}_{c \in C})$
    $MaxSVAcc \leftarrow Acc^{c_{best}}_{te}$
    return $MaxSVAcc, \widehat{Y}^{c_{best}}_{te}$
end procedure
```

In order to test the validity of the ensemble classification methods that we propose, we need to compare the classification accuracy of the best performing base classifier with the ensemble classification methods. If the best performing base classifier provides better overall accuracy than an ensemble that combines multiple supervoxels, such as *BREParam* method that is described in this section, it means we fail to fuse the mental representations captured by the base layer classifiers at the supervoxels. In order to perform that comparison, we find the best performing base classifier and the corresponding supervoxel in the set of all supervoxels $C = \bigcup_{\theta \in \Theta}(C^\theta)$, where $\theta \in \Theta = \{\theta_1, \theta_2, ..., \theta_{N_\theta}\}$. The selection algorithm is as follows.

Given the set of supervoxels C, the best performing supervoxel is selected by using the accuracy ratings of the base layer classifiers for the validation set. Then the accuracy ($BestSV$) for the best performing supervoxel is reported for the test set (Algorithm 3.7).

3.5 Diversity Measures for Classifier Ensembles

One of the premises of BRE method is that supervoxels can capture the mental representations that are distributed across specialized brain regions. In order to test this premise, we propose to use the diversity of the base layer classifiers as a measure, where each base layer classifier receives input from a specific supervoxel. If the classifiers are diverse, it would mean that the supervoxels can provide diverse representations of the underlying mental states. In order to set a baseline for the classifier diversity of the base layer classifiers, we use the diversity of the base layer classifiers that use regions specified by AAL as supervoxels. The reason for that is, we already know the regions specified by AAL have distinct functional and representational properties.

One other promise of BRE is to increase the diversity of the classifier ensemble with respect to the ensemles that are based on the selected voxels. The comparison is based on the classifier diversity of random subspace ensembles created by selected voxels and RSS-BRE of supervoxels.

In order to test these two premises we use two different diversity measures for classifier pairs: Q statistic [99] and the disagreement measure [47]. In order to find the diversity of a set of classifiers, the measures are averaged over all pairs in the set.

Q statistic is a symmetric measure of diversity between a pair of classifiers CF_i and CF_j (each of which is based on a random subspace of voxels, or random subset of supervoxels) which is defined as:

$$Q_{i,j} = \frac{ad - be}{ad + be} \tag{3.19}$$

where, a is the probability that both classifiers make the correct classification, b is the probability that CF_i is correct and CF_j is wrong, e is the probability that CF_j is correct and CF_i is wrong, and lastly d is the probability that both classifiers are wrong. Q statistic is calculated as an average for all classifier pairs over all samples at a classifier ensemble. The absolute values of Q statistics are used in this calculation.

The disagreement measure is defined as $D_{i,j} = b + e$ for a pair of classifiers.

3.6 Region Specification with BRE

One of our aims is to validate our approach by locating the brain areas that have functional properties that are important for the underlying stimului or mental tasks using supervoxels. We propose to locate such regions by finding the discriminative supervoxels which are specified in this section. In order to compare them with the existing neuroscientific literature, we use the labels generated by AAL [92] as discussed below.

In this section, we provide a method to select the supervoxels that contain discriminative information regarding the binary classification of two distinct mental states. Similar to BRE, this is also a generic approach that is applicable to a wide range of brain decoding applications.

In order to find the supervoxels that are discriminative for a pair of distinc mental states as specified by an fMRI experiment, we use the chance level accuracy of the base layer classifier that receives input from a specific supervoxel as a measure. We calculate a discriminative stability threshold, where the base layer classifier should correctly classify the test samples $N_{correct}$ times over all of the runs in order to be discriminative. The discriminative stability threshold for a base layer classifier is calculated as follows: For each supervoxel c, N_m samples are classified by using the base layer classifier trained using that supervoxel. Suppose that,the class assignment is performed by the random variable $B_{m,c}$ at the cluster c for each sample s. Then, the Bernoulli random variable $B_{m,c} \sim Bernoulli(p = 1/N_l)$ (where N_l is the total number of class labels) is:

$$B_{m,c} = \begin{cases} 1, & \text{class assignment is correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.20}$$

The probability of having the number of successful guesses more than or equal to $N_{correct}$ at cluster c, over N_m samples , is given by the probability density function of the of Binomial random variable $Z_c = \sum_{r=1}^{N_m} B_{m,c}(N_m, p = 1/N_l)$. which is :

$$P(Z_c \geq N_{correct}) = \sum_{r=N_{correct}}^{N_m} \left(\begin{array}{c} N_m \\ r \end{array} \right) p^r \, (1-p)^{(N_m - 1)} \, . \tag{3.21}$$

under the assumption that Z_c are independent and identically distributed (IID). Using Equation 3.21, stable supervoxels for each class, across the validation runs can be selected by setting a threshold ($N_{correct}$) on the number of successful classifications. Note that, the IID assumption makes the probability given by (Equation 3.21) to be an upper bound over the cases where the independence assumption does not hold. Thus, the threshold $N_{correct}$ is a conservative one when rejecting the null-hypothesis.

We use the overlaps of the discriminativle supervoxels with the regions specified by AAL in order to compare our results with the existing neuroscience literature. For that purpose, we first select the supervoxels for which the base layer classifier performs better than chance level by using Equation 3.21, where $(Z_c \geq N_{correct}) \geq 0.01$. Then, for each discriminative supervoxel $c^{disc} \in C^{\theta}$ for the clustering parameter θ and within each AAL region ω, the number of voxels that belong to the discriminative supervoxels $N_v^{(\omega,disc)} = \sum_{c^{disc}} |c^{disc} \cap c^{\omega}|$ are determined. Here, $| \cdot |$ signifies cardinality, c^{ω} signifies the supervoxel that belongs to a specific AAL region. Lastly, for each AAL region ω we calculate the ratio:

$$h(\theta, \omega) = \frac{N_v^{(\omega,disc)}}{\sum_{\omega} N_v^{(\omega,disc)}}, \tag{3.22}$$

where we determine the distribution of the voxels that belong to the discriminative supervoxels (i.e. discriminative voxels) across AAL regions for the clustering parameter θ. We perform this operation for the brain partitions that are obtained using every clustering parameter in order to observe the effect of clustering parameter over the distribution of discriminative voxels.

3.7 Chapter Summary

In this chapter we have presented our method called Brain Region Ensembles (BRE) for decoding mental states from the fMRI data. The suggested method consists of several steps. First, the brain volumes are partitioned into homogenous regions, called supervoxels. The supervoxels can be considered as an alternative to brain partitions formed by functional brain atlases such as AAL. At the second step, the voxel intensity values at each supervoxel is fed to a logistic regression classifier, which we call a base layer classifier. The output of each classifier represents the posteriori probabilities of mental states under consideration. The posteriori probabilities are then concatenated for a set of base layer classifiers, which correspond to a set of supervoxels, and fed to a meta classifier in order to form a brain region ensemble.

The composition of the set of supervoxels is deemed to be critical for the success of a BRE. However, selecting the supervoxels with respect to the accuracy of the base layer classifiers that are tied to them have proven ineffective for finding the right composition of supervoxels, where simply using all supervoxels for construction of a BRE yielded to higher classification accuracies than using a selected set of supervoxels. In order to bypass the need of finding the right composition of supervoxels, we have suggested a method that randomly samples subsets of supervoxels for each which, a BRE is constructed. We call this method RSS-BRE. With RSS-BRE, the outputs of multiple BRE are combined in a majority voting scheme in order to obtain the final predictions for the class labels of the mental states under consideration.

The method for forming random subsets of supervoxels for brain region ensembles (RSS-BRE) is based on the random subspaces method for classifier ensembles (RSE), with which, random subspaces of selected voxels are constructed in order to form a classifier ensemble [56]. We compare our RSS-BRE method with the RSE that uses voxels selected by voxel selection algorithms presented in the previous chapter. The comparison is done not only in terms of classification accuracy that each of these methos provide. But also, we analyze the diversity of the ensemble of classifiers for these two methods in order to justify our claims regarding the ability of the supervoxels to capture the diverse representations of the mental states in the brain. In this chapter, we have provided two diversity measures to compare the diversity of RSS-BRE that is based on supervoxels with respect to RSE that is based on selected voxels.

In addition to the diversity comparison of the classifier ensembles RSS-BRE and RSE, we suggested to perform a comparison of the diversity of the base layer classifiers that are tied to the supervoxels formed by the clustering algorithms, and the base layer classifiers that are tied to the supervoxels formed using AAL. With this comparison, our aim is to compare the brain partitions formed by the clustering algorithms with brain regions that are already considered diverse, which are AAL regions, in terms of their functional properties.

Lastly, we have proposed a method for the specification of the discriminative supervoxels that relies on the accuracy ratings of the base layer classifiers for classification of a given set of mental states. With this method, suppervoxels for which the base layer classifier accuracy is higher than the threshold for chance level accuracy are marked as discriminative supervoxels. The discrimative supervoxels are then used for the specification of the brain regions that are differentially active for the given categories of mental states.

CHAPTER 4

VALIDATION, VERIFICATION AND EMPIRICAL ANALYSIS OF BRAIN REGION ENSEMBLES (BRE) METHOD

In this chapter we present the applications of our methods on three datasets namely Objects, Emotion, and TOL datasets. We begin our analysis with the comparison of BRE and random subspace ensembles applied to the selected voxels with respect to the diversity measures that are described in the previous section. Then, we proceed to the classification experiments and comparative results of the experiments in all datasets. Lastly we present the voxel distributions of the selected regions for Objects and Emotion, and TOL datasets.

4.1 Classifier Diversity Analysis

It is well-known that the measure of diversity among the classifiers within a classifier ensemble is an indicator to the success of the ensemble learning methods. Therefore, in this first group of computational experiments, we analyze the difersity of classifiers in the ensembles. In this section, two different sets of experiments are performed to test the diversity. First, we compare the diversity of the base layer classifiers with two different types of inputs. The first type of inputs are the supervoxels formed by clustering algorithms, where the second type are the supervoxels formed by AAL. Here, we expect diversity of the classifiers that uses the supervoxels generated by the clustering algorithms to be similar to those that are specified by AAL regions.

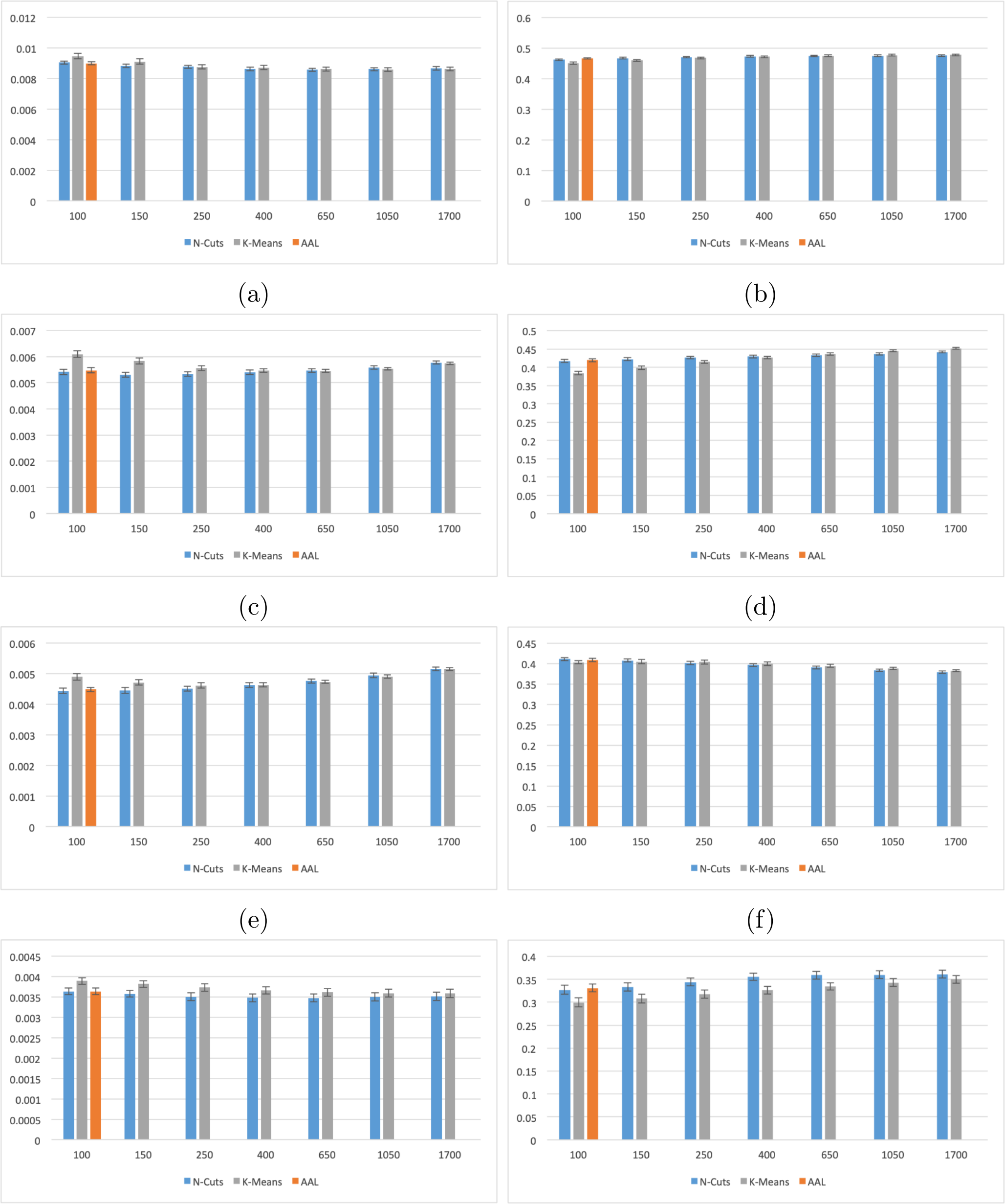
Second, we compare the diversity of the classifiers formed by RSS-BRE (random subsets of supervoxels for brain region ensembles) method to the classsifiers that are formed by RSE which uses random subspaces of selected voxels. The comparison is based on the random subsets of supervoxels, where, for each one, a BRE is formed, and the random subspaces of voxels, where, for each one, an SVM classifier is trained. Since the outputs of the diversity calculations can be affected by the number of classifiers, we have an equal number of subsets of supervoxels for RSS-BRE, and subspaces of voxels for RSE for this comparison. With this comparison, we expect classifiers of RSS-BRE method to have higher diversity with respect to RSE of voxel subspaces as the number of subsets/subspaces increase. That is due to the fact of the number of subsets being proportional to the number of brain partitions (i.e. supervoxels). In other words, as the number of brain partitions increase, each supervoxel represents a smaller region of the brain, which would increase the likelihood of a more diverse composition of supervoxels within a subset of supervoxels.

4.1.1 Diversity of Base Layer Classifiers that Use Supervoxels

As we have mentioned in the previous sections, FSG is an ensemble learning algorithm that is well suited to fuse the information embedded in different feature spaces by first estimating and then concatenating the posteriori probabilities for each mental task or stimulus. The feature vector, which consist of all the posteriori probabilities obtained from each mental task or stimulus for each region is then fed to a meta classifier to estimated the final prediction. The classifier diversity heavily depends on the input feature space of the base layer classifiers and it is an important measure for the success of the meta classifier. For AAL regions, it is shown that each region processes a different aspect of the task or stimulus provided by a mental task or stimulus of an fMRI experiment. Thus, AAL regions can be considered a natural alternative to the set supervoxels, which are generated by a brain partitioning process such as clustering, for the base layer classifiers as inputs. Yet, the diversity of the base layer classifiers that uses the supervoxels obtained as the outputs clustering algorithms (K-Means and constrained N-Cuts) as their inputs is not studied until now. In this section, we provide a comparative analysis of classifier diversity for the base layer classifiers using supervoxels that are specified by AAL and those generated by clustering algorithms.

Given a set of supervoxels C^θ generated by the clustering parameter $\theta \in \{100, 150, 250, 400, 650, 1050, 1700\}$, or supervoxels specified by AAL, classifier diversity measures Q-Statistic, and Disagreement measure are calculated for the set of base layer classifiers that are formed using the set of supervoxels C^θ, or for the AAL regions, C^{AAL}. Here, each clustering parameter is equal to the number of generated supervoxels by clustering the whole brain volume. For each base classifier, regularization parameters of the classifiers $\phi \in \{0.1, 1, 10\}$ are optimized by using training and validation sets. The diversity values of supervoxels specified by AAL regions (100 regions in total, depending on the fMRI dataset) are compared with the supervoxels that are generated when $\theta = 100$.

The diversity measures for all datasets (Q-Statistic and Disagreement measure, see Section 3.5) are shown in Figure 4.1. Q-Statistic is displayed on the left column (lower is better), while Disagreement Measure (higher is better) is displayed on the left one.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 4.1: Q-Statistic (column on the left, lower is better) and Disagreement Measure (column on the right, higher is better) for subspaces specified with N-Cuts, K-Means and AAL. The first row (a,b) shows the diversity values for the Objects dataset while the second row (c,d) is for the Emotion 2-Class experiment (e,f), the third one is for the Emotion 4-Class experiment, and the last one is for the TOL dataset (g,h). The number of supervoxels in the brain partitions that are generated by clustering algorithms are given at the bottom of every graph (100 to 1700).

4.1.1.1 Base Layer Diversity for Objects Dataset

For the Objects dataset (Figure 4.1.a,b), the diversity of the base layer classifiers that uses AAL regions are higher than those uses supervoxels generated by the two clustering algorithms for the both diversity measures by small margins. When the supervoxels formed by the clustering algorithms are considered; as the number of supervoxels increases, the diversity of the classifiers increases. This result can be explained with the nature of the dataset, where both of the stimulus classes used in this fMRI experiment were visual stimuli, for which the diversity in the visual representations can be captured better by the smaller supervoxels.

4.1.1.2 Base Layer Diversity for Emotion Dataset with 2 Classes

For the Emotion dataset with 2 classes (Figure 4.1.c,d), similar to the Objects dataset, base layer classifiers that use supervoxels formed by the clustering algorithms have lower classifier diversity than those use AAL regions. For this dataset, classifier diversity, as measured by the Q statistic (Figure 4.1.c) increases as the number of supervoxels increase up to 400 voxels, where the diversity starts to decrease as the number of supervoxels is increased further. Whereas, classfier diversity measured by the Disagreement measure increases monotonically as the number of supervoxels increase. Since the Disagreement measure steadily increases, the decrease in the diversity as measured by the Q-statistic is either caused by the increase in the probability of the classifier pairs both being wrong, or classifier pairs both being correct, or both (see 3.5 for the definition of the two measures). When we consider the fact that, as the number of supervoxels increase, they get smaller and they would less likely to include a voxel that is critical for the classification task, we can conclude that the decrease of the diversity as measured by Q-Statistic is due to appearence of smaller supervoxels which does not correctly classify some of the samples. In conrtast to the Objects dataset, where the task was to recognize visual objects (bird or flower), Emotion 2 class dataset require the classification of mental states (fear or disgust) that have emotional content versus the visual and semantic representations of furniture and kitchen appliances. Given this fact, the increase in the diversity measured by the Q-statistic can be explained by the observation that as the supervoxels get smaller, the details of visual representations can be more likely to be captured by individual supervoxels within the areas that are dedicated to visual processing such as visual blobs, textures, hence the diversity increase for the Objects dataset.

The decrease in the diversity as measured by the Q-Statistic can stem from the decrease in the likelihood of smaller supervoxels to capture the emotional states, where smaller supervoxels would not capture any specialized activity regarding emotions, except for the regions that directly responsible for the processing of the emotions. In other words, smaller supervoxels does not necessarily capture smaller components of emotional representations in contrast to the visual object representations where smaller components such as color, texture, visual blobs, are known to exist in the brain. Thus, as the supervoxels get smaller, the voxels that are critical in processing emotional stimuli would be excluded from some of the supervoxels, while the emotional representations captured by the supervoxels that retain a critical voxel would not be entirely different thant a larger supervoxel that contain a multitude of such critical voxels, causing the decrease in the diversity as measured by the Q-statistic by increasing the likelihood of incorrect classifications.

4.1.1.3 Base Layer Diversity for Emotion Dataset with 4 Classes

For the Emotion dataset with 4 classes, where the mental states to be classified are fear, disgust, kitchen appliances, and furniture, the effect of emotional brain activity versus the activity of processing visual stimuli is more prominent. In this case, as the supervoxels get smaller, classifier diversity measured by the both Q-statistic and Disagreement measure decrease (Figure 4.1.e,f). We explain this effect by the observation that while emotional activity affect a larger portion of the brain volume than the visual representations of the visual objects, smaller supervoxels does not necessarily capture different aspects of the emotional representations in contrast to visual representations where small supervoxels can capture sub-components of a visual object representation within the brain volume.

4.1.1.4 Base Layer Diversity for TOL Dataset

The TOL dataset is essentially different from the other datasets that we analyze in this study, where the visual stimuli does not change completely across the mental states under consideration. For this dataset, our task is to determine the state of the puzzle solving process of the participants whether they are in the planning stage, or in the execution stage. For us, it is unclear whether the supervoxels should have more diverse representations or not, as the size of the supervoxels get smaller. However, our results suggest that diversity of the base layer classifiers increase as the supervoxels get smaller (Figure 4.1.g,h) for the both of the diversity measures that we consider in this study. With that in mind, we can suggest that planning and execution states for problem soving can be represented by the diverse activity in relatively smaller, specialized brain regions.

To sum up, we can conclude that, for all datasets, constrained N-Cuts and K-Means clustering provide sets of supervoxels that are comparable to the AAL regions in terms of diversity diversity of the base layer classifiers that they are tied to. For the three experiments (Objects, Emotion 2-Classes, and TOL) the diversity within the set of base layer classifiers increases as the number of supervoxels (thus, the number of base layer classifiers) increases. For the Emotion 4-Classes experiment, the diversity within the subsets of supervoxels decrease while the number of supervoxels in the subsets increases. This result is reflected in the classification experiments (Section 4.2), where the classification accuracy of BRE with optimal clustering parameter is lower than the one provied by the BRE that receives input from the AAL regions for Emotion 4-Classes experiment. For the other experiments, performance of the BRE that receives inputs from the set of supervoxels that are generated by the optimal clustering parameter outperforms the performance of the BRE that receives input from AAL regions.

4.1.2 Diversity comparison of Brain Region Ensembles that use Random Subsets of Supervoxels (RSS-BRE) with respect to Random Subspace Ensembles of Voxels (RSE)

In order to determine the classifier diversity of brain region ensembles based on random subsets of supervoxels (RSS-BRE) and random subspace ensembles of voxels selected by the voxel selection methods (Section 2.4.2), we used two different statistics that are explained in Section 3.5. The classifier pairs we use for the calculation of the diversity measures are the BREs that use different subsets of supervoxels as their inputs and the SVMs that use different subspaces of voxels as their inputs. The calculated diversity measures are averaged for the all BRE pairs within a RSS-BRE, and all SVM pairs within a RSE. For comparison, we use the same set of numbers for the number of generated supervoxels and selected voxels $\{100, 150, 250, 400, 650, 1050, 1700\}$ to form the supersets of voxels, and supervoxels within which the random sampling procedure is performed. The supervoxels are formed by partitioning the brain volume of all voxels into homogenous regions (supervoxels), and the number of supervoxels is determined by the clustering parameter of a clustering algorithm.

The voxel selection methods select voxels from the brain volumes that contain all voxels. For that purpose, we use three different voxel selection methods that are prominent in the brain decoding literature: Voxel selection by using SVM, MI and ANOVA (Section 2.4.2).

Recall that, in order to form a BRE, we need to select a set of supervoxels for which we collect the posteriori probabilities from the base layer classifiers that are tied to each of such supervoxels and then train a meta classifier using the posteriori probability outputs. This task is achieved by concatenating the estimated posteriori probabilities of mental tasks generated by the selected base layer classifiers and feeding these posteriors to the input of a meta classifier. In the previous chapter (3.4.2), we suggest to randomly select subsets of supervoxels from within the set of supervoxels formed by the brain partitions as the result of a clustering procedure that uses a particular clustering parameter θ. That procedure forms subsets of supervoxels that are selected from within a number of supervoxels that is determined by the θ parameter, where for each subset of supervoxels a BRE is built. Similarly, in order to form random subspace ensembles of voxels, we randomly sample subspaces of voxels from within a set of voxels that is originally formed by a voxel selection procedure, where for each subspace of voxels, an SVM classfier is trained. Thus, we form a set of supervoxel sets each of which are generated by a clustering process that is tied to a specific θ parameter and we form a set of voxel sets each of which results from selecting a number of voxels using a particular voxel selection algorithm. As a result, we have voxel spaces, and supervoxel sets that have equal number of voxels/supervoxels in them, depending on the number of selected voxels and the θ parameter, where we form supervoxel sets with cardinalities: $\{100, 150, 250, 400, 650, 1050, 1700\}$, and voxel spaces with voxel counts: $\{100, 150, 250, 400, 650, 1050, 1700\}$. We form a RSS-BRE for each one of the said supervoxel sets and form a RSE for each of the voxel spaces. The number of voxels/supervoxels in the sets start from 100, which is roughly equal the number of brain regions specified by AAL, and increased in the multiples of the Fibonnacci sequence in order to obtain sets with meaningfully different number of elements in order to observe the trends in the change of classifier diversities.

Classification for supervoxel sets is done by the RSS-BRE that is built for that supervoxel set. Likewise, an SVM classifier is trained for each subspace of selected voxels by the above stated algorithms. For each method, regularization parameters of the classifiers $\phi \in \{0.1, 1, 10\}$ are optimized by using training and validation sets. The reported test set results are used for calculating Q-statistics [99] and the disagreement measures [47] which are presented in Section 3.5.

In order to form a random subset of supervoxels, we randomly sample half of the supervoxels whithout replacement whose are generated by a clustering algorithm with a specific parameter, where we perform this operation 100 times as suggested in [56]. Similarly in order to form a random subspace of voxels, we sample half of the voxels at random without replacement within a voxel space, where we repeat this procedure 100 times to form 100 different random subspaces of voxels.

To sum up, we calculate the classifier diversity within the RSS-BRE built using the set of supervoxels generated by the clustering process that uses a specific clustering algorithm. This process is repeated for all clustering algorithms and all clustering parameters, where for each of them we build a seperate RSS-BRE and calculate classfier diversity for them. Also, we build a RSS-BRE using the brain regions specified by AAL, and calculate the diversity within RSS-BRE. Also, we build a RSE for a voxel space generated by a set number of voxels selected by a specific voxel selection algorithm and calculate diversity within that ensemble. We repeat this process for all voxel selection algorithms and the voxel numbers that we mentioned above.

When we compare the voxel spaces formed by the selected voxels, and the supervoxel sets formed by brain partitioning, we can make the following observations.

First, voxel spaces for RSE contain much less voxels than the sets of supervoxels for RSS-BRE, which contain the voxels from the whole brain volume partitioned into a number of supervoxels. In other words, a voxel space that contains 100 voxels only include the voxels that are most relevant to mental tasks or stimuli specified by the fMRI experiment with respect to the voxel selection measure, whereas a supervoxel set containing 100 supervoxels is formed by partitioning the whole brain volume into 100 seperate regions. Consequently, it is not possible to form random subsets of supervoxels and random subspaces of voxels that are based on an equal number of voxels let alone an equal set of voxels unless we select the set of all voxels to form the voxel space, which makes voxel selection irrelevant.

Second, as the number of selected voxels as well as the number of partitions through clustering are both increased, for instance, when we use 200 voxels and 200 supervoxels instead of 100 to form the voxel space, and the set of supervoxels, the diversity of classifiers, which are brain region ensembles each of which uses a random subset of supervoxels, and SVM classifiers each of which uses a random subspace of voxels, expected to decrease. That is because the inclusion or exclusion of a specific voxel/supervoxel to a subspace/subset would have a lesser effect to the performance of the classifiers when the number of subspaces/subsets are increased in the original voxel/supervoxel sets.

Third, as the number of voxels selected by the voxel selection algorithm increases, the diversity of the classifiers based on random subspaces would get lower. That is because, as the set of selected voxels gets larger, the voxels with activity patterns that are less discriminative with respect to the mental states will get likely to be included in the set. Since the number of voxels that are included in all of the supervoxels stays the same (the whole brain volume), such an effect would not be observed for random subspaces of supervoxels.

Fourth, as we have presented in the Section 4.1.1, the diversity of the base layer classifiers increase as the supervoxels, which are generated by partitioning the whole brain, get smaller.

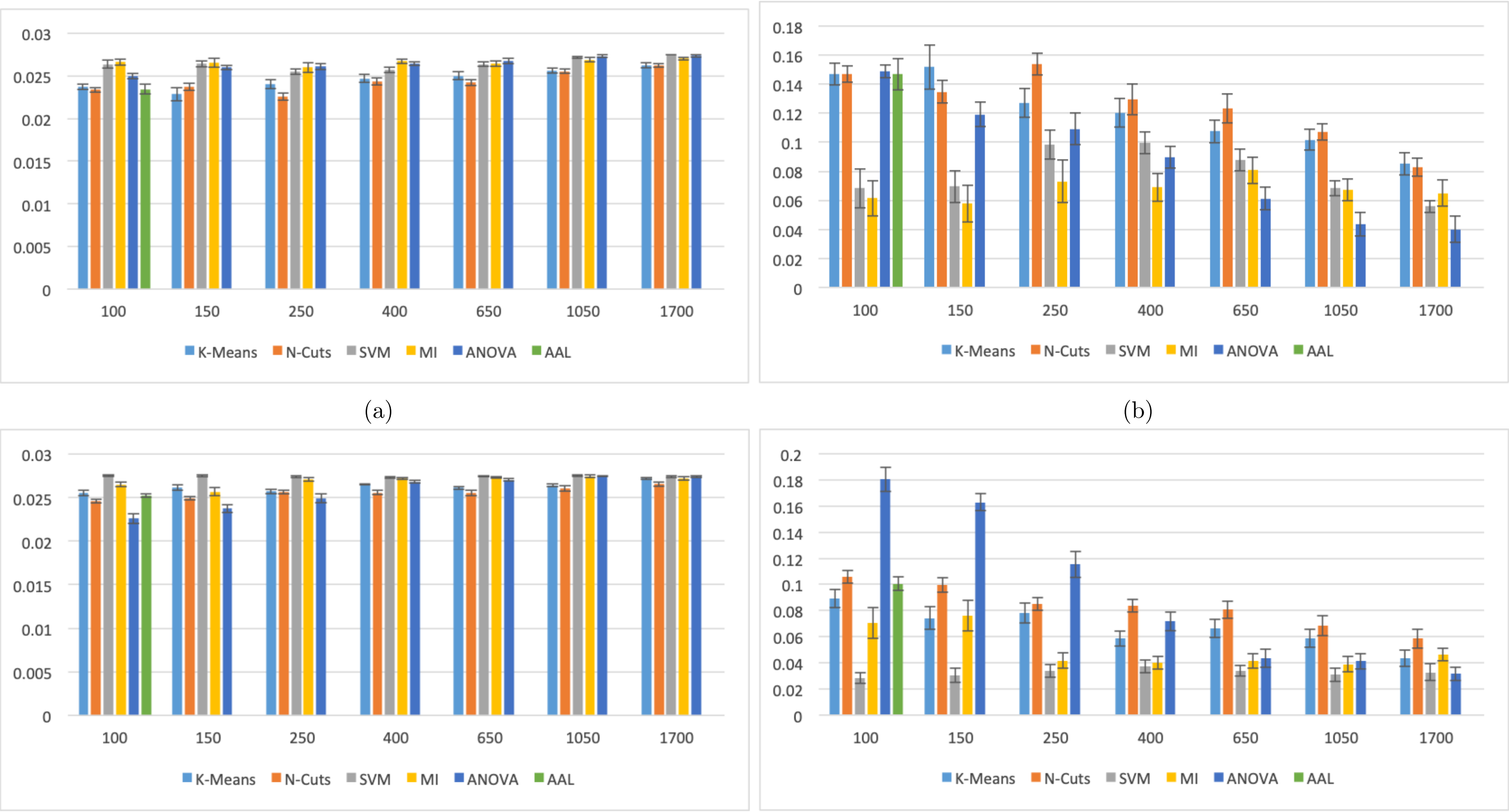
When we consider the second and the third observations stated above, we expect a decrease in classifier diversity for classifiers based on random subspaces as the number of voxels/supervoxels increase. However, when we consider the third, and fourth observations, we expect classifiers based on random subspaces of voxels to get much less diverse as the number of selected voxels increases when compared to RSS-BRE. Depending on how these factors interact, we would even observe an increase in the diversity for RSS-BRE as the number of supervoxels that partition the brain is increased.

In the following sections we present the calculated classifier diversity measures for Objects and Emotion (for 2 and 4 classes) datasets for all subjects.

4.1.2.1 Diversity Comparison of RSS-BRE vs. RSE for the Objects Dataset

The classifier diversities of RSS-BRE and RSE which are measured by Q-Statistic and the Disagreement Measure for each subject of the Objects dataset are given in Figures 4.2 and 4.3. Here, the lower the Q-statistic and the higher the disagreement measure, the diverse the ensemble of classifiers. The classifier ensembles are formed using sets of supervoxels (RSS-BRE) and voxes (RSE) where number of voxels/supervoxels are specified at the bottom of every graph.

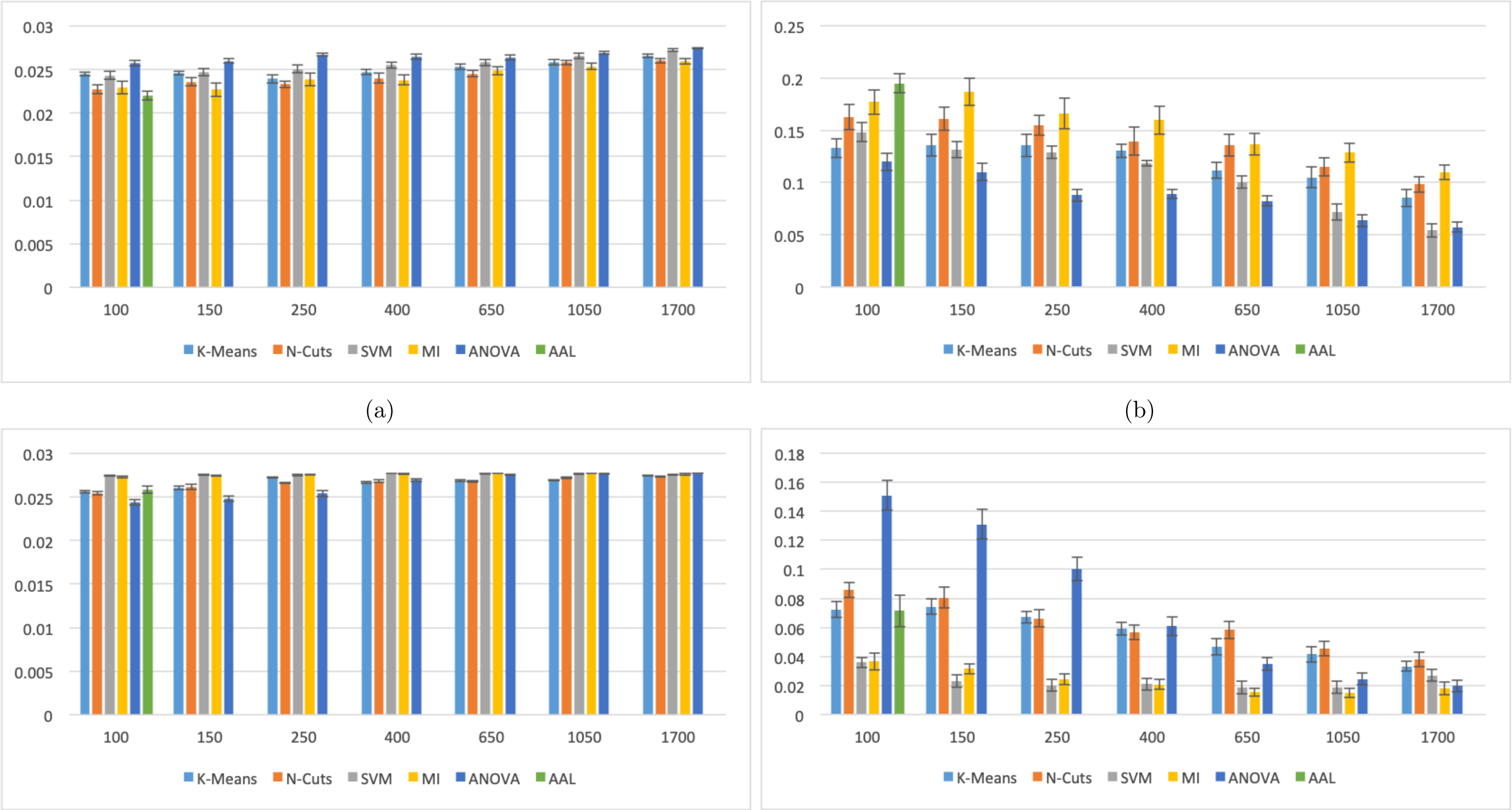
When compared to RSE, RSS-BRE generally yield higher classifier diversities (Figure 4.2 and 4.3). Here, K-Means, N-Cuts, and AAL specify the methods by which the supervoxels are formed while SVM, and MI, and ANOVA specify the methods by which the voxels are selected (Section 2.4.2). RSE that uses voxels selected by using ANOVA for subjects 2 and 4, and using SVM for subject 3



(a) (b)

(c) (d)

Figure 4.2: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Objects dataset. The top row of figures (a,b) represents the results for subject 1, while the bottom row (c,d) represents the results for subject 2. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).



(a) (b)

(c) (d)

Figure 4.3: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Objects dataset. The top row of figures (a,b) represents the results for subject 3, while the bottom row (c,d) represents the results for subject 4. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).

yields to more diverse ensembles than the RSS-BRE formed using the sets of supervoxels generated using K-Means and N-Cuts clustering for some clustering parameters. However, in general, RSS-BRE that are built using supervoxels generated by the clustering methods K-Means and N-Cuts have higher classifier diversitiy (Figure 4.2 and 4.3) when compared to RSE that uses voxels selected by the voxel selection algorithms. The RSS-BRE built using AAL regions provided diversity values similar to the RSS-BRE built using supervoxels generated by clustering for all Subjects (Figure 4.2 and 4.3).

The diversity of the classifiers in both RSS and RSE decreases as the supervoxel sets and voxel spaces get larger. However, diversity among the classifiers of RSE declines faster when compared to RSS-BRE as the voxel space or set of supervoxels get larger, as we have predicted in the beginning of this section.

4.1.2.2 Diversity Comparison of RSS-BRE vs. RSE for the Emotion Dataset: 2 Classes

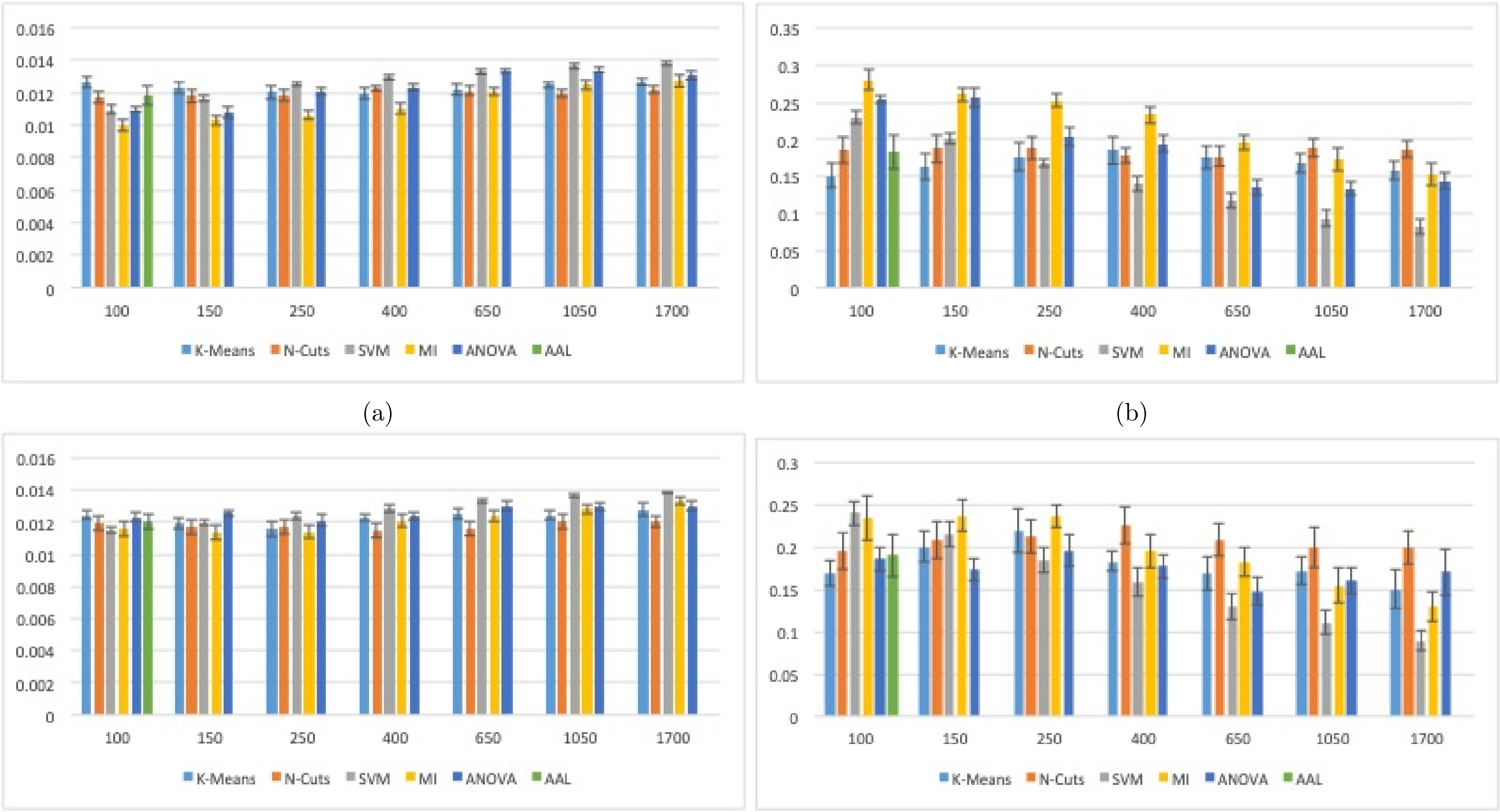
The classifier diversities of RSS-BRE and RSE which are measured by Q-Statistic and the Disagreement Measure for each subject of the Emotion dataset with 2 classes are given in Figures 4.4, 4.5, and 4.6. Here, the lower the Q-statistic and the higher the disagreement measure, the diverse the ensemble of classifiers. The classifier ensembles are formed using sets of supervoxels (RSS-BRE) and voxes (RSE) where number of voxels/supervoxels are specified at the bottom of every graph.

The two class version of the Emotion dataset has emotional (fear, disgust) and non-emotional (furniture, kitchen appliances) categories. For this dataset, the voxel selection methods provide a more diverse set of classifiers for RSE for up to 650 voxels when compared to the RSS-BRE built with same number of supervoxels. Beyond that point, RSS-BRE start to generate more diverse classifiers than BRE (Figures 4.4, 4.5, and 4.6).

For this dataset, the diversity of RSS-BRE classifiers increase as the number of supervoxels are increased up until 400 supervoxels. The increase in the diversity of RSS-BRE classifiers would be the result of the increase in the diversity of the base layer classifiers as the supervoxels get smaller (the fourth observation in Section 4.1.1).

4.1.2.3 Emotion Dataset: 4 Classes

The classifier diversities of RSS-BRE and RSE which are measured by Q-Statistic and the Disagreement Measure for each subject of the Emotion dataset with 4 classes are given in Figures 4.7, 4.8, and 4.9. Here, the lower the Q-statistic and the higher the disagreement measure, the diverse the ensemble of classifiers. The classifier ensembles are formed using sets of supervoxels (RSS-BRE) and voxes (RSE) where number of voxels/supervoxels are specified at the bottom of every graph.



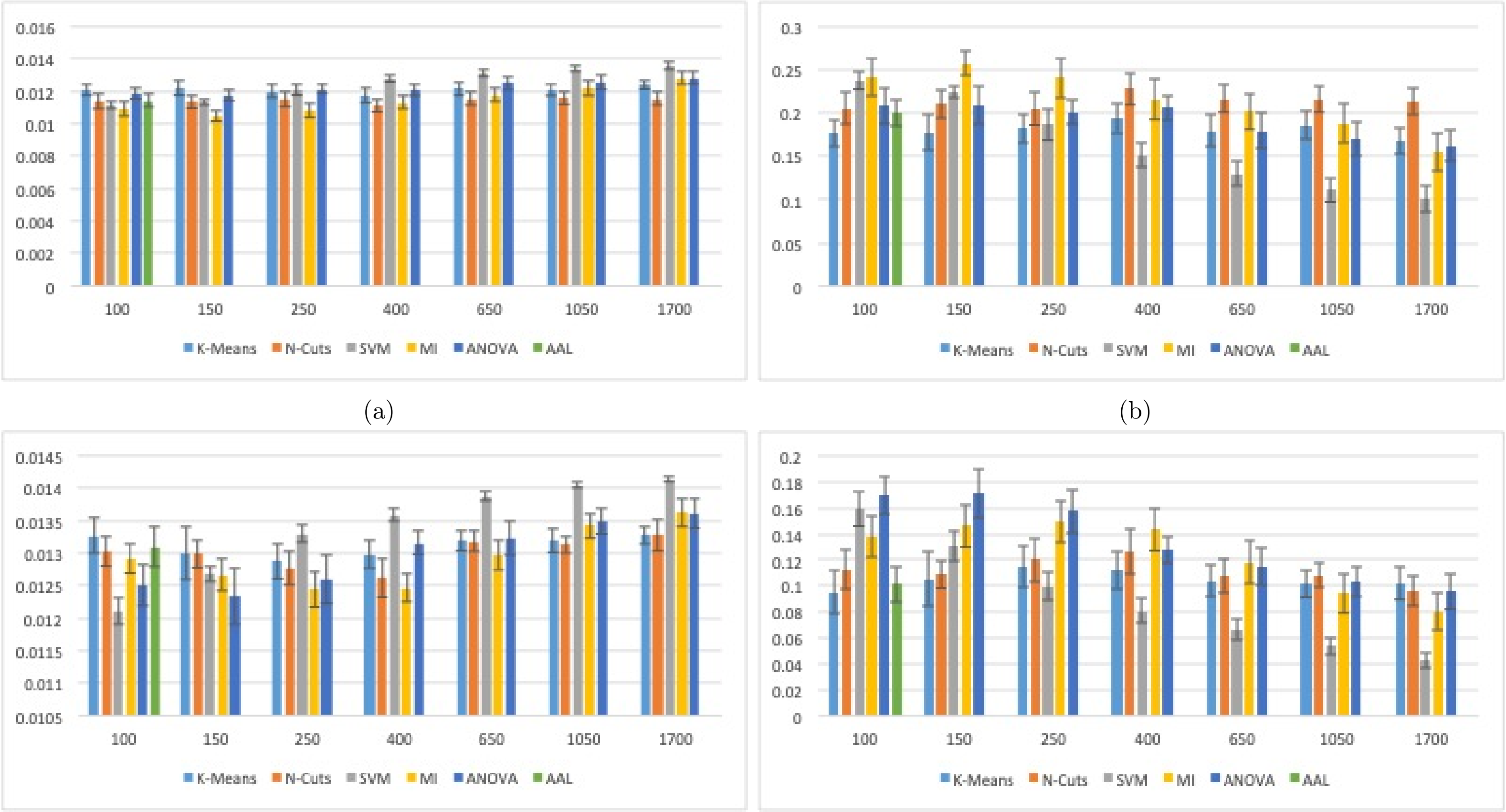
(a)

(b)

(c)

(d)

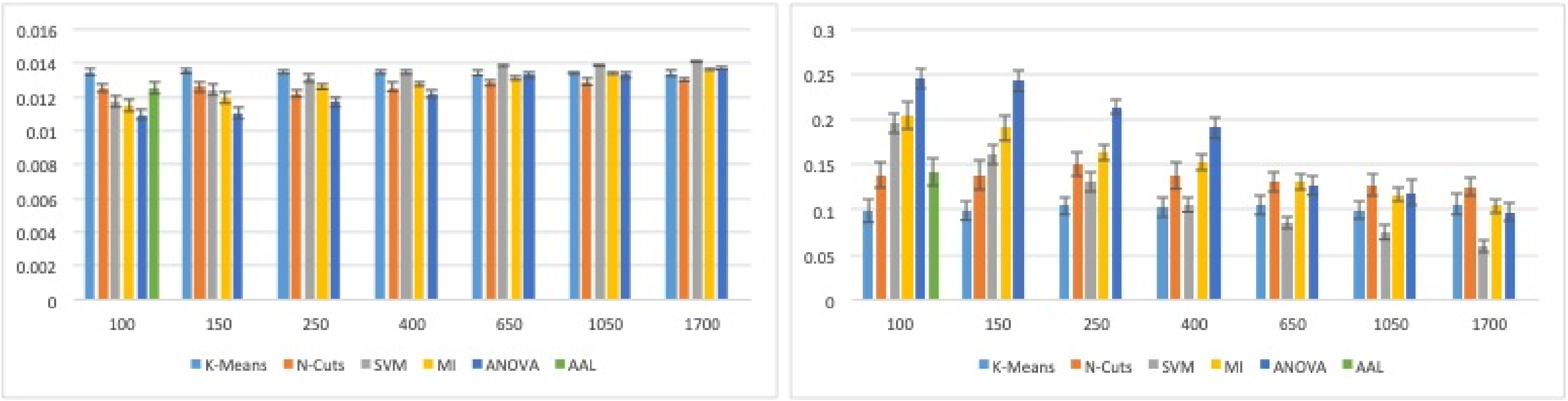
Figure 4.4: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 2-Classes dataset. The top row of figures (a,b) represents the results for subject 1, while the bottom row (c,d) represents the results for subject 2. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).



(a) (b)

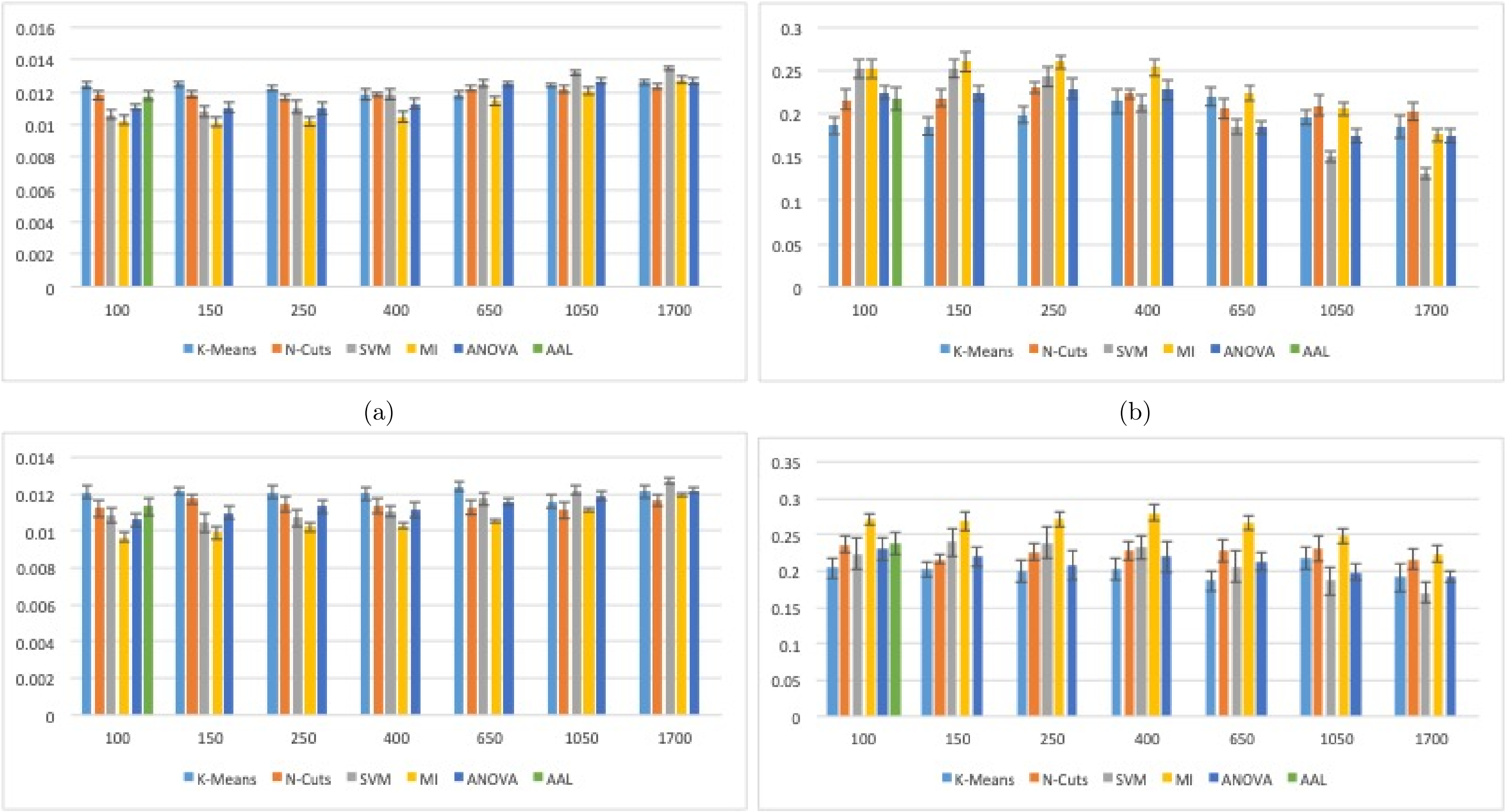
(c) (d)

Figure 4.5: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 2-Classes dataset. The top row of figures (a,b) represents the results for subject 3, while the bottom row (c,d) represents the results for subject 4. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).



(a) (b)

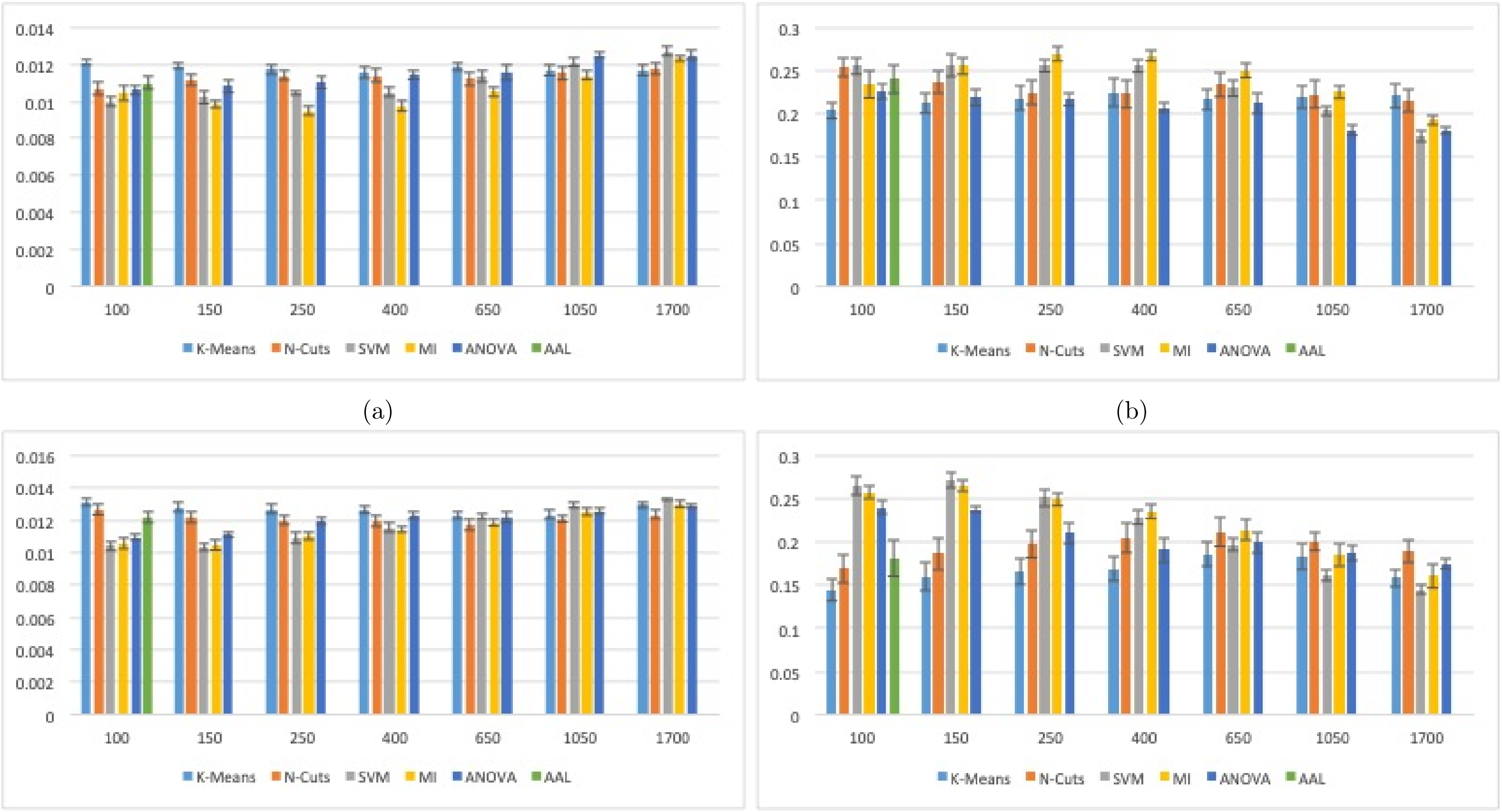
Figure 4.6: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 2-Classes dataset. The results are given for Subject 5. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).



(a) (b)

(c) (d)

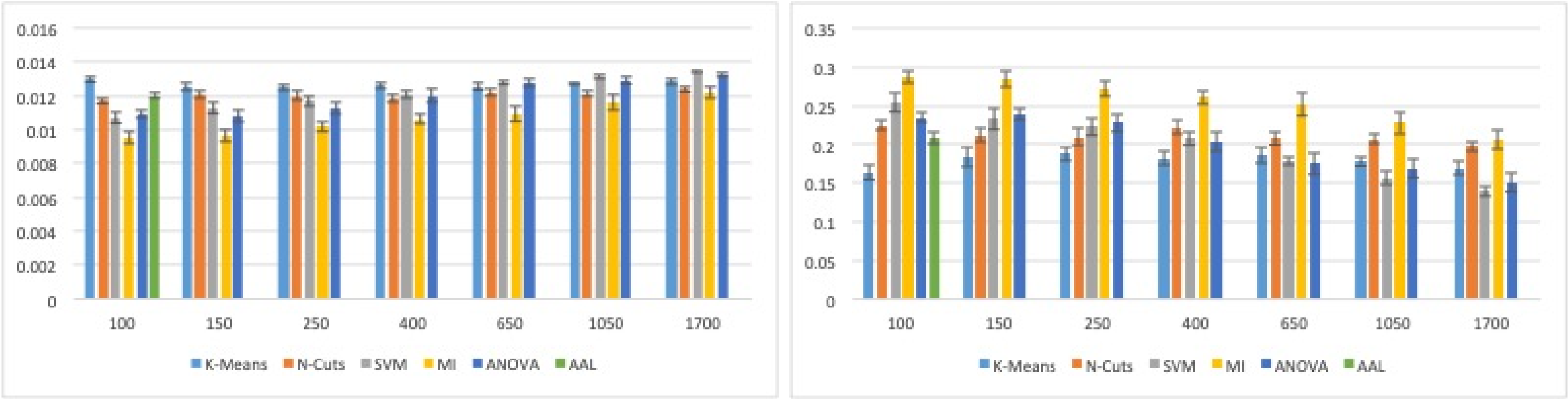
Figure 4.7: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 4 Classes dataset. The results are given for the Subjects 1 (a,b) and 2 (c,d). The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).



(a) (b)

(c) (d)

Figure 4.8: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 4 Classes dataset. The results are given for the Subjects 3 (a,b) and 4 (c,d). The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).



(a) (b)

Figure 4.9: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for Emotion 4 Classes dataset. The results are given for the Subject 5. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).

Similar to Emotion dataset with 2 classes, the diversity of RSS-BRE classifiers get higher relative to the diversity of the random subspace ensembles that are based on selected voxels as the number of clusters/voxels increase. The inflection point is between 1050 to 1700 voxels/clusters (Figures 4.7, 4.8, and 4.9).

4.1.2.4 TOL Dataset

The classifier diversities of RSS-BRE and RSE, which are measured by Q-Statistic and the Disagreement Measure, averaged for all Subjects of the TOL dataset given in Figure 4.10. Here, the lower the Q-Statistic and the higher the disagreement measure, the diverse the ensemble of classifiers. The classifier ensembles are formed using sets of supervoxels (RSS-BRE) and voxes (RSE) where number of voxels/supervoxels are specified at the bottom of every graph.

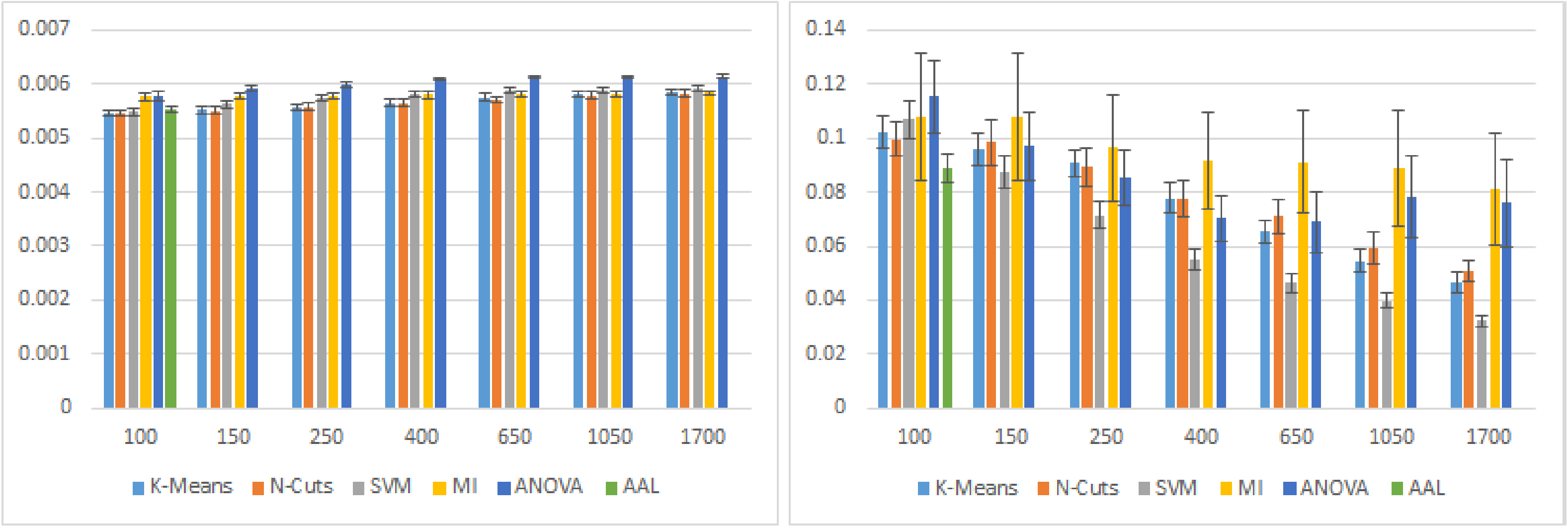
For this dataset when the disagreement measure is considered, the diversity of RSS-BRE get higher relative to the diversity of the random subspaces that are based on voxels selected by SVM. The inflection point is 100 voxels/supervoxels. For the other voxel selection methods, diversity measured by disagreement measure is higher than RSS-BRE. However, that result is due to the nature of disagreement measure, which does not take misclassified samples into consideration, where the increase in the diversity is the result of inaccuracy of the classifiers in RSE when MI or ANOVA is used for voxel selection (see the next section for the classification results). When the Q-Statistic is considered, where the effect of misclassified samples are taken into consideration, RSE based on voxels selected by MI and ANOVA have a less diverse set of classifiers compared to RSS-BRE (Figure 4.10).

4.2 Decoding Mental States by BRE

In this section we present the implementation details and the performance of BRE on the aforementioned datasets. First of all, we briefly mention the data preparation for clustering and classification. Second, we specify the parameters for clustering algorithms, which are common for Objects, Emotion, and TOL datasets. Third, we briefly mention the classification methods we described in the previous chapter. Lastly, we present the classification results for individual datasets.

4.2.1 Data Preparation for Clustering and Classification

During the classification experiments, we use a cross-validation scheme where we separate the data in train, validation and test epochs. For each cross-validation, four out of the six experimental epochs are reserved randomly for training while one is set aside for validation and one is set aside for testing for Emotion and Objects datasets. TOL dataset only has four epochs per subject. Thus, for that dataset, two epochs are used for training and one each for validation and testing.



(a) (b)

Figure 4.10: Q-Statistic (column on the left, lower is better) and disagreement measure (column on the right, higher is better) for RSS-BRE built using supervoxels specified by clustering methods K-Means, N-Cuts, and AAL, and RSE built using selected voxels (SVM, MI, and ANOVA) for TOL dataset. The results are given for the average of all Subject. The number of selected voxels and generated supervoxels are the same for each method, which are given at the bottom of every graph (100 to 1700).

Using training, validation and test epochs, the data matrices $\mathbf{X}_{tr}$, $\mathbf{X}_{val}$, $\mathbf{X}_{te}$, as well as their respective label vectors $\mathbf{Y}_{tr}$, $\mathbf{Y}_{val}$, and $\mathbf{Y}_{te}$ are formed.

Clustering for the specification of the supervoxels is performed only using the samples in the training set of voxel intensity values $\mathbf{X}_{tr}$, and their respective class labels $\mathbf{Y}_{tr}$.

4.2.2 Clustering Parameters and Methods for Generating Supervoxels

Recall that the preliminary step of BRE is to partition the brain volume into a set of regions called supervoxels. This task is achieved by using several clustering algorithms, namely K-Means and N-Cuts. In order to investigate the effect of the number of supervoxels on BRE, we experimented with each clustering algorithm with an array of clustering parameters θ. The baseline for the number of supervoxels we choose is in the order of the number distinct brain regions specified by AAL ($\sim$100). Using that baseline we then proceeded to increase the number of supervoxels at each clustering level by using a Fibonacci sequence in order to have distinct supervoxels at each level, while searching the space of different clustering parameters.

K-Means clustering is done by using one minus Pearson correlation $1-\rho_{ij}$, as the distance metric between the voxel pairs i and j. The clustering was performed for target number of final clusters equals to $\theta \in \{100, 150, 250, 400, 650, 1050, 1700\}$.

Similarly, spatially constrained N-Cuts clustering was applied to the distance graph that is formed by the similarity metric $\rho_{ij} - 1$. This time however, the similarities of non-adjacent voxels (in three dimensions) were set to 0 [21], in order to test the effect of enforcing spatial proximity while forming the voxel clusters. The clustering was performed for target number of final clusters $\theta \in \{100,\ 150, 250, 400, 650, 1050, 1700\}$.

The primary difference between K-Means and N-Cuts clustering algorithms that we use in this study is their approach to the spatial connectivity of the individual voxels. N-Cuts clustering requires the voxels in a cluster to be a sub-graph of the voxel graph where each voxel is connected to its spatial neighbors in the 3D voxel space. K-Means clustering does not take spatial proximity into account. Therefore, the supervoxels generated by K-Means are more likeliy to be equal in size and have ellipsoidal shapes in the domain specified by the pairwise voxel similarities, due to the Gaussian misture assumption of the algorithm [41].

In addition to the tests with the clustering algorithms, we employed automated anatomical labeling (AAL) to the fMRI data and used the generated regions to set a baseline for our cluster analysis.

4.2.3 Classification Parameters and Methods for the Validation of BRE

In the classification experiments, our primary aim is to validate the classification performances of BRE based classification strategies with respect to the methods that are based on voxel, and ROI selection.

In this section, we present a brief reminder for the classification strategies that are based on BRE and the methods that use voxel selection.

4.2.3.1 Classification Strategies Based on BRE

In order to test the validity of the proposed computational model BRE, we use three different approaches. In the first approach, we apply BRE to the sets of supervoxels C^{θ} each of which are generated by a specific clustering method that uses a particular clustering parameter θ. Given an array of such clustering parameters $\theta \in \Theta$, a BRE for each clustering parameter is built. The classification performance of the best performing BRE is then reported as the BRE that uses the set of supervoxels generated with the optimal clustering parameter ($BestParamBRE$), using the Algorithm 3.5. This approach sets a baseline for the performance of BRE, where all of the supervoxels generated by a brain partitionining process is used for the construction of a BRE.

Similarly, a BRE is built for the brain regions specified by AAL (C^{AAL}) in order to observe the performance of a BRE that uses AAL regions, and compare it to the performance of BREs that use supervoxels. The classification accuracy for this method is then reported ($BRE - AAL$) for the test set samples.

The second approach that we use is to form random subsets of supervoxels (RSS) within the set of supervoxels that are specified by a clustering parameter θ, and use RSS for building one RSS-BRE for each clustering parameter (see Section 3.4.2.2). The classification accuracy of the RSS-BRE for the optimal clustering parameter is then reported ($BestParamRSS - BRE$) for the test set.

For the brain regions specified by AAL, one RSS-BRE is built and the classification accuracy is reported for the test set samples ($RSS - BRE - AAL$)

The third approach is to use the set of all supervoxels $C = \bigcup_{\theta \in \Theta}(C^{\theta})$, which are generated by a clustering algorithm that uses the set of available clustering parameters $\Theta = \{\theta_1, \theta_2, ...\theta_{N_\theta}\}$, where N_θ is the number of clustering parameters. With this approach, random subsets of supervoxels are selected from within C, and RSS-BRE is built (see Section 3.4.2.1). The classification accuracy of the predictions obtained by RSS-BRE for the test set is then reported ($RSS{-}BRE$).

In addition to the methods that use BRE, we have selected the base layer classifier that performs best, where each of them receives input from a specific supervoxel, or AAL region. We report the classification accuracy of the best base layer classifier for supervoxels generated by each clustering method ($BestSV$), as well as AAL regions ($BestAAL$). We use these results in order to compare the performance of our methods with respect to the most relevant region of interest that can either be formed by clustering, or by using AAL.

4.2.3.2 Classification Strategies Based on Voxel Selection

The voxel selection based MVPA algorithms are implemented with $N_s \in \{100, 200, 300, 500, 800, 1300, 2100, 3400, 5100\}$ where N_s is the number of selected voxels. After the feature selection phase, random subspaces of voxels are formed from the set of selected voxels. Recall that each random subspace of voxels is formed by randomly sampling voxels without replacement, where using each random subspace, an SVM classifier is trained. Majority voting is then performed for the outputs of the SVMs for each set of selected voxels and classification accuracy for the RSE for the given set of voxels are acquired. The best set of selected voxels is determined using the classification performance of RSE on the validation set. For all classifiers, the regularization parameter $\phi \in \{0.1, 1, 10\}$ for the SVM classifiers is determined by using train and validation sets which are then applied to the test set. The accuracy results for test set ($BestSet$) are reported for the number of selected features N_s that achieves the highest accuracy in the validation set. For each voxel selection algorithm (SVM, MI, and $ANOVA$), this process is repeated.

4.2.4 A Comparative Analysis of BRE and Baseline Methods on Objects Dataset

Objects dataset is a two class dataset as explained in the Section 2.2.1. For this dataset, the average classification accuracies of the test samples are presented in Table 4.1. In the following subsections we provide our analysis of individual methods using this dataset.

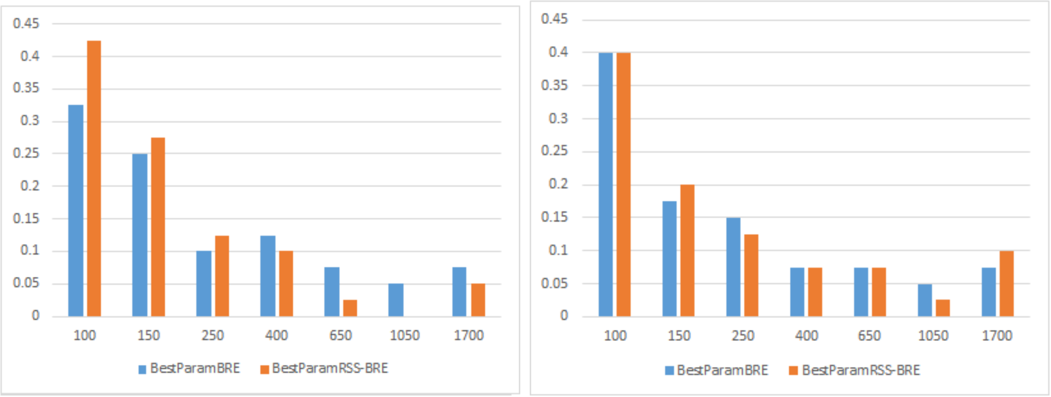
4.2.4.1 BRE Performances Obtained by Different Sets of Supervoxels

For this dataset, the overall performance of the BRE that receives inputs from the supervoxels formed by K-Means clustering in terms of accuracy is slightly superior to those formed by N-Cuts clustering. For both of those methods, the performance of BRE for the set of supervoxels that are generated by the best clustering parameter is ($BestParamBRE$) significantly better than the regions specified by AAL ($BRE - AAL$) (90.69%, 90.35% vs. 87.02% overall accuracies respectively). These results suggest that the AAL parcellation do not necessarily provide the best set of supervoxels for BRE for this dataset. A similar result can be observed with $BestParamRSS - BRE$ formed using the optimal clustering parameter for clustering methods, and with RSS-BRE formed using AAL regions ($RSS - BRE - AAL$).

In general, the clustering algorithm that does not take voxel locations into account (K-Means) provides slightly better sets of supervoxels for classification purposes than the algorithm that ensure spatial connectivity within the voxel

Table 4.1: The average accuracies for 10 validation runs are presented via per subject basis for the Objects dataset. The maximum accuracy acquired for each subject is marked with bold text.

		Subject1		Subject2		Subject3		Subject4		Overall	
		Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Rank
N-Cuts	BestParamBRE	88.61%	2.01%	92.78%	1.39%	86.39%	2.47%	93.61%	1.55%	90.35%	5
	RSS-BRE	88.33%	2.06%	94.44%	1.17%	**87.78%**	2.87%	94.72%	1.40%	91.32%	2.75
	BestParamRSS-BRE	86.67%	2.44%	93.33%	1.26%	85.56%	2.98%	**95.00%**	1.75%	90.14%	4.5
	BestSV	79.17%	2.85%	80.56%	2.84%	68.89%	2.89%	85.83%	2.28%	78.61%	13
K-Means	BestParamBRE	88.33%	1.54%	92.78%	1.77%	87.22%	2.59%	94.44%	1.31%	90.69%	4
	RSS-BRE	89.44%	1.80%	**95.00%**	1.08%	86.94%	3.68%	**95.00%**	1.48%	**91.60%**	**1.75**
	BestParamRSS-BRE	89.17%	1.92%	92.78%	1.67%	86.11%	2.87%	94.44%	1.10%	90.63%	4.25
	BestSV	81.67%	2.49%	86.39%	3.60%	72.50%	1.52%	82.22%	1.95%	80.69%	12.25
AAL	BREAAL	85.56%	1.71%	93.61%	1.06%	80.83%	2.29%	91.67%	1.59%	87.92%	7.5
	RSS-BRE-AAL	80.83%	1.21%	92.78%	0.93%	80.00%	1.75%	92.22%	1.34%	86.46%	9.75
	BestAAL	74.44%	2.56%	77.22%	1.90%	71.67%	2.04%	84.72%	2.60%	77.01%	13.5
SVM + RSE	BestSet	82.50%	1.40%	88.61%	1.09%	77.78%	2.43%	92.78%	1.25%	85.42%	9.75
MI + RSE	BestSet	84.44%	1.55%	89.17%	1.47%	76.94%	3.16%	92.78%	1.41%	85.83%	9.75
ANOVA + RSE	BestSet	**89.72%**	2.19%	90.00%	1.95%	83.89%	3.72%	93.89%	1.80%	89.38%	5.75



(a) N-Cuts (b) K-Means

Figure 4.11: The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and all 4 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for Objects dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS-BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter.

clusters (N-Cuts). In other words, there are supervoxels composed of fuctionally correlated voxels that are spatially separated, which make up a more suitable basis for classification than supervoxels for which the spatial connectivity is enforced for the voxels make up them.

In the next group of experiments, we examine the optimal number of supervoxels which yielded the best decoding performance in $BestParamRSS-BRE$ method and $BestParamBRE$ method. Figure 4.11 shows the distribution of the optimal clustering parameters over all classification experiments for two different clustering algorithms (N-Cuts and K-Means) and two different classification methods ($BestParamBRE$, and $BestParamRSS-BRE$) is given. For the both clustering methods (N-Cuts and K-Means), we see a dominance of lower number of supervoxels, which indicates that setting the number of supervoxels at around the number of AAL parcels is effective. Still, we cannot ignore the distribution of the optimal clustering parameter over all parameter choices for different cross validation runs and different subjects meaning that clustering with multiple parameters is a correct approach.

When we consider the $BestParamBRE$ accuracies for N-Cuts, K-Means, and meta classifier accuracy ($BRE-AAL$) for AAL regions we can see that unsupervised clustering is much more effective for generating supervoxels that are useful for BRE than only using AAL regions. Furthermore, the classification performance of the supervoxel with the optimal performance ($BestSV$) is higher for the supervoxels that are generated by the clustering algorithms when compared to supervoxels specified by AAL $BestAAL$. This finding suggest that clustering can be used to specify some supervoxels that are highly correlated with the specific mental states in the fMRI experiment.

4.2.4.2 BRE Performances Obtained by Random Subsets of Supervoxels

For the Objects dataset, RSS-BRE formed with the set of supervoxels that are generated by the optimal clustering parameter ($BestParamRSS - BRE$) performed on par with the BRE formed using the set of supervoxels generated by the optimal clustering parameter ($BestParamBRE$) for both N-Cuts and K-Means clustering. However, random subsets of supervoxels that are generated from all supervoxels ($RSS - BRE$) slightly outperforms the methods that depend on clustering parameters ($BestParamRSS - BRE$, $BestParamBRE$). That result suggests that a combination of supervoxels, which are generated by different clustering paramenters, can be better for building classifier ensembles than a selected set of voxels.

4.2.4.3 Comparison of BRE with Voxel Selection

In Objects dataset, $RSS-BRE$ based on K-Means clustering outperform all of the voxel-selection strategies 4.1 with a significant margin. Also all BRE based methdos that use supervoxels generated by clustering significantly outperforms voxel selection methods that use MI or SVM. Even a single meta classifier trained using the supervoxels formed by AAL regions performed better than voxel selection with SVM and MI. Thus, for this dataset we can conclude that supervoxel based BRE is effective for classification of mental states, which can be further improved by using supervoxels generated by clustering instead of using AAL regions.

4.2.5 A Comparative Performance Analysis of BRE and Baseline Methods using Emotion 2 Classes Dataset

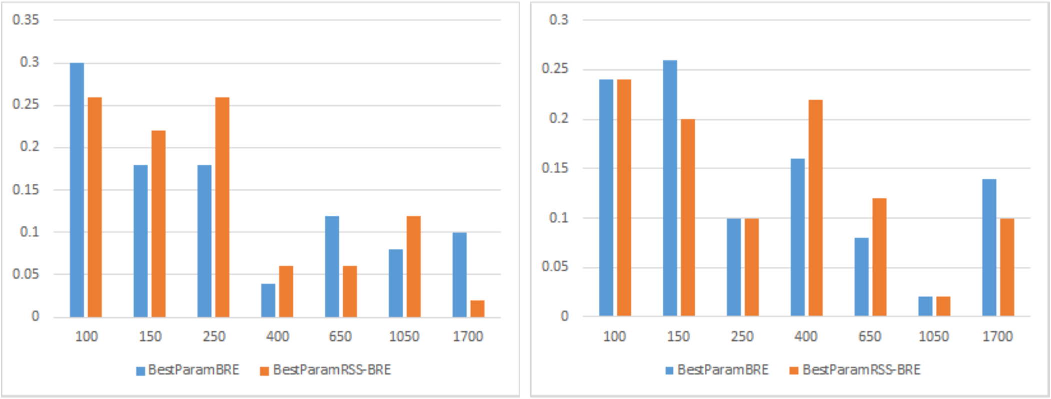
Emotion dataset can be used either as a 2-class, or a 4-class brain decoding problem that contains scans regarding participants viewing either emotionally stimulating images (fear, or disgust inducing) or non stimulating ones (furniture and kitchen appliances). In this section we present the classification results for the 2-class case for emotionally stimulating, or non stimulating images. In the following subsections we provide our analysis regarding the effects of individual methods on the brain decoding performance for this dataset.

4.2.5.1 BRE Performances Obtained by Different Sets of Supervoxels

For this dataset, $RSS - BRE$ formed using the supervoxels generated by the K-Means clustering has relatively better performance, compared to N-Cuts clustering by a small margin. This result supports the case that clustering with functional correlation similarity metric alone can provide better classification results than clustering that also takes spatial proximity into account [21].

Table 4.2: The average accuracies for 10 validation runs are presented via per subject basis for the Emotion dataset with two classes (emotional - non emotional). The maximum accuracy acquired for each subject is marked with bold text.

		Subject1		Subject2		Subject3		Subject4		Subject5		Overall	
		Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Rank
N-Cuts	BestParamBRE	74.71%	3.76%	65.71%	3.33%	72.43%	2.90%	84.86%	3.26%	82.71%	2.93%	76.09%	7.4
	RSS-BRE	75.43%	3.65%	67.57%	3.37%	74.43%	2.71%	86.57%	3.80%	**85.29%**	2.59%	77.86%	3.4
	BestParamRSS-BRE	73.86%	3.84%	68.71%	3.15%	73.86%	2.70%	84.57%	2.81%	82.86%	3.06%	76.77%	6.6
	BestSV	62.14%	2.95%	62.71%	2.92%	63.86%	2.61%	71.71%	1.85%	68.14%	2.38%	65.71%	13.6
K-Means	BestParamBRE	73.71%	3.33%	**69.86%**	3.32%	74.29%	2.16%	86.29%	1.78%	84.00%	1.95%	77.63%	4.2
	RSS-BRE	77.29%	3.44%	65.86%	3.85%	74.00%	3.15%	**86.71%**	3.63%	83.71%	2.93%	77.51%	4.2
	BestParamRSS-BRE	75.29%	3.30%	69.43%	3.29%	74.86%	2.75%	86.29%	2.08%	84.00%	2.58%	**77.97%**	**3.2**
	BestSV	60.57%	1.92%	59.43%	1.82%	66.86%	1.79%	80.29%	2.23%	72.57%	2.62%	67.94%	13.0
AAL	BRE-AAL	72.29%	3.03%	62.86%	3.03%	**77.86%**	2.33%	85.00%	1.97%	80.86%	3.58%	75.77%	7.4
	RSS-BRE-AAL	70.14%	2.94%	63.71%	2.94%	74.57%	2.06%	85.00%	0.93%	80.43%	2.70%	74.77%	8.4
	BestAAL	70.86%	3.23%	65.00%	3.23%	77.29%	2.07%	85.57%	1.43%	80.14%	3.10%	75.77%	7.4
SVM + RSE	BestSet	**77.71%**	3.65%	65.57%	3.65%	71.57%	3.14%	86.43%	3.51%	81.43%	2.83%	76.54%	5.8
MI + RSE	BestSet	77.43%	2.73%	68.71%	2.73%	66.29%	2.74%	82.57%	4.20%	79.29%	2.87%	74.86%	7.8
ANOVA + RSE	BestSet	68.43%	3.13%	64.86%	3.59%	65.86%	3.63%	82.00%	3.88%	73.43%	3.94%	70.91%	11.8



(a) N-Cuts (b) K-Means

Figure 4.12: The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and all 5 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for Emotion 2 Class dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS - BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter.

Figure 4.12 shows the relative frequency of clustering parameters to be selected optimal with respect to the classification performances of BRE and $RSS-BRE$ that use the set of supervoxels generated by the clustering parameters. In the figures, we can see that the optimal clustering parameters are more uniformly distributed across all clustering parameters for K-Means clustering when compared to the N-Cuts clustering. As the number of supervoxels gets closer to number of AAL regions the parameters has a higher frequency to be selected as optimal. When compared to subspaces generated using AAL regions, optimal clustering levels for the two suggested methods $BestParamBRE$, and $BestParamRSS - BRE$ had higher performance when K-Means clustering is considered. For the spatially constrained N-Cuts clustering, $BestParamBRE$ performance is similar with $BRE - AAL$ performance of AAL regions, while $BestParamRSS - BRE$ performance is significantly higher than $RSS - AAL$ performance of AAL regions. We suggest that these results are correlted with the increased diversity between base layer classifiers when the number of supervoxels generated by clustering increases (Figure 4.1).

4.2.5.2 BRE Performances Obtained by Random Subsets of Supervoxels

For the this dataset, RSS-BRE using the set of supervoxels generated by the optimal clustering parameter $BestParamRSS - BRE$ yielded the best overall performance. Also, random subsets of supervoxels that are sampled from within the set of all supervoxels ($RSS-BRE$) performed similarly for K-Means clustering. For N-Cuts clustering the best result is achieved by $RSS - BRE$

method. These results show the effectiveness random subsets of supervoxels for building brain region ensembles using the supervoxels generated by clustering algorithms. For the supervoxels specified by AAL regions, this effect cannot be observed.

4.2.5.3 Comparison of BRE with Voxel Selection

For this dataset, $BestParamRSS - BRE$ based on K-Means clustering outperform all of the voxel-selection strategies (Table 4.2). Also, brain region ensembles that are trained using supervoxels generated by the optimal clustering parameters $BestParamBRE$ and BRE that use AAL regions $BRE - AAL$, and $RSS - BRE$ with K-Means and N-Cuts clustering are able outperform two of the voxel selection strategies (ANOVA and MI) by significant margins. Only, voxel selection with SVM provided accuracy results comparable to our suggested methods.

4.2.6 A Comparative Performance Analysis of BRE and Baseline Methods using Emotion 4 Classes Dataset

In this subsection we present the classification results for the 4-class case where we decode fear, or disgust inducing images, or images containing kitchen appliances, or furniture. In Table 4.3, classification results for all 5 participants are presented. In the following subsections we provide our analysis regarding the effects of individual methods.

4.2.6.1 BRE Performances Obtained by Different Sets of Supervoxels

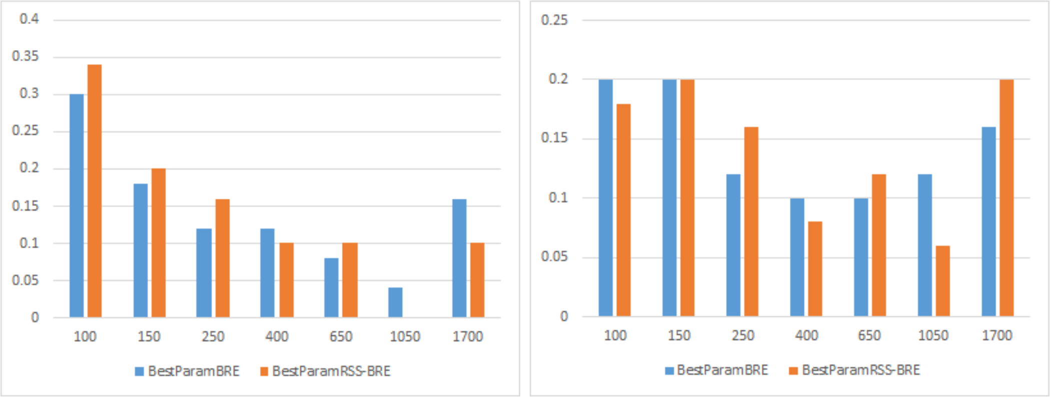
For this dataset, similar to the 2 class case, $RSS - BRE$ formed using supervoxels generated by K-Means clustering provides the best performance, followed by N-Cuts clustering, providing further support for the use of functional relations between the voxels when forming supervoxels.

Figure 4.13 shows the distribution of the optimal clustering parameters for K-Means and N-Cuts clustering algorithms with two different ($BestParamBRE$, and $BestParamRSS - BRE$) classification methods. In the figures, we can see that the most effective clustering parameters are almost uniformly distributed across all clustering parameters for K-Means clustering, where supervoxels formed by N-Cuts clustering yielded better results when they are larger, which would explain better performance of $RSS - BRE$ using the supervoxels generated by all clustering parameters for K-Means clustering. Also, when compared to N-Cuts clustering, supervoxels formed by K-Means clustering are more effective for forming $BestParamBRE$ and $BestParamRSS - BRE$.

When the AAL regions are considered, it can be seen that BRE formed using AAL regions ($BRE - AAL$) perform better than $BestParamBRE$, and

Table 4.3: The average accuracies for 10 validation runs are presented via per subject basis for the emotion dataset with four classes (furniture - kitchen appliances - fear - disgust). The maximum accuracy acquired for each subject is marked with bold text.

		Subject1		Subject2		Subject3		Subject4		Subject5		Overall	
		Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Accuracy	Rank
N-Cuts	BestParamBRE	55.86%	4.16%	43.43%	3.37%	43.57%	3.09%	60.71%	4.22%	57.00%	3.17%	52.11%	7.6
	RSS-BRE	55.00%	3.98%	46.43%	3.70%	46.71%	3.12%	64.86%	5.75%	63.57%	3.44%	55.31%	2.8
	BestParamRSS-BRE	53.71%	3.63%	44.86%	3.74%	43.29%	2.46%	60.86%	4.40%	57.71%	2.78%	52.09%	6.8
	BestSV	38.00%	3.81%	38.00%	2.44%	35.00%	2.46%	42.86%	2.49%	44.57%	4.21%	39.69%	13.8
K-Means	BestParamBRE	52.29%	4.73%	43.43%	3.04%	48.00%	3.01%	61.43%	5.35%	62.57%	3.84%	53.54%	5.6
	RSS-BRE	**56.00%**	5.11%	**47.86%**	3.99%	**49.86%**	3.64%	**66.71%**	5.57%	**64.00%**	4.16%	**56.89%**	**1.0**
	BestParamRSS-BREE	53.00%	4.39%	45.29%	3.13%	48.00%	3.43%	62.71%	4.74%	61.14%	3.07%	54.03%	4
	BestSV	42.57%	3.32%	38.57%	1.94%	42.00%	1.80%	46.29%	2.52%	50.86%	2.41%	44.06%	12.4
AAL	BRE-AAL	52.29%	3.76%	44.57%	3.25%	43.86%	3.19%	63.00%	3.14%	59.71%	3.68%	52.69%	5.8
	RSS-BRE-AAL	52.00%	2.90%	43.71%	2.49%	42.43%	2.99%	61.57%	3.10%	56.43%	2.22%	51.23%	8.8
	BestAAL	51.57%	3.56%	44.71%	2.66%	43.29%	2.87%	61.71%	2.93%	57.43%	3.36%	51.74%	7.8
SVM + RSE	BestSet	54.14%	3.06%	42.00%	3.49%	47.00%	3.09%	60.00%	3.47%	62.00%	4.93%	53.03%	6.6
MI + RSE	BestSet	50.43%	4.01%	37.57%	1.95%	43.86%	3.03%	55.57%	3.72%	59.43%	3.30%	49.37%	10.2
ANOVA + RSE	BestSet	50.71%	4.49%	43.43%	3.14%	36.29%	2.63%	57.00%	4.84%	50.14%	3.81%	47.51%	11.2



(a) N-Cuts (b) K-Means

Figure 4.13: The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and all 4 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for Emotion 4 Class dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS-BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter.

$BestParamRSS-BRE$ formed by the sets of supervoxels generated by optimal clustering parameters of both N-Cuts, and K-Means clustering. This result is in parallel with the diversity analysis made for this dataset. For this dataset, none of the clustering methods could generate a set of supervoxels using a clustering parameter for which the diversity of the base layer classifiers is higher than that of AAL regions (Figures 4.7, 4.8, and 4.9).

When the nature of the dataset is considered, the results can be better interpreted. The brain decoding task for this dataset is to classify the brain patterns formed due to two visual object classes (furniture, and kitchen appliences), and two types of emotion inducing visual stimuli (fear, and disgust). It is known that emotional responses within the brain are widespread across multiple brain regions (amygdala, insula, preforontal cortex, parietal cortex) [88], while representations of visual object classes are expected to confined within primary visual areas and temporal cortex [25]. While emotions having large scale representations, visual object representations would be localized within small areas of the brain. That discrepancy would create a problem of scale when a decision for clustering parameter is to be made. On one hand, if the brain volume is divided into supervoxels that are too fine in scale, brain regions with representations for emotional processing would be oversegmented. On the other hand, if the size of the supervoxels is kept large, the representations concerning different visual aspects of the visual object classes (color, texture, shape, etc.) could be lost within a supervoxel. Thus, for this dataset it is only natural for $RSS-BRE$ method, which forms random subspaces of supervoxels from all clustering levels, to be the most accurate.

4.2.6.2 BRE Performances Obtained by Random Subsets of Supervoxels

For this dataset, random subsets of supervoxels that are sampled from the set of all supervoxels ($RSS-BRE$) performed better than all other methodologies. Moreover, the performance of RSS-BRE using the set of supervoxels generated by an optimal clustering parameter ($BestParamRSS-BRE$) yielded better overall accuracies than $BestParamBRE$ (which do not utilize random subsets). In this dataset, with four classes, using random subsets of supervoxels provide a significant improvement in classification performance for all cases, except for the supervoxels formed by AAL regions.

4.2.6.3 Comparison of BRE with Voxel Selection

In this dataset, $RSS-BRE$ that uses supervoxels generated by K-Means, and N-Cuts clustering algorithms outperform all of the voxel-selection strategies by significant margins (Table 4.3), and has the best overall performance. This result indicate the success of our methodology, where fusion of the distributed mental representations being more effective in brain decoding than a selected set of voxels.

4.2.7 A Comparative Performance Analysis of BRE and Baseline Methods using Tower of London (TOL) Dataset

In this subsection we present the brain decoding results for the TOL dataset. The dataset is generated using block design where planning and action phases for the solution of a Tower of London puzzle were recorded. In the following experiments the two phases are considered to be seperate classes. Every brain volume captured during a phase is used as a sample, and labelled with the corresponding class label. The data from 18 subjects are used in the following experiments.

4.2.7.1 BRE Performances Obtained by Different Sets of Supervoxels

Figure 4.12 shows the distribution of the optimal clustering parameters for BRE that uses sets of supervoxels specified by K-Means and N-Cuts clustering algorithms with two different classification methods ($BestParamBRE$, and $BestParamRSS-BRE$). In the figures, we can see that the optimal clustering parameters are uniformly distributed across all clustering parameters for the both clustering methods. As for the classification performances, BRE methods based on the supervoxels formed by clustering methods do not provide significantly better results than those specified by AAL regions (Table 4.7). Similar accuracies for the methods that use BRE regardless of the procedure that is used to specify the supervoxels suggest that individual AAL regions can already capture the essential components of the mental tasks.

Table 4.4: Average accuracy results over 6 cross-validation runs are presented individually for subjects 1-6 in the TOL dataset.

		Subject1		Subject2		Subject3		Subject4		Subject5		Subject6	
		Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste
N-Cuts	BestParamBRE	93.27%	0.84%	93.57%	0.94%	90.06%	1.73%	80.41%	2.75%	85.00%	2.17%	78.13%	5.61%
	RSS-BRE	**94.05%**	0.71%	94.54%	0.85%	88.30%	2.80%	82.26%	3.77%	82.81%	4.59%	78.44%	6.82%
	BestParamRSS-BRE	93.47%	0.83%	93.66%	0.67%	89.86%	1.45%	82.46%	3.11%	86.04%	2.32%	78.65%	5.32%
	BestSV	88.21%	2.58%	87.43%	1.11%	85.38%	1.14%	74.46%	7.45%	84.79%	2.06%	79.17%	4.18%
K-Means	BestParamBRE	92.69%	0.54%	93.66%	1.41%	88.11%	2.74%	83.82%	2.62%	84.90%	3.38%	74.48%	5.25%
	RSS-BRE	93.66%	0.63%	93.96%	1.45%	88.30%	3.65%	**83.43%**	3.46%	80.42%	5.46%	78.13%	5.60%
	BestParamRSS-BRE	92.69%	0.52%	93.86%	1.18%	89.67%	2.10%	83.24%	3.07%	85.73%	2.55%	78.13%	5.14%
	BestSV	86.26%	2.79%	88.89%	1.28%	82.85%	2.81%	79.04%	1.64%	84.79%	1.43%	77.19%	4.85%
AAL	BRE-AAL	89.96%	1.02%	94.54%	0.82%	88.40%	1.72%	78.27%	2.48%	78.65%	5.62%	78.75%	6.93%
	RSS-BRE-AAL	90.84%	0.88%	**94.93%**	0.90%	**90.16%**	1.20%	81.19%	2.83%	83.85%	3.76%	78.54%	6.83%
	BestAAL	85.87%	2.96%	84.31%	4.84%	87.04%	1.28%	71.25%	3.09%	81.67%	3.87%	80.63%	1.78%
SVM + RSE	BestSet	87.43%	1.72%	78.27%	6.63%	84.70%	1.92%	77.78%	5.61%	**86.56%**	2.01%	**82.60%**	2.44%
MI + RSE	BestSet	89.18%	0.75%	83.43%	5.93%	80.70%	5.31%	77.68%	3.55%	85.31%	2.25%	78.44%	4.47%
ANOVA + RSE	BestSet	87.23%	1.20%	81.97%	6.34%	80.51%	5.22%	68.42%	6.63%	81.46%	3.73%	51.35%	8.24%

Table 4.5: Average accuracy results over 6 cross-validation runs are presented individually for subjects 7-12 in the TOL dataset.

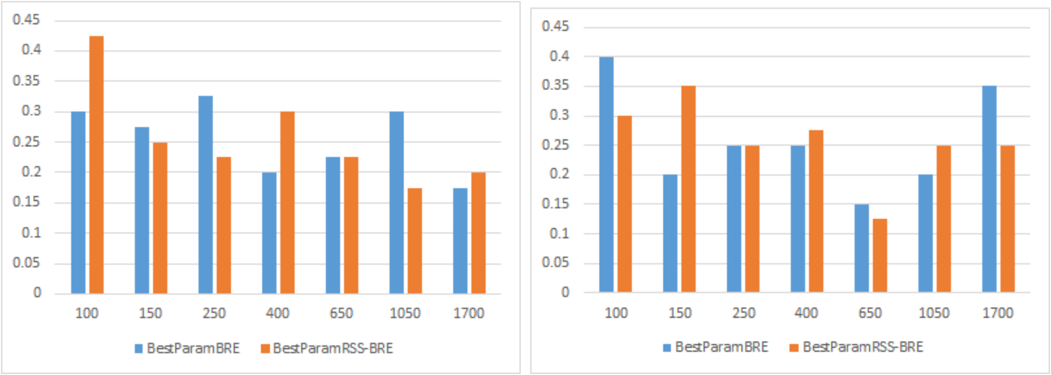
		Subject7		Subject8		Subject9		Subject10		Subject11		Subject12	
		Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste
N-Cuts	BestParamBRE	89.08%	2.71%	88.38%	1.95%	87.19%	0.75%	79.53%	12.58%	61.21%	4.31%	88.50%	2.95%
	RSS-BRE	89.77%	1.89%	**92.00%**	1.32%	84.90%	1.27%	78.85%	11.78%	58.67%	4.40%	88.11%	3.25%
	BestParamRSS-BRE	89.67%	2.74%	89.05%	1.86%	**88.33%**	0.55%	81.58%	11.13%	61.21%	2.89%	**89.38%**	2.59%
	BestSV	85.77%	4.73%	84.10%	3.02%	73.85%	4.28%	**87.52%**	2.27%	58.09%	2.75%	87.23%	1.19%
K-Means	BestParamBRE	89.57%	2.64%	84.38%	1.48%	84.17%	3.49%	82.65%	10.50%	**66.18%**	5.63%	88.89%	3.09%
	RSS-BRE	88.79%	3.47%	89.62%	1.65%	84.79%	2.54%	79.14%	10.75%	60.62%	5.46%	86.84%	3.57%
	BestParamRSS-BRE	89.96%	2.91%	87.62%	1.64%	86.35%	2.63%	77.29%	14.19%	65.79%	5.04%	88.79%	3.00%
	BestSV	84.80%	1.39%	74.29%	4.67%	76.77%	4.81%	81.97%	7.63%	64.42%	4.47%	78.46%	3.85%
AAL	BRE-AAL	88.89%	2.81%	89.62%	1.17%	86.15%	1.88%	76.02%	11.26%	65.59%	5.34%	84.21%	3.83%
	RSS-BRE-AAL	**90.45%**	2.09%	91.43%	0.99%	84.90%	2.50%	80.80%	8.96%	63.35%	3.41%	86.45%	3.23%
	BestAAL	83.53%	4.71%	79.52%	2.67%	69.79%	6.88%	82.94%	12.21%	63.45%	4.33%	84.60%	2.99%
SVM + RSE	BestSet	85.96%	3.54%	83.43%	3.15%	76.15%	5.56%	82.16%	10.60%	62.67%	3.83%	87.13%	3.01%
MI + RSE	BestSet	85.58%	3.32%	70.10%	7.52%	58.75%	5.30%	84.11%	4.56%	60.14%	4.71%	84.60%	3.22%
ANOVA + RSE	BestSet	80.02%	5.60%	68.38%	4.31%	52.40%	1.33%	66.37%	8.50%	60.04%	2.71%	80.80%	4.45%

Table 4.6: Average accuracy results over 6 cross-validation runs are presented individually for subjects 13-18 in the TOL dataset.

		Subject13		Subject14		Subject15		Subject16		Subject17		Subject18	
		Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste	Acc	Ste
N-Cuts	BestParamBRE	**85.83%**	1.71%	92.59%	1.36%	85.19%	3.49%	91.23%	0.92%	89.18%	0.74%	78.44%	2.73%
	RSS-BRE	83.13%	2.38%	92.30%	2.11%	83.53%	4.13%	91.52%	0.84%	87.82%	1.01%	76.56%	5.46%
	BestParamRSS-BRE	85.00%	2.22%	93.66%	0.90%	84.50%	2.77%	91.81%	1.24%	89.77%	0.78%	84.38%	2.51%
	BestSV	79.48%	3.36%	86.74%	1.60%	84.21%	1.88%	85.77%	1.48%	85.77%	1.06%	83.13%	2.25%
K-Means	BestParamBRE	81.88%	1.65%	92.59%	1.67%	85.28%	3.39%	91.42%	0.91%	89.47%	1.10%	77.92%	5.83%
	RSS-BRE	85.10%	1.65%	91.52%	2.58%	82.94%	4.64%	91.33%	0.80%	88.21%	0.80%	75.31%	6.79%
	BestParamRSS-BRE	80.63%	2.12%	92.79%	1.68%	**85.87%**	3.31%	91.42%	1.10%	89.57%	1.09%	80.31%	2.78%
	BestSV	79.90%	3.95%	85.28%	1.61%	**85.87%**	1.24%	85.77%	1.93%	87.72%	0.87%	75.94%	5.07%
AAL	MetaAAL	81.35%	2.38%	93.86%	1.01%	83.33%	2.53%	90.16%	0.86%	87.62%	1.66%	**90.73%**	0.67%
	RSS-BRE-AAL	82.29%	2.38%	**94.54%**	0.88%	84.02%	3.37%	**92.11%**	1.02%	89.86%	1.30%	87.60%	1.48%
	BestAAL	76.88%	3.90%	85.96%	3.15%	80.12%	2.33%	86.16%	1.94%	85.96%	1.69%	81.35%	4.47%
SVM + RSE	BestSet	80.63%	6.13%	92.01%	1.37%	82.46%	3.16%	83.04%	5.44%	**91.13%**	0.68%	89.06%	2.26%
MI + RSE	BestSet	69.58%	5.04%	88.60%	1.64%	83.04%	4.18%	86.06%	3.62%	88.21%	1.31%	51.67%	0.75%
ANOVA + RSE	BestSet	57.92%	2.59%	77.68%	3.03%	83.43%	3.35%	83.63%	1.62%	85.77%	4.67%	51.98%	0.83%

Table 4.7: Accuracy results averaged over all subjects and 6 cross-validation runs are presented for TOL dataset

		Overall	
		Accuracy	Rank
N-Cuts	BestParamBRE	85.38%	5.61
	RSS-BRE	84.86%	6.11
	BestParamRSS-BRE	**86.25%**	**3.83**
	BestSV	82.28%	8.89
K-Means	BestParamBRE	85.11%	5.56
	RSS-BRE	84.56%	7.22
	BestParamRSS-BRE	85.54%	4.78
	BestSV	81.12%	9.67
AAL	BRE-AAL	84.78%	6.83
	RSS-BRE-AAL	85.96%	4.44
	BestAAL	80.61%	9.94
SVM + RSE	BestSet	82.95%	7.89
MI + RSE	BestSet	78.07%	10.06
ANOVA + RSE	BestSet	72.19%	13.11



(a) N-Cuts (b) K-Means

Figure 4.14: The relative frequencies for clustering parameters to be selected as the optimal clustering parameter over all cross-validation runs and 18 subjects are presented for (a) N-Cuts and (b) K-Means clustering methods for TOL dataset. The optimal clustering parameters are selected with respect to the classification performance of BRE, and $RSS-BRE$ methods using a set of supervoxels specified by that particular clustering parameter. The numbers below each column denote the number of supervoxels generated using that clustering parameter.

4.2.7.2 BRE Performances Obtained by Random Subsets of Supervoxels

For this dataset, random subspace ensembles within the optimal clustering level (*RSELevel*) yielded the best overall performance for all clustering methods. For this dataset, using supervoxels that are formed using different clustering parameters was ineffective in terms of accuracy results. The performance of *RSE* method is worse than both *BestParamBRE*, and *RSELevel* methods.

4.2.7.3 Comparison of BRE with Voxel Selection

In this dataset, regardless of the particular method for the specification of supervoxels (via clustering, or based on AAL regions), suggested BRE methods perform significantly better than the voxel selection based methods.

For this dataset, inter subject difference in the accuracy ratings for the classification methods considered in this study is higher than the other datasets. Especially, random subspace ensembles based on voxels selected by ANOVA and MI algorithms have a high variance in their classification performace across subjects. For some of the subjects, they perform at the chance level classification accuracy (Figures 4.4, 4.5, 4.6). For this dataset, accuracy ratings below 58% were considered to be no different than chance level accuracy using the procedure explained in 3.6). Given that, RSE based on voxels selected with ANOVA fails to pass chance level for Subjects 6, 9, 13, and 18 and RSE based on voxels selected with MI fails to pass the chance level for the Subjects 9 and 18. Since RSE based on voxels selected by SVM does not have such an issue, we can suggest that the univariate nature of voxel selection processes employed by MI and ANOVA to be the cause of them being failed.

The low performance of RSE based on voxels selected by ANOVA and MI is also hinted at the section where we compare the classifier diversities. In Figure 4.10 it can be seen that the diversity of RSE based on ANOVA and MI is high when the disagreement measure is considered. Whereas, when Q-Statistic is used, their diversity is measured below the other methods. That discrepancy is caused by the high number of samples misclassified by RSE based on voxels selected by these two methods, where Q-statistic is a diversity measure that is normalized by the frequency of the misclassified samples, while disagreement measure is not.

4.3 Selection of Brain Regions that Contribute to the Task of Classification of Mental States

Recall that, some of the brain regions may not participate to the processing or representation of the mental task or stimulus under consideration, but rather engaged by the other neural activities. One of the aims of this study is to select out the regions that are relevant to the tasks or stimuli specified by the fMRI experiments.

In order to select the brain regions that contribute to the classification of the mental states specified by the fMRI experiments, we eliminated the supervoxels for which, the corresponding base layer classifiers perform below the chance level classifier accuracy given the number of test samples (see Section 3.6). We call the remaining supervoxels to be effective in the classification task. The voxels in the remaining supervoxels are then grouped within AAL regions and the ratio of the effective voxels within each AAL region to number of effective voxels obtained from the whole brain are then calculated for the brain partitioning performed using a clustering parameter. The relative frequency of the number of effective voxels within a specific AAL region with respect to the number of all effective voxels from the whole brain is then plotted, where the relative frequencies obtained for each AAL region using different clustering parameters are stacked and plotted in the Figures (4.15, 4.16, 4.17, 4.18, and 4.19) of this section.

4.3.1 Comparison of the Regions Specified using Base Layer Classifiers with the Existing Neuroscience Literature with respect to the FMRI Datasets used in this Study

For the experiments with the objects dataset, we can observe that the regions that contribute to the classification task are the occipital lobe and surrounding areas including cuneus, calcerine cortex, and inferior and medial temporal lobes (Figures 4.15 and 4.16). These findings comply with the existing knowledge about the processing of visual stimuli where the visual information is processed through the occipital lobe and visual object classification is done at the temporal lobe [25]. Furthermore, our findings suggest that it is possible to decipher object specific information from the voxels that correspond to the *early* visual areas such as primary visual cortex. Also, we can observe that the visual areas that correspond to the *where* pathway (a pathway that is known to be involved with the processing of the spatial aspects of visual information) such as calcerine cortex also contains object specific information.

For the Emotion dataset, we observe that, there is a wide distribution of brain regions that are affected by the emotional stimuli in such a way that they respond differentially to emontion arousing stimuli (Figures 4.17 and 4.18). In contrast to Objects dataset, the most prominent regions for this classification task are the middle frontal gyri (Frontal_Mid_L, and Frontal_Mid_R in Figure **??**), where voxels in these regions lacked any discriminative information (Figures 4.15 and 4.16) for visual object classification task. On the other hand, the regions that are most effective in discriminating fear and disgust emotions (middle frontal gyrus - Frontal_Mid -, fusiform gyrus - Fusiform -, insula - Insula -, middle occipital gyrus - Occipital_Mid -, middle temporal gyrus - Temporal_Mid, middle occipital gyrus - Occipital_Mid, prenucleus - Prenucleus -) comply with the previous study on this very subject [88].

The relative voxel frequencies for TOL dataset indicate highly selective activity in sensory-motor areas (Precentral and Postcentral sulci), as well as supplementary motor areas, and some specific areas of cerebelum (Cerebelum 6, and Cerebelum Crus 1). Also, visual areas such as calcerine cortex, occipital cortex and cuneus are selective in discriminating planning and action phases of the experiment. Moreover, the cortical structures that are implicated in spatial processing are also involved (Parietal Cortex, Angular Gyrus). Furthermore, Middle Frontal Gyrus, Lingual Gyrus, Parietal Inferior Lobule, Medial Superior Frontal Gyrus, which are related to logical processing [79] is also found to be highly selective for this experiment. When compared to the original article [71], our algorithm seems to be more sensitive in the specification of regions with selective activity.

4.3.2 Comparison of the Regions Specified using Base Layer Classifiers with the Regions Specified with ANOVA with respect to the FMRI Datasets used in this Study

In order to compare our region specification procedure, we have used analysis of variance (ANOVA) to calculate p-values for each voxel. Then, we selected the voxels with $p \leq 0.01$ and plotted the relative frequencies of the selected voxels to be in an AAL region (Figures 4.20 to 4.24).

When we compare the regions selected by p-values with the regions selected by the base layer classifiers, for the Objects dataset we can observe that the most critical brain regions for the task are the same for the both methods. For the Objects dataset, there are some differences can be observed regarding the less critical (less frequently seleced) regions when the two methods are compared. The accuracy ratings of the base layer classifiers are more selective than ANOVA when this dataset is considered since the critical voxels are more spread out among the brain regions when ANOVA is used. This difference can be explained using the fact that the both of the stimulus classes are visual objects (bird-flower), where ANOVA selects the voxels that show significantly different activity than the baseline activity of the all voxels, whereas our method only selects the supervoxels that can discriminate the given classes. As long as the voxels are involved in processing the visual stimuli, ANOVA can specify the voxels to be effective, however, our method adds another constraint for a supervoxel to be effective, which is, to be able to discriminate the two visual stimuli that belong to different classes (Figures 4.15, 4.16, 4.20, and 4.21).

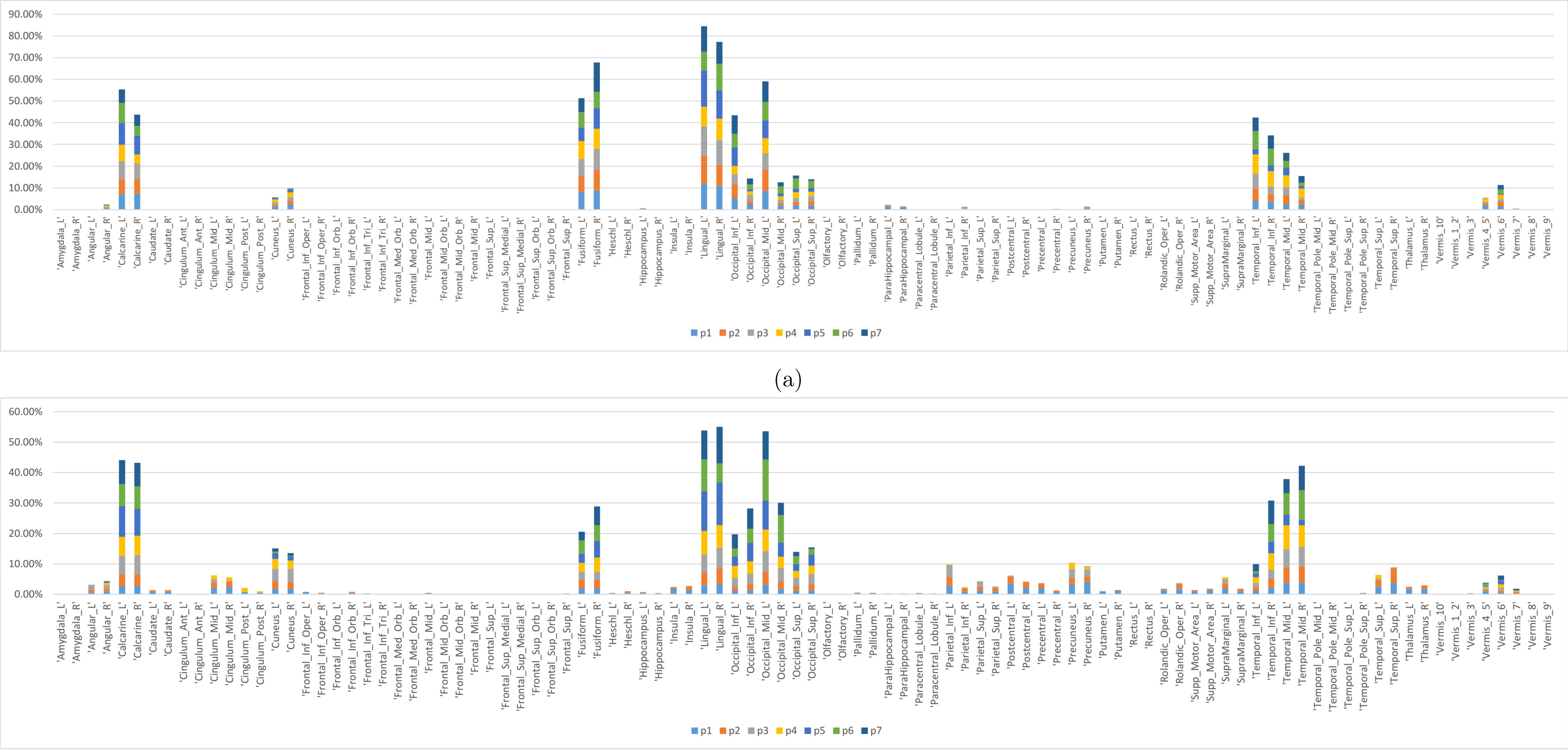
When the Emotion 2 Class dataset is considered, the relative voxel frequencies for the regions specified by the base layer classifiers that use supervoxels as their inputs correlate well with the relative voxel frequencies for the regions specified by ANOVA (Figures 4.17, 4.18, 4.22, and 4.23). One small difference of our suggested method with ANOVA is that, our method can show small differences across Subjects with respect to the relative voxel frequencies for the specified brain regions, while with ANOVA, the relative voxel frequencies for the brain regions stay almost the same across Subjects. In contrast to the Objects dataset, since the stimuli for the different classes (visual objects vs. emotional stimuli) used in this dataset are not expected to be processed/represented by the same regions of the brain, except for the visual cortex, thus, the correlation between our method and ANOVA is higher for this dataset.

For the TOL dataset, we calculated relative voxel frequencies for brain regions averaged across all 18 Subjects. When we compare the relative voxel frequencies of the effective voxels per region specified by the base layer classifiers with the relative voxel frequencies specified by ANOVA, we can observe highly correlated patterns (Figures 4.19 and 4.24).

When all datasets are considered, we can see that base layer classifiers that receive inputs from supervoxels can be used to identify the relative contributions of the brain regions to the classification task. The relative contributions of the individual brain regions as determined by base layer classifiers are similar to the ones determined using ANOVA, which is the current standard. Also, our method can be used to specifiy brain regions that contribute to the classification of the mental states while with ANOVA, all the regions that are involved in the processing or representation of the experimental stimuli are specified.

4.4 Chapter Summary

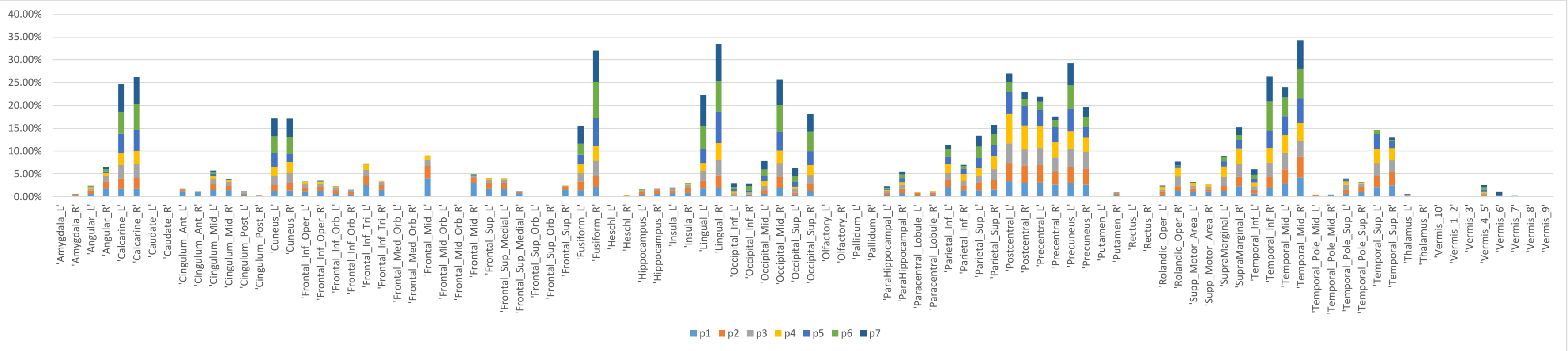
In this chapter, we have first analyzed the classifier diversity of BRE and R. The method we presented resulted in highly diverse sets of classifiers as the number of supervoxels increased, when compared to state of the art voxel selection methods. Also, we have shown that BRE is more effective for decoding mental states from brain images where the classification accuracy is higher than the state of the art methods. Finally, we compared the brain regions that are found to be selective across the tasks presented in the fMRI experiments. The supervoxels formed by functional clustering of the voxels were found to be a viable alternative to the methods that are used in the specification of the regions that are correlated with the experimental tasks.



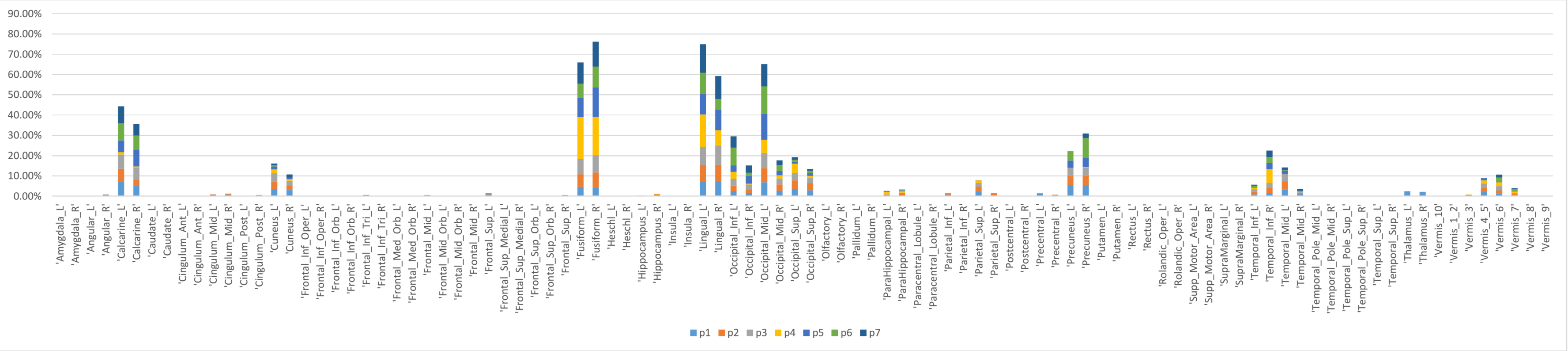
(a)

(b)

Figure 4.15: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 2 (b) of the Objects dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner.

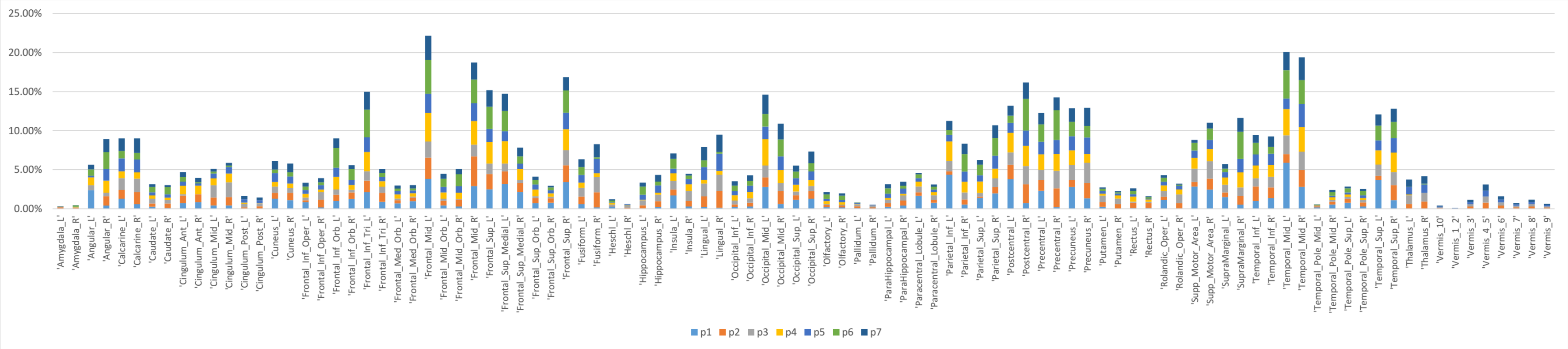


(a)

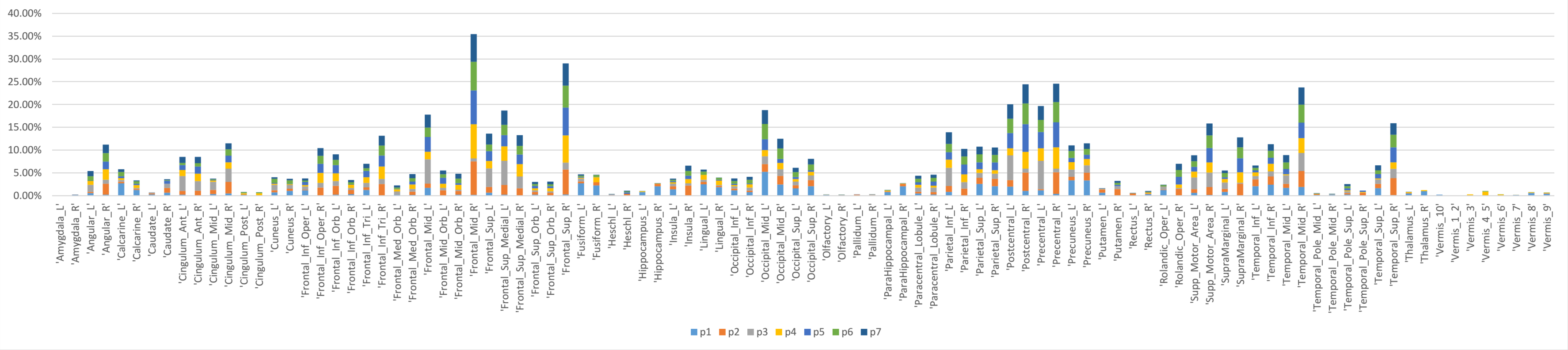


(b)

Figure 4.16: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 3 (a) and 4 (b) of the Objects dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner.

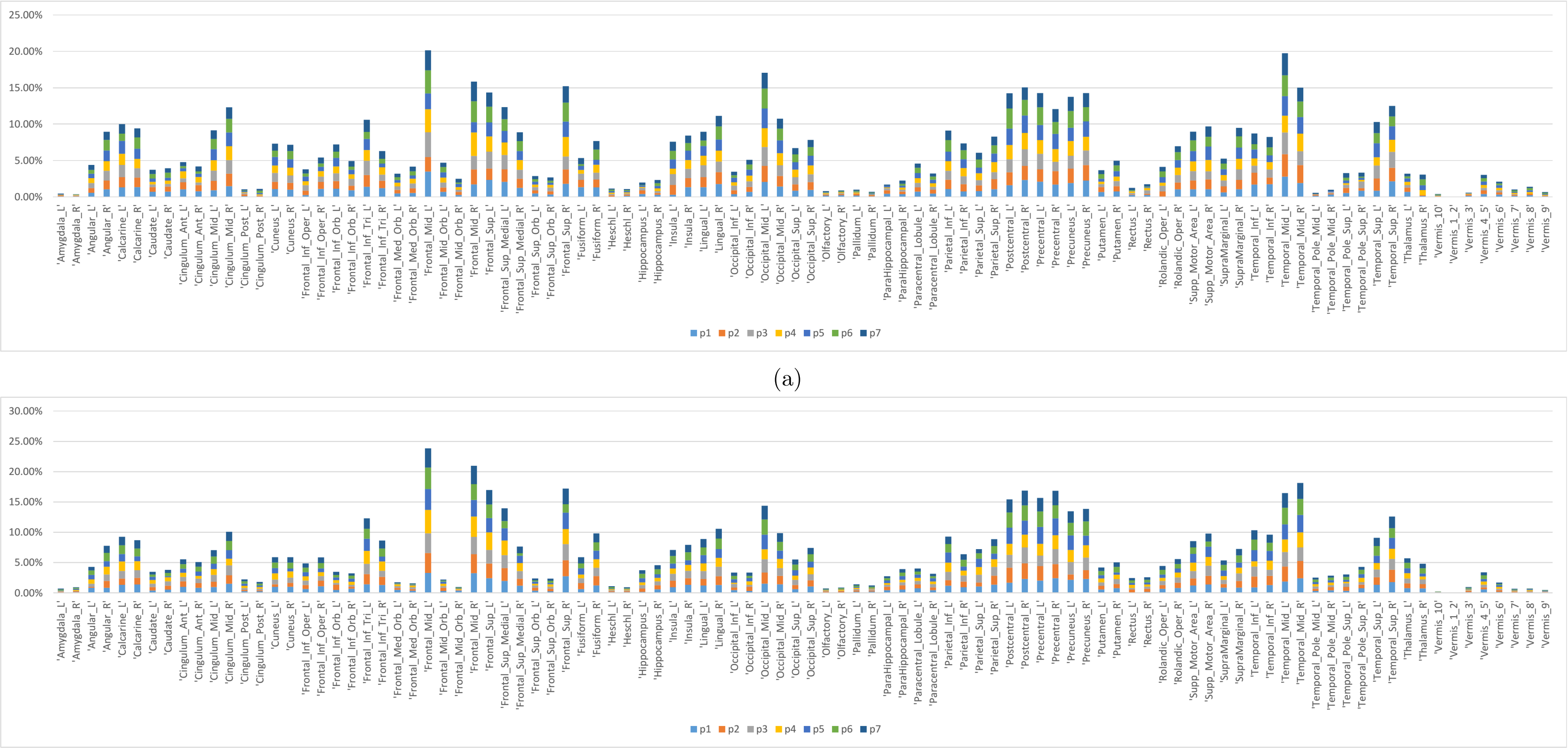


(a)



(b)

Figure 4.17: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 1 (b) of the Emotion 2 Class dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner.



(a)

(b)

Figure 4.18: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 1 (b) of the Emotion 2 Class dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner.

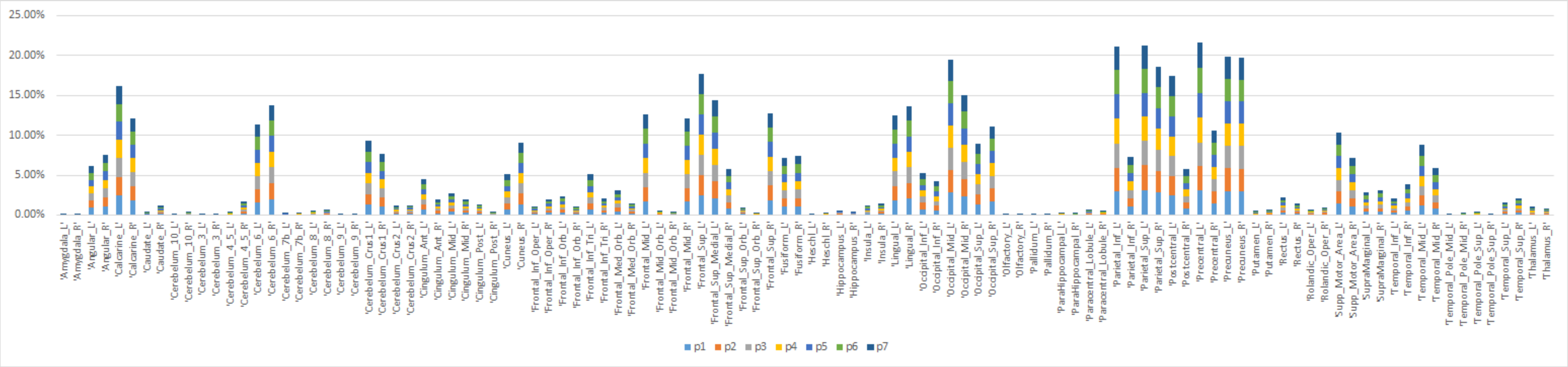
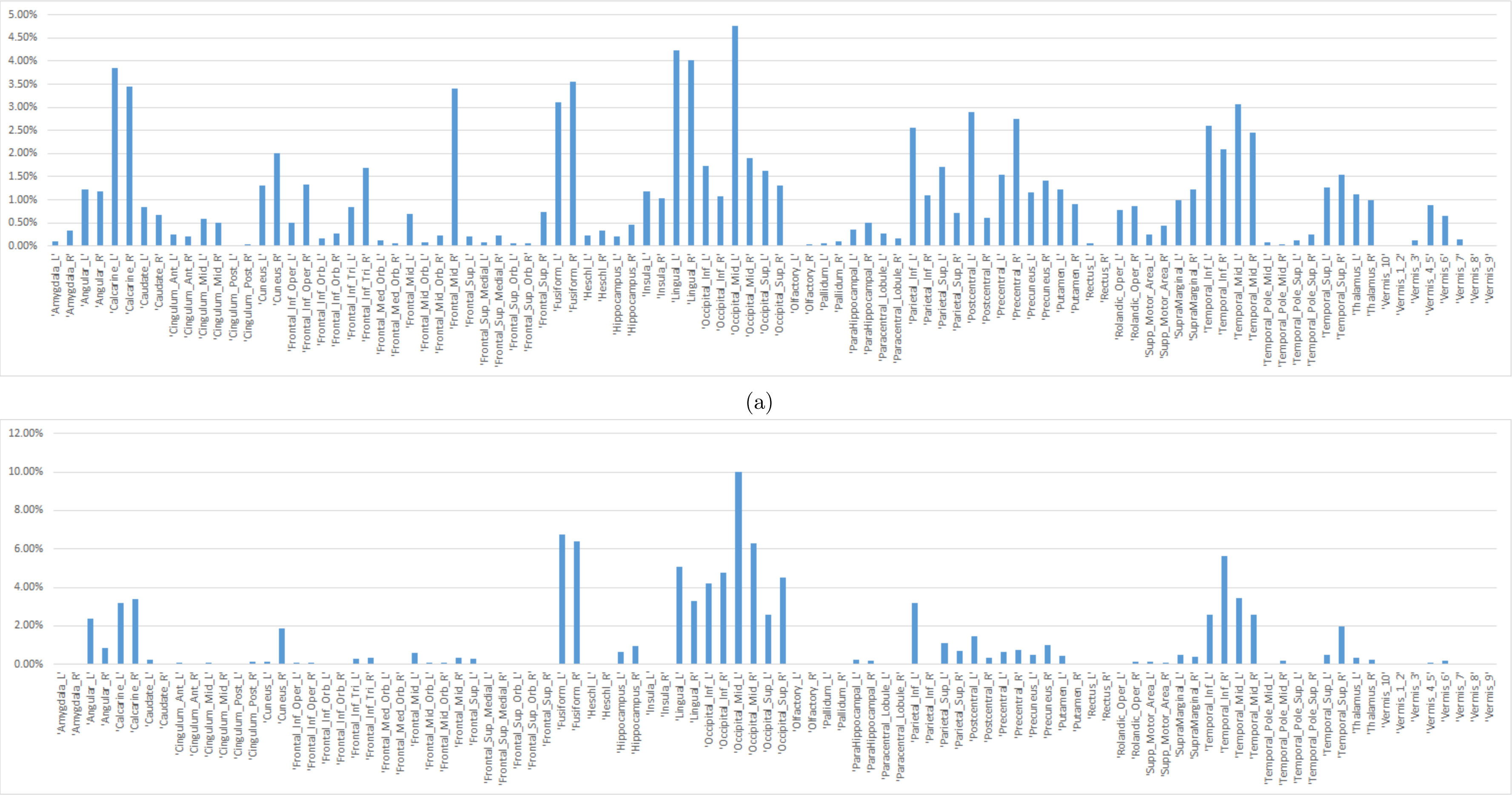


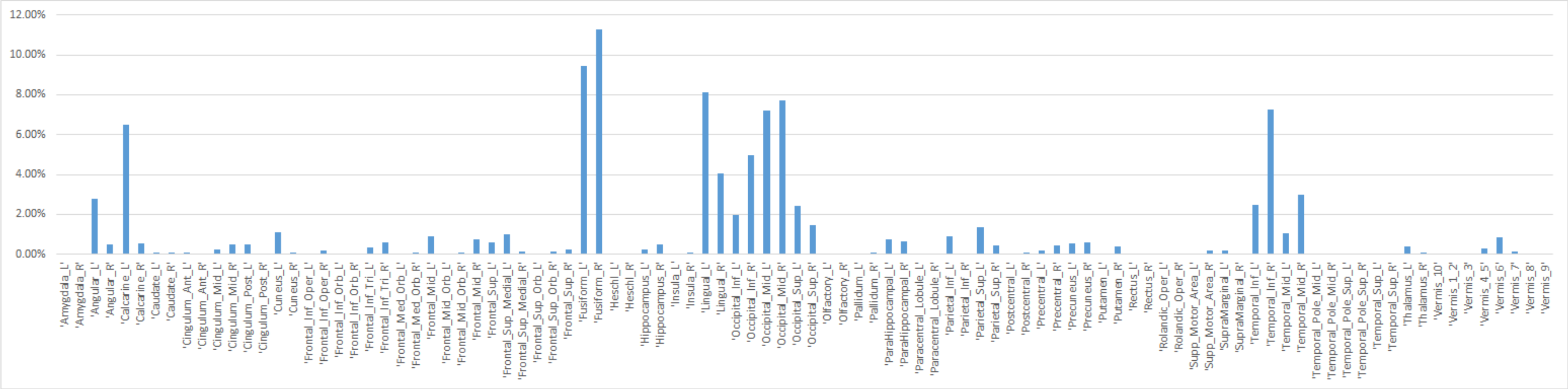
Figure 4.19: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for all subjects from TOL dataset are shown. The relative frequencies of effective voxels for each AAL region using different clustering parameters (p1 = 100 supervoxels to p7 = 1700 supervoxels) are plotted in a stacked manner that are averaged across all 18 Subjects.



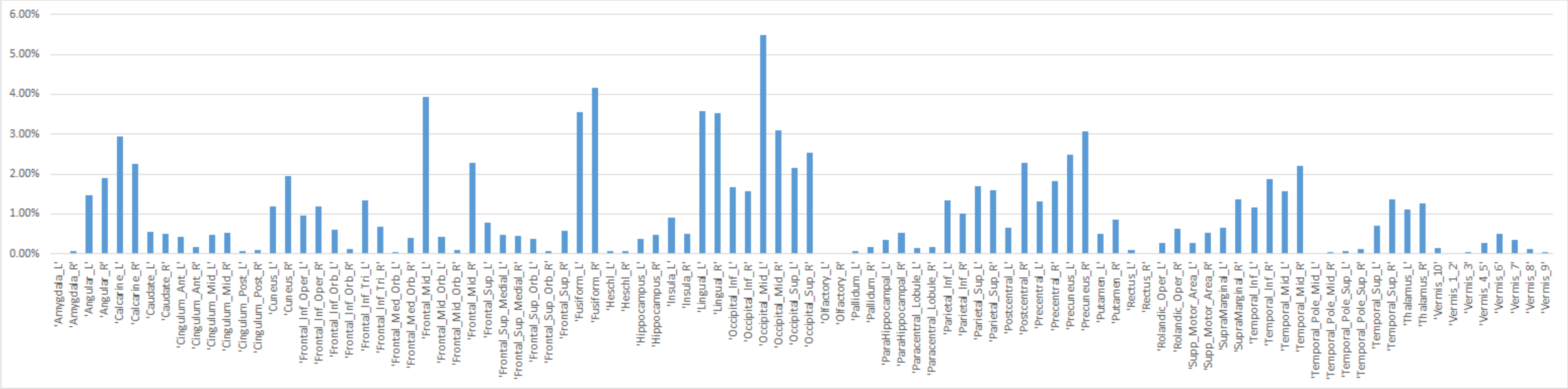
(a)

(b)

Figure 4.20: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 2 (b) of the Objects dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$.

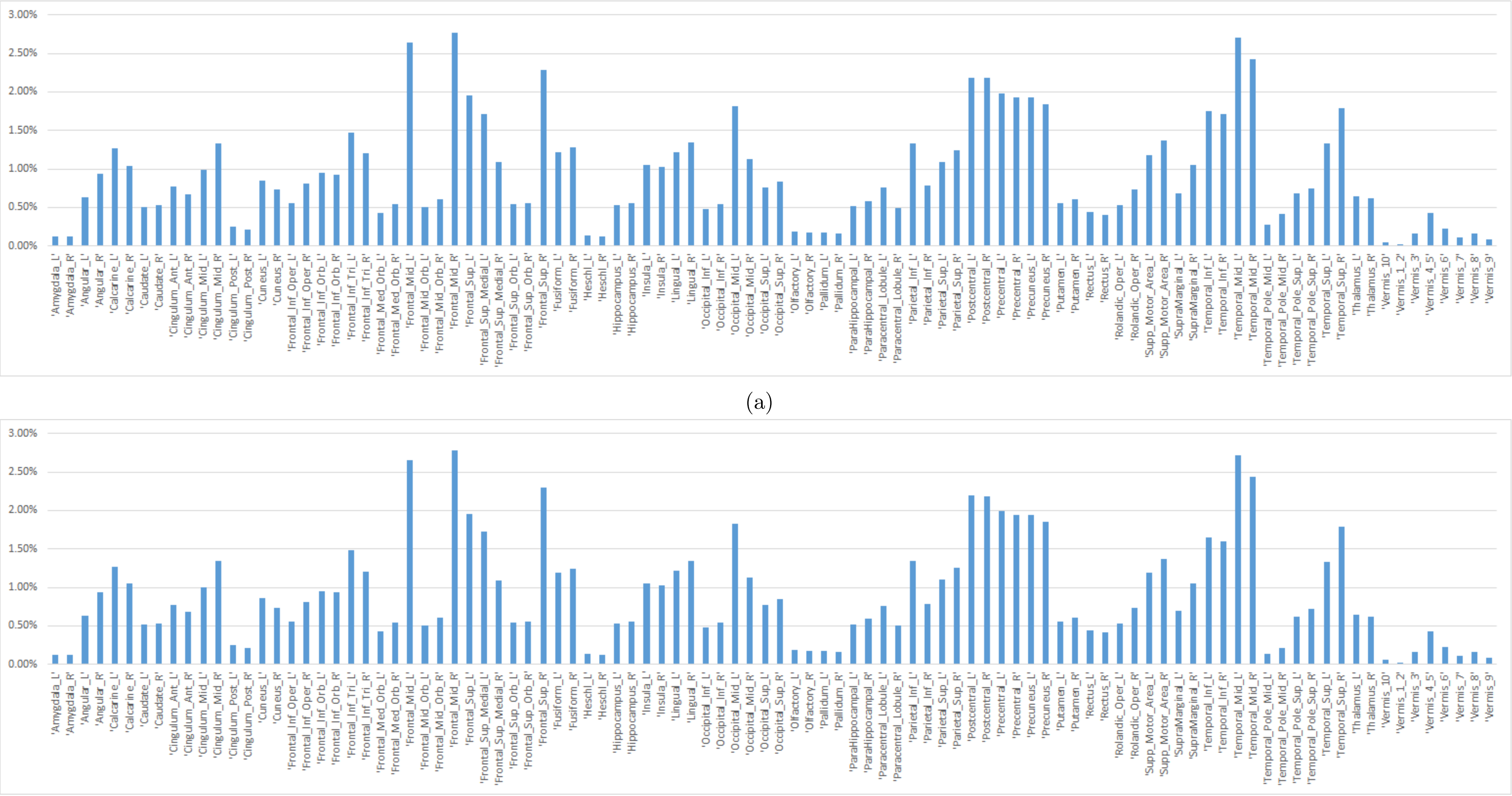


(a)



(b)

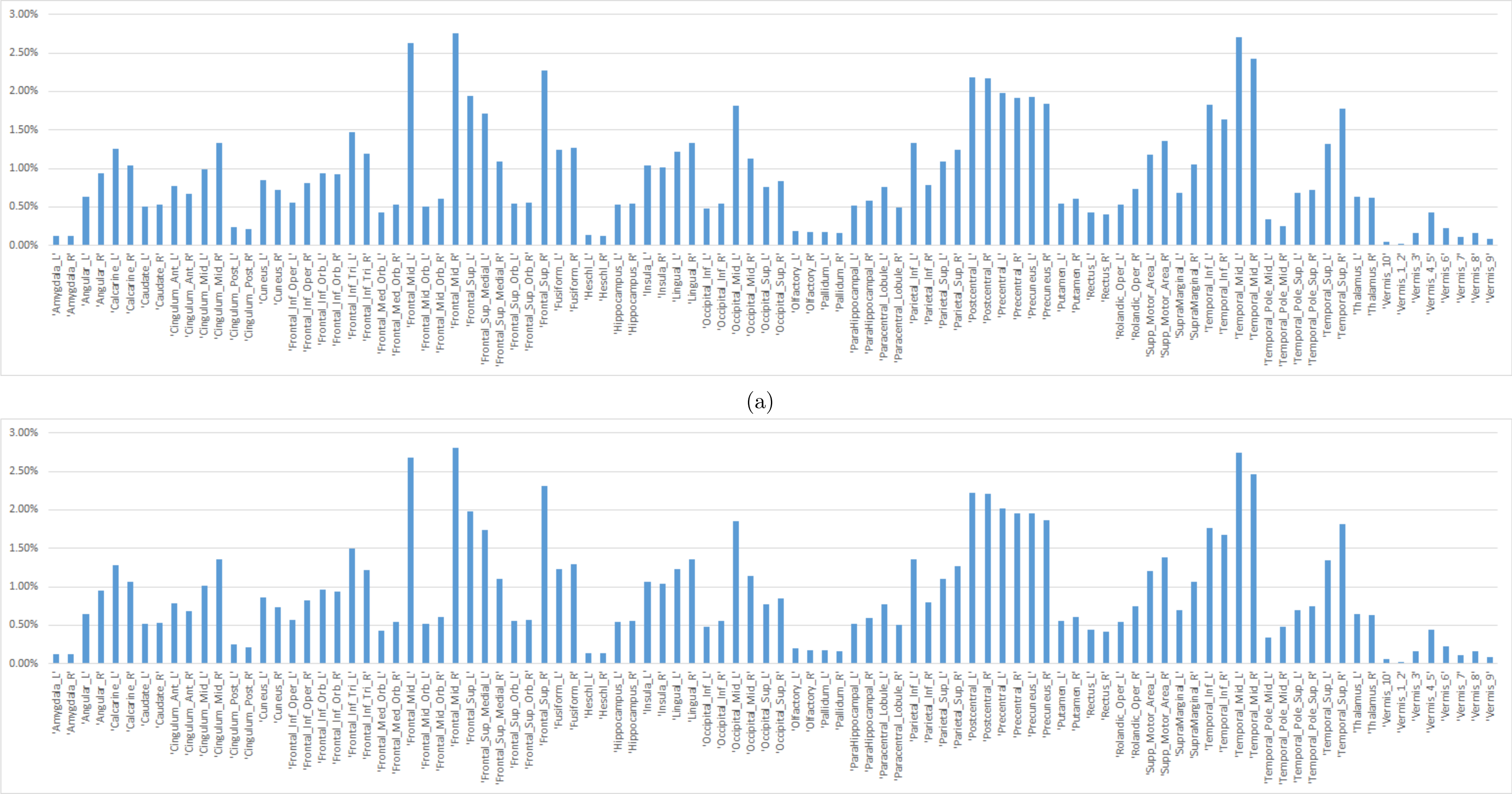
Figure 4.21: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 3 (a) and 4 (b) of the Objects dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$.



(a)

(b)

Figure 4.22: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 1 (a) and 2 (b) of the Emotion 2 Class dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$.



(a)

(b)

Figure 4.23: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for Subjects 4 (a) and 5 (b) of the Emotion 2 Class dataset are shown. The effective voxels are selected using ANOVA with $p \leq 0.01$.

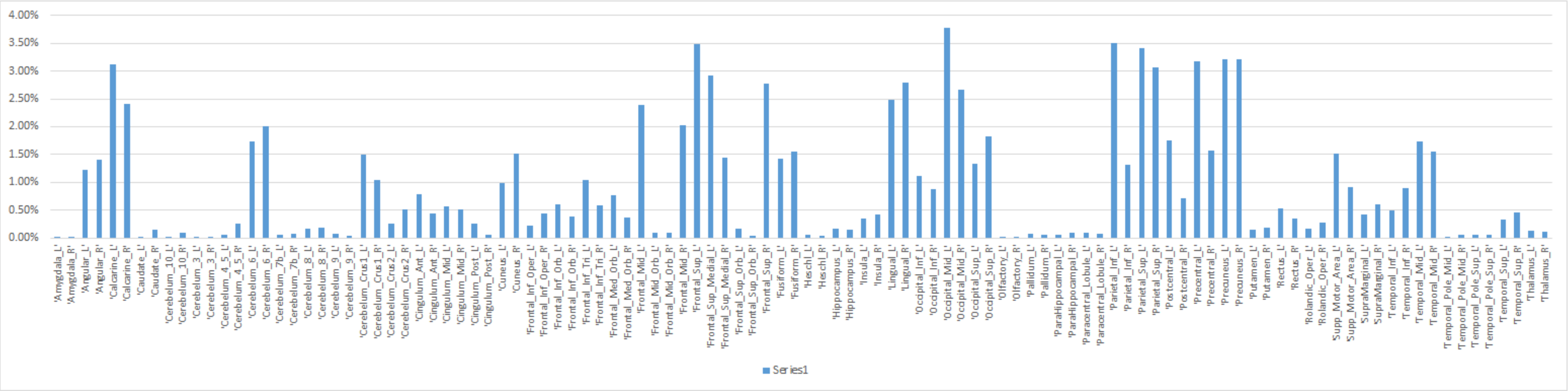


Figure 4.24: Relative voxel frequencies of effective voxels within each AAL region with respect to the number of all effective voxels from the whole brain for all Subjects of TOL dataset, averaged over the Subjects. The effective voxels are selected using ANOVA with $p \leq 0.01$.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

In this study, we have presented a computational model (BRE) of human brain to decode mental states from fMRI images. With BRE, it is our aim to capture distributed representations of mental states such as their visual (shapes, textures, colors) and emotional components within voxel groups using supervoxels and combine them using classifier ensembles for brain decoding.

Motivated by the distributed representations of mental states in the human brain [66, 81, 43], we suggested clustering methods to isolate the distinct aspects of mental representations, which are obtained using fMRI, within homogenous voxel groups that we call supervoxels. We define supervoxels to be groups of voxels that are specified using either a clustering algorithm that uses functional correlation between the voxels, or specified using a functional brain atlas (AAL). We use clustering to specify supervoxels which include groups of voxels which respond similarly to the stimuli or the mental tasks that are used during an fMRI experiment. With clustering, our aim is to specify supervoxels that are relevant to the tasks or stimuli used in a particular fMRI experiment. AAL regions on the other hand, are not specific to the tasks or stimuli of a particular fMRI experiment, but using AAL regions as supervoxels offer a baseline to compare them with the supervoxels obtained with clustering algorithms.

We suggested to combine the mental representations captured within the supervoxels using a classifier ensemble that we call a Brain Region Ensemble (BRE). We consider the activity of voxels within each supervoxel as a distinct feature that corresponds to a specific aspect of a mental state recorded during an fMRI experiment. For BRE, a base layer classifier that outputs class posteriori probabilities is trained, usin the inputs from a specific supervoxel in terms of voxel activity values correspond to that supervoxel. Then, for a set of such supervoxels, the outputs of base layer classifiers in the form of class posteriori probabilities are concatenated and fed to a meta classifier that correspond to the BRE. We proposed a classifier ensemble (RSS-BRE) that is composed of a set of BREs, each of which uses a set of randomly sampled supervoxels, or random subsets of supervoxels (RSS), which uses a voting strategy for final classification. We also explored possible methods for forming random subsets of supervoxels for RSS-BRE, where we formed RSS-BRE for all supervoxels, and we formed a RSS-BRE for each set of supervoxels generated using a specific clustering parameter.

When the classification accuracy results are compared, we show that RSS-BRE performs significantly better than the widely used brain decoding algorithms that use voxel selection and random subspace ensembles. In all datasets that we used, we could directly observe this result.

We introduced diversity as the indicator for the success of an ensemble learning algorithm. We used Q-Statistic and disagreement measure to compute the diversity among the base layer classifiers. Diversity among base layer classifiers is used to compare the diversity of the classifiers that are based on supervoxels generated by clustering algorithms, and supervoxels specified by AAL regions. Also, we compared the diversity of brain region ensembles in a RSS-BRE with random subspace ensembles of support vector machines that use voxels selected by a voxel selection algorithm.

We postulated that partitioning the brain volume into smaller regions that we call supervoxels would allow us to capture distinct aspects of mental states within the supervoxels. We argued that ensembles of classifiers each of which receive input from such supervoxels would provide us with an ensemble with a diverse set of classifiers when compared to the ensembles formed by classifiers that receive inputs from AAL regions. Also, we argued that RSS-BRE formed using smaller supervoxels would have more diverse set of classifiers than random subspace ensembles of selected voxels. In the experiments regarding classifier diversity, we observed supervoxels generated by clustering methods provide sets of supervoxels with corresponding sets of base layer classifiers for which the diversity among the classifiers increases as the number of supervoxels generated during the clustering phase increases. When compared to BREs that are formed using AAL regions as supervoxels, BREs formed by supervoxels that are generated by clustering have a higher diversity among the base layer classifiers when the number of supervoxels are increased. Also, when we used random subsets of supervoxels in order to form a RSS-BRE, we have shown that the diversity of the classifiers within the RSS-BRE becomes higher than the classifier ensembles of support vector machines that use random subspaces of selected voxels, as the number of supervoxels and voxels are both increased, thus, confirming our postulate.

For the specification of the brain regions that are relevant to the classification of the mental states that are under consideration for the fMRI experiment, we proposed to use the classification accuracies of the base layer classifiers that receive input from each supervoxel. We eliminated the supervoxels for which the base layer classifiers that receive input from them do not have a higher accuracy than the chance level. We obtained the relative contribution of each brain region specified by AAL to the classification task that is under consideration by eliminating the irrelevant supervoxels. We also obtained relative contribution of AAL regions to the experimental tasks or stimuli of fMRI using ANOVA. We observed that our method could provide us with the relative contribution of brain regions with respect to the classification task, whereas ANOVA would provide us with the relative contribution of the brain regions when considering all of the stimuli or mental tasks presented during the experiment. In that sense, our method is more specific with respect to the classification task when the relative contribution of the brain regions are considered.

When compared to voxel selection based methods, BRE, especially RSS-BRE provide a better classification accuracy for the classification of mental states under consideration by an fMRI experiment. However, when compared to methods that use voxel selection, BRE is computationally expensive due to the coss-validation phase of the base layer classifiers, where in order to obtain the class posteriori probabilities of the training samples, a classifier is trained for each of the training samples. In order to deal with this problem, a parallel computation scheme can be used, where the classifiers can be trained and the class posteriori probability of each training sample are obtained in parallel.

In contrast to searchlight methods, our method does not assume spatial proximity for voxel groups that are used in the classification tasks. Also, our method is less susceptible for inclusion of irrelevant voxels than searchlight methods since the voxel groups are determined by their activity patterns using clustering. Moreover, our methodology uses multi-voxel analysis for the region specification which makes our method more sensitive for detecting relevant voxel groups when compared to univariate analysis such as ANOVA. These effects can be observed in the region specification results of TOL dataset when compared to the original article.

5.1 Future Directions

In a future study, a dedicated method for selecting the best subset of supervoxels, among all supervoxels that are generated using every clustering parameter, can be developed. Such a set of supervoxels could provide not only a better classification accuracy for decoding mental states, but also it could also be used to provide a complete map of brain regions that is concerned with the representation of the mental states that are discrimative across the experimental tasks.

The performance of BRE for the classification of mental states is higher than the widely used methods of voxel selection. However, BRE is much slower than these methods due to the high number of cross-validations required when forming the base layer fuzzy stacked generalization classifiers. In order for BRE to work in large datasets, a parallel implementation is necessary.

A brain decoding strategy that is similar to BRE can be developed, which uses features generated by an autoencoder [96] that is developed for processing brain images, instead of supervoxels formed by clustering. Features generated by autoencoders could capture the distinct aspects of mental states better than a clustering algorithm that relies only on the similarities between the voxels in terms the correlation of their activities.

REFERENCES

[1] R. Adolphs, D. Tranel, H. Damasio, and A. R. Damasio. Fear and the human amygdala. *The Journal of neuroscience*, 15(9):5879–5891, 1995.

[2] A. Afrasiyabi, I. Onal, and F. T. Y. Vural. Effect of voxel selection on temporal mesh model for brain decoding. In *Signal Processing and Communication Application Conference (SIU), 2016 24th*, pages 2249–2252. IEEE, 2016.

[3] E. Aksan and F. T. Yarman Vural. *An fMRI segmentation method under markov random fields for brain decoding. [Electronic resource].* Ankara : METU, 2015., n.d.

[4] S. Alkan and F. T. Yarman-Vural. Localization of semantic category classification in fmri images. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pages 2178–2181. IEEE, 2014.

[5] S. Alkan and F. T. Yarman-Vural. Ensembling brain regions for brain decoding. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 2948–2951. IEEE, 2015.

[6] E. Amaro and G. J. Barker. Study design in fmri: basic principles. *Brain and cognition*, 60(3):220–232, 2006.

[7] I. Barandiaran. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell*, 20(8):1–22, 1998.

[8] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

[9] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. Behrens, and S. M. Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fmri. *NeuroImage*, 2013.

[10] M. Brett, J.-L. Anton, R. Valabregue, and J.-B. Poline. Region of interest analysis using the marsbar toolbox for spm 99. *Neuroimage*, 16(2):S497, 2002.

[11] V. D. Calhoun, R. F. Silva, T. Adalı, and S. Rachakonda. Comparison of pca approaches for very large group ica. *NeuroImage*, 118:662–666, 2015.

[12] C. Cavina-Pratesi, R. Kentridge, C. Heywood, and A. Milner. Separate processing of texture and form in the ventral stream: evidence from fmri and visual agnosia. *Cerebral Cortex*, 20(2):433–446, 2009.

[13] C. Cavina-Pratesi, R. Kentridge, C. Heywood, and A. Milner. Separate channels for processing form, texture, and color: evidence from fmri adaptation and visual object agnosia. *Cerebral cortex*, 20(10):2319–2332, 2010.

[14] E. Challis, P. Hurley, L. Serra, M. Bozzali, S. Oliver, and M. Cercignani. Gaussian process classification of alzheimer's disease and mild cognitive impairment from resting-state fmri. *Neuroimage*, 112:232–243, 2015.

[15] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[16] P.-H. Chen, X. Zhu, H. Zhang, J. S. Turek, J. Chen, T. L. Willke, U. Hasson, and P. J. Ramadge. A convolutional autoencoder for multi-subject fmri data aggregation. *arXiv preprint arXiv:1608.04846*, 2016.

[17] P.-H. C. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. Haxby, and P. J. Ramadge. A reduced-dimension fmri shared response model. In *Advances in Neural Information Processing Systems*, pages 460–468, 2015.

[18] C.-A. Chou, K. Kampa, S. H. Mehta, R. F. Tungaraza, W. A. Chaovalitwongse, and T. J. Grabowski. Voxel selection framework in multi-voxel pattern analysis of fmri data for prediction of neural response to visual stimuli. *Medical Imaging, IEEE Transactions on*, 33(4):925–934, 2014.

[19] J. D. Cohen, N. Daw, B. Engelhardt, U. Hasson, K. Li, Y. Niv, K. A. Norman, J. Pillow, P. J. Ramadge, N. B. Turk-Browne, et al. Computational approaches to fmri analysis. *Nature neuroscience*, 20(3):304, 2017.

[20] I. Costantini, P. Filipiak, K. Maksymenko, S. Deslauriers-Gauthier, and R. Deriche. fmri deconvolution via temporal regularization using a lasso model and the lars algorithm. In *40th International Engineering in Medicine and Biology Conference*, 2018.

[21] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012.

[22] T. Cukur, S. Nishimoto, A. G. Huth, and J. L. Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(6):763, 2013.

[23] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[24] M. R. Delgado, L. E. Nystrom, C. Fissell, D. Noll, and J. A. Fiez. Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of neurophysiology*, 84(6):3072–3077, 2000.

[25] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[26] Y. Du, G. D. Pearlson, J. Liu, J. Sui, Q. Yu, H. He, E. Castro, and V. D. Calhoun. A group ica based framework for evaluating resting fmri markers when disease categories are unclear: application to schizophrenia, bipolar, and schizoaffective disorders. *Neuroimage*, 122:272–280, 2015.

[27] H. Eidenberger. Statistical analysis of content-based mpeg-7 descriptors for image retrieval. *Multimedia Systems*, 10(2):84–97, 2004.

[28] O. Ekmekci, O. Firat, M. Ozay, I. Oztekin, F. T. Y. Vural, and U. Oztekin. Mesh learning for object classification using fmri measurements. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 2631–2634. IEEE, 2013.

[29] I. O. Ertugrul, M. Ozay, and F. T. Y. Vural. Hierarchical multi-resolution mesh networks for brain decoding. *Brain imaging and behavior*, pages 1–17, 2017.

[30] I. O. Ertugrul, M. Ozay, and F. T. Y. Vural. Encoding the local connectivity patterns of fmri for cognitive task and state classification. *Brain Imaging and Behavior*, pages 1–12, 2018.

[31] A. Etkin, T. Egner, and R. Kalisch. Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in cognitive sciences*, 15(2):85–93, 2011.

[32] J. A. Etzel, J. M. Zacks, and T. S. Braver. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage*, 78:261–269, 2013.

[33] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[34] Y. Fan, S. M. Resnick, and C. Davatzikos. Feature selection and classification of multiparametric medical images using bagging and svm. In *Medical Imaging 2008: Image Processing*, volume 6914, page 69140Q. International Society for Optics and Photonics, 2008.

[35] E. Feczko, N. Balba, O. Miranda-Dominguez, M. Cordova, S. Karalunas, L. Irwin, D. Demeter, A. Hill, B. Langhorst, J. G. Painter, et al. subtyping cognitive profiles in autism spectrum disorder using a functional random forest algorithm. *Neuroimage*, 172:674–688, 2018.

[36] O. Firat, M. Ozay, I. Onal, I. Oztekiny, and F. T. Y. Vural. Functional mesh learning for pattern analysis of cognitive processes. In *12th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 161–167. IEEE, 2013.

[37] O. Firat, L. Oztekin, and F. T. Y. Vural. Deep learning for brain decoding. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 2784–2788. IEEE, 2014.

[38] J. Gu, L. Jiao, F. Liu, S. Yang, R. Wang, P. Chen, Y. Cui, J. Xie, and Y. Zhang. Random subspace based ensemble sparse representation. *Pattern Recognition*, 74:544–555, 2018.

[39] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[40] A. R. Hariri, A. Tessitore, V. S. Mattay, F. Fera, and D. R. Weinberger. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323, 2002.

[41] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[42] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37:435–456, 2014.

[43] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[44] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[45] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.

[46] C. Herff and T. Schultz. Automatic speech recognition from neural signals: a focused review. *Frontiers in neuroscience*, 10:429, 2016.

[47] T. K. Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[48] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[49] Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679, 2005.

[50] Y. Kamitani and F. Tong. Decoding seen and attended motion directions from activity in the human visual cortex. *Current biology*, 16(11):1096–1102, 2006.

[51] M. L. Keightley, G. Winocur, S. J. Graham, H. S. Mayberg, S. J. Hevenor, and C. L. Grady. An fmri study investigating cognitive modulation of brain regions associated with emotional processing of visual stimuli. *Neuropsychologia*, 41(5):585–596, 2003.

[52] D. Kemmerer, J. G. Castillo, T. Talavage, S. Patterson, and C. Wiley. Neuroanatomical distribution of five semantic components of verbs: evidence from fmri. *Brain and Language*, 107(1):16–43, 2008.

[53] E. Kim and H. Park. Pairwise classifier ensemble with adaptive subclassifiers for fmri pattern analysis. *Neuroscience bulletin*, 33(1):41–52, 2017.

[54] T. Kim, W. Bair, and A. Pasupathy. Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, pages 3073–18, 2019.

[55] L. I. Kuncheva and J. J. Rodríguez. Classifier ensembles for fmri data analysis: an experiment. *Magnetic Resonance Imaging*, 28(4):583–593, 2010.

[56] L. I. Kuncheva, J. J. Rodríguez, C. O. Plumpton, D. E. Linden, and S. J. Johnston. Random subspace ensembles for fmri classification. *Medical Imaging, IEEE Transactions on*, 29(2):531–542, 2010.

[57] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[58] R. D. Lane, E. M. Reiman, M. M. Bradley, P. J. Lang, G. L. Ahern, R. J. Davidson, and G. E. Schwartz. Neuroanatomical correlates of pleasant and unpleasant emotion. *Neuropsychologia*, 35(11):1437–1444, 1997.

[59] Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı. Quantifying the interaction and contribution of multiple datasets in fusion: application to the detection of schizophrenia. *IEEE transactions on medical imaging*, 36(7):1385–1395, 2017.

[60] J. A. Lewis-Peacock and K. A. Norman. Multi-voxel pattern analysis of fmri data. *The cognitive neurosciences*, pages 911–920, 2013.

[61] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[62] A. Mahmoudi, S. Takerkart, F. Regragui, D. Boussaoud, and A. Brovelli. Multivoxel pattern analysis for fmri data: a review. *Computational and mathematical methods in medicine*, 2012, 2012.

[63] A. Martin, C. L. Wiggs, L. G. Ungerleider, and J. V. Haxby. Neural correlates of category-specific knowledge. *Nature*, 379(6566):649–652, 1996.

[64] P. M. Matthews and P. Jezzard. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12, 2004.

[65] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fmri-based prediction of behavior. *IEEE transactions on medical imaging*, 30(7):1328–1340, 2011.

[66] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

[67] E. Mızrak and I. Öztekin. Relationship between emotion and forgetting. *Emotion*, 16(1):33, 2016.

[68] H. Moğultay, S. Alkan, and F. T. Yarman-Vural. Classification of fmri data by using clustering. In *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, pages 2381–2383. IEEE, 2015.

[69] H. Mogultay and F. T. Y. Vural. Cognitive learner: An ensemble learning architecture for cognitive state classification. In *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, pages 1–4. IEEE, 2017.

[70] H. Moğultay and F. T. Yarman Vural. *A Hierarchical representation and decoding of fMRI data by partitioning a brain network. [Electronic resource].* Ankara : METU, 2017., n.d.

[71] S. D. Newman, J. A. Greco, and D. Lee. An fmri study of the tower of london: a look at problem structure differences. *Brain research*, 1286:123–132, 2009.

[72] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006.

[73] I. Onal, M. Ozay, and F. T. Y. Vural. Functional mesh model with temporal measurements for brain decoding. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2624–2628. IEEE, 2015.

[74] M. Ozay, I. Öztekin, U. Öztekin, and F. T. Yarman Vural. Mesh learning for classifying cognitive processes. *arXiv preprint arXiv:1205.2382*, 2012.

[75] M. Ozay and F. T. Yarman-Vural. Hierarchical distance learning by stacking nearest neighbor classifiers. *Information Fusion*, 29:14–31, 2016.

[76] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[77] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[78] C. O. Plumpton, L. I. Kuncheva, N. N. Oosterhof, and S. J. Johnston. Naive random subspace ensemble with linear classifiers for real-time classification of fmri data. *Pattern Recognition*, 45(6):2101–2108, 2012.

[79] J. Prado and I. A. Noveck. Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, 19(4):642–657, 2007.

[80] J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, and D. Van De Ville. Decoding brain states from fmri connectivity graphs. *Neuroimage*, 56(2):616–626, 2011.

[81] J. Rissman and A. D. Wagner. Distributed representations in memory: insights from functional brain imaging. *Annual review of psychology*, 63:101–128, 2012.

[82] L. Rokach. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 2016.

[83] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[84] A. Sarica, A. Cerasa, and A. Quattrone. Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, 9:329, 2017.

[85] J. J. Shih, D. J. Krusienski, and J. R. Wolpaw. Brain-computer interfaces in medicine. In *Mayo Clinic Proceedings*, volume 87, pages 268–279. Elsevier, 2012.

[86] M. Skurichina and R. P. Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.

[87] K. Smith. Brain decoding: reading minds. *Nature News*, 502(7472):428, 2013.

[88] R. Stark, M. Zimmermann, S. Kagerer, A. Schienle, B. Walter, M. Weygandt, and D. Vaitl. Hemodynamic brain correlates of disgust and fear ratings. *Neuroimage*, 37(2):663–673, 2007.

[89] J. Talairach and P. Tournoux. Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging. 1988.

[90] F. Tong and M. S. Pratte. Decoding patterns of human brain activity. *Annual review of psychology*, 63:483–509, 2012.

[91] L. Tyler, E. Stamatakis, E. Dick, P. Bright, P. Fletcher, and H. Moss. Objects and their actions: evidence for a neurally distributed semantic system. *Neuroimage*, 18(2):542–557, 2003.

[92] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[93] P. Vaidyanathan. The theory of linear prediction. *Synthesis lectures on signal processing*, 2(1):1–184, 2007.

[94] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[95] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

[96] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[97] T. D. Wager and M. A. Lindquist. Principles of fmri. *New York: Leanpub*, 2015.

[98] Z. Yu, D. Wang, J. You, H.-S. Wong, S. Wu, J. Zhang, and G. Han. Progressive subspace ensemble learning. *Pattern Recognition*, 60:692–705, 2016.

[99] G. U. Yule. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194:257–319, 1900.

[100] X. Zhu, X. Du, M. Kerich, F. W. Lohoff, and R. Momenan. Random forest based classification of alcohol dependence patients and healthy controls using resting state mri. *Neuroscience letters*, 676:27–33, 2018.

[101] I. Önal Ertuğrul. *Representation of human brain by mesh networks.* PhD thesis, Middle East Technical University, 2017.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Alkan, Sarper
Nationality: Turkish (TC)
Date and Place of Birth: 03.07.1980, Gemerek-SİVAS
Marital Status: Married
e-mail: sarperalkan@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
M.S.	Department of Cognitive Sciences	2005
B.S.	Department of Mechanical Engineering	2002
High School	Ankara Cumhuriyet High School	1996

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2016 -	Biokido Medical Engineering	Partner
2013 - 2016	Cankaya University	Specialist
2006 - 2010	ODTÜ	Research Assistant

PUBLICATIONS

S. Alkan, F. T. Yarman Vural. "Ensembling brain regions for brain decoding",37th annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015.

H. Mogultay, S. Alkan, F. T. Yarman Vural. "Classification of fMRI data by using clustering", 23rd IEEE Conference on Signal Processing and Communications Applications (SIU), 2015.

S. Alkan, F. T. Yarman Vural, "Localization of semantic category classification in fMRI images", 22nd IEEE Conference on Signal Processing and Communications Applications (SIU), 2014.

S. Alkan, M. Gülgeç, K. W. Schmidt, 7. Engineering and Technology Symposium, Book of Poceedings (Editor), Ankara, 2014