AUTOMATIC DETECTION OF BRAIN FUNCTIONAL DISORDER USING IMAGING DATA

by

SOUMYABRATA DEY
B.Tech., West Bengal University of Technology, India 2005
M.S., University of Central Florida, US 2011

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2014

Major Professor: Mubarak Shah

ABSTRACT

Recently, Attention Deficit Hyperactive Disorder (ADHD) is getting a lot of attention mainly for two reasons. First, it is one of the most commonly found childhood behavioral disorders. Around 5-10% of the children all over the world are diagnosed with ADHD. Second, the root cause of the problem is still unknown and therefore no biological measure exists to diagnose ADHD. Instead, doctors need to diagnose it based on the clinical symptoms, such as inattention, impulsivity and hyperactivity, which are all subjective.

Functional Magnetic Resonance Imaging (fMRI) data has become a popular tool to understand the functioning of the brain such as identifying the brain regions responsible for different cognitive tasks or analyzing the statistical differences of the brain functioning between the diseased and control subjects. ADHD is also being studied using the fMRI data. In this dissertation we aim to solve the problem of automatic diagnosis of the ADHD subjects using their resting state fMRI (rs-fMRI) data.

As a core step of our approach, we model the functions of a brain as a connectivity network, which is expected to capture the information about how synchronous different brain regions are in terms of their functional activities. The network is constructed by representing different brain regions as the nodes where any two nodes of the network are connected by an edge if the correlation of the activity patterns of the two nodes is higher than some threshold. The brain regions, represented as the nodes of the network, can be selected at different granularities e.g. single voxels or cluster of functionally homogeneous voxels. The topological differences of the constructed networks of the ADHD and control group of subjects are then exploited in the classification approach.

We have developed a simple method employing the Bag-of-Words (BoW) framework for the classification of the ADHD subjects. We represent each node in the network by a 4-D feature vector: node degree and 3-D location. The 4-D vectors of all the network nodes of the training data are then grouped in a number of clusters using K-means; where each such cluster is termed as a

word. Finally, each subject is represented by a histogram (bag) of such words. The Support Vector Machine (SVM) classifier is used for the detection of the ADHD subjects using their histogram representation. The method is able to achieve 64% classification accuracy.

The above simple approach has several shortcomings. First, there is a loss of spatial information while constructing the histogram because it only counts the occurrences of words ignoring the spatial positions. Second, features from the whole brain are used for classification, but some of the brain regions may not contain any useful information and may only increase the feature dimensions and noise of the system. Third, in our study we used only one network feature, the degree of a node which measures the connectivity of the node, while other complex network features may be useful for solving the proposed problem.

In order to address the above shortcomings, we hypothesize that only a subset of the nodes of the network possesses important information for the classification of the ADHD subjects. To identify the important nodes of the network we have developed a novel algorithm. The algorithm generates different random subset of nodes each time extracting the features from a subset to compute the feature vector and perform classification. The subsets are then ranked based on the classification accuracy and the occurrences of each node in the top ranked subsets are measured. Our algorithm selects the highly occurring nodes for the final classification. Furthermore, along with the node degree, we employ three more node features: network cycles, the varying distance degree and the edge weight sum. We concatenate the features of the selected nodes in a fixed order to preserve the relative spatial information. Experimental validation suggests that the use of the features from the nodes selected using our algorithm indeed help to improve the classification accuracy. Also, our finding is in concordance with the existing literature as the brain regions identified by our algorithms are independently found by many other studies on the ADHD. We achieved a classification accuracy of 69.59% using this approach. However, since this method represents each voxel as a node of the network which makes the number of nodes of the network several thousands. As a result, the network construction step becomes computationally very expensive.

Another limitation of the approach is that the network features, which are computed for each node of the network, captures only the local structures while ignore the global structure of the network.

Next, in order to capture the global structure of the networks, we use the Multi-Dimensional Scaling (MDS) technique to project all the subjects from an unknown network-space to a low dimensional space based on their inter-network distance measures. For the purpose of computing distance between two networks, we represent each node by a set of attributes such as the node degree, the average power, the physical location, the neighbor node degrees, and the average powers of the neighbor nodes. The nodes of the two networks are then mapped in such a way that for all pair of nodes, the sum of the attribute distances, which is the inter-network distance, is minimized. To reduce the network computation cost, we enforce that the maximum relevant information is preserved with minimum redundancy. To achieve this, the nodes of the network are constructed with clusters of highly active voxels while the activity levels of the voxels are measured based on the average power of their corresponding fMRI time-series. Our method shows promise as we achieve impressive classification accuracies (73.55%) on the ADHD-200 data set. Our results also reveal that the detection rates are higher when classification is performed separately on the male and female groups of subjects.

So far, we have only used the fMRI data for solving the ADHD diagnosis problem. Finally, we investigated the answers of the following questions. Do the structural brain images contain useful information related to the ADHD diagnosis problem? Can the classification accuracy of the automatic diagnosis system be improved combining the information of the structural and functional brain data? Towards that end, we developed a new method to combine the information of structural and functional brain images in a late fusion framework. For structural data we input the gray matter (GM) brain images to a Convolutional Neural Network (CNN). The output of the CNN is a feature vector per subject which is used to train the SVM classifier. For the functional data we compute the average power of each voxel based on its fMRI time series. The average power of the fMRI time series of a voxel measures the activity level of the voxel. We found significant differences

in the voxel power distribution patterns of the ADHD and control groups of subjects. The Local binary pattern (LBP) texture feature is used on the voxel power map to capture these differences. We achieved 74.23% accuracy using GM features, 77.30% using LBP features and 79.14% using combined information.

In summary this dissertation demonstrated that the structural and functional brain imaging data are useful for the automatic detection of the ADHD subjects as we achieve impressive classification accuracies on the ADHD-200 data set. Our study also helps to identify the brain regions which are useful for ADHD subject classification. These findings can help in understanding the pathophysiology of the problem. Finally, we expect that our approaches will contribute towards the development of a biological measure for the diagnosis of the ADHD subjects.

To my wife Sampita, my parents Tapan Kanti Dey and Shrabani Dey, sister Sayantani, brother-in-law Anupam and niece Megh.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Mubarak Shah, Dr. Ravi Rao and the committee for their guidance throughout the research process. I would like to also acknowledge my family, friends, and lab mates throughout the years that have accompanied me on this journey of self-discovery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Motivation

ADHD is rated as one of the most commonly found childhood behavioral brain disorders. Around 5-10% of the children all over the world are diagnosed with ADHD [4]. Children diagnosed with ADHD may suffer from learning difficulties, developing behavioral abnormalities or fidgety, disobedience or aggression towards authorities. They often face difficulties in understanding instructions, concentrating on a task and remembering important things. The children also suffer from anxiety and depression and cannot control their emotions.

Recently, researchers are putting a lot of effort to discover the root cause of this problem which is still unknown. No well known biological measure exists to date to detect ADHD. Instead, clinical symptoms, such as inattention, impulsivity and hyperactivity are used to characterize the subjects affected with this problem. The ADHD diagnosis process is often questioned for various reasons. Many times the diagnosis is performed by general pediatricians and family doctors who do not have extensive training required for the task. Scarcity of psychiatrists and neurologists, lack of knowledge of the problem and instinctive judgment make the situation even worse. As a result, according to the Centers for Disease Control and Prevention, one in seven children in the United States and almost 20 percent of all boys receive a diagnosis of ADHD by the time they turn 18. Many experts believe that this one in five ratio is a clear sign of over-diagnosis of the problem. All these facts motivate us to develop an automatic diagnosis process using brain functional activity data which can standardize the detection process and reduce the dependency on the human expertise. Dr. Thomas Insel, Director of the National Institute of Mental Health (NIMH) also shares the same view as he mentioned - ”We need to begin collecting the genetic,

imaging, physiologic, and cognitive data to see how all the data - not just the symptoms - cluster and how these clusters relate to treatment response" [39], while talking about the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). These issues have motivated us to ask two major questions. Can we create a framework for automatic classification of ADHD subjects that performs better than the current best algorithms? How can we identify brain regions that contain significant differences between the ADHD and control groups? For our studies we used rs-fMRI and sMRI data of the brain.

1.2 fMRI Overview

The main part of our brain activity is performed in terms of communication among the neurons. Neurons communicate among each other by transporting charged particles or ions through their synapsis. This activity results in an increase of energy requirements for the brain regions. The brain produces this energy by consuming glucose and oxygen transported through blood vessels. Hence, the measurement of the blood oxygen level in a brain region can be considered as an indirect measure of the activity level of the region. Blood Oxygen Level Dependent(BOLD) fMRI is a technique to measure the brain activity by measuring the blood oxygen concentration [35]. The fMRI data can be considered as a video where each frame of the video is a 3D image of the brain activity. The regions with higher activity levels are captured with brighter intensity. The brain volume is divided into small cubicle regions called voxels. Hence, the fMRI data can also be viewed as an intensity time series observed for each voxel of the brain volume.

1.3 Previous Works

Recently, fMRI has become a very popular tool for the analysis of brain functional activities. It has extensive use in identifying the brain regions responsible for particular cognitive activities (task-related fMRI). Researchers also used it to better understand different brain func-

tional diseases like Dementia [61] based on the functional activity pattern differences from the control group. Likewise, structural and functional brain imaging techniques are also being used to analyze the group level statistics of the ADHD and control subjects. Studies using structural MRI (sMRI) data on ADHD subjects found abnormalities in different brain regions, specifically in the frontal lobes, basal ganglia, parietal lobe, occipital lobe, and cerebellum (Castellanos et al., 1996, Overmeyer et al., 2001, Seidman et al. 2006, Sowell et al. 2003 [12, 57, 65, 71]). In a different set of studies, task-related fMRI analysis is used on ADHD subjects. Bush et al., 1999 [8] found significant low activity in the anterior cingulate cortex when ADHD subjects are asked to perform the Counting Stroop during fMRI. Durston, 2003 [29] showed that ADHD conditioned children have difficulties performing the go/no-go task and display decreased activity in the frontostriatal regions. Teicher et al., 2000 [75] demonstrated that boys with ADHD have higher T2 relaxation time in the putamen region of brain which is directly connected to a child's capacity to sit still.

A third set of works is done using the resting state brain fMRI to locate any abnormalities in the Default Mode Network (DMN) [59]. Castellanos et al., 2008 [13] performed the Generalized Linear Model based regression analysis on the whole brain with respect to three frontal foci of DMN, and found low negative correlated activity in precuneus/anterior cingulate cortex in ADHD subjects. Tian et al., 2006 [76] found functional abnormalities in the dorsal anterior cingulate cortex; Cao et al., 2006 [10] showed decreased regional homogeneity in the frontal-striatal-cerebellar circuits, but increased regional homogeneity in the occipital cortex among boys with ADHD. Zang et al., 2007 [81] verified decreased Amplitude of Low-Frequency Fluctuation (ALFF) in the right inferior frontal cortex, left sensorimotor cortex, bilateral cerebellum, and the vermis, as well as increased ALFF in the right anterior cingulate cortex, left sensorimotor cortex, and bilateral brainstem.

Studies of group level statistics are successfully able to indicate the abnormal regions of the ADHD subjects but still these techniques lack the ability of automatic diagnosis of the disordered subjects. There have been relatively few investigations at the individual level of classification of

the ADHD subjects. One of the first attempts is made by Zhu et al., 2010 [82] where regional homogeneity of the fMRI data is used as the feature to classify the ADHD subjects.

Recently, there is a global competition (ADHD-200) organized for automatic diagnosis of ADHD subjects as well as understanding the pathophysiology of the problem. Researchers from different disciplines of science are involved in this work. The organizers released a data-set [53] containing rs-fMRI data, sMRI data and phenotypic information of a large number of ADHD and control subjects. In total, eight different data collection centers contributed for the data set. Since subjects from different demographic and different experimental protocols are used by different data centers for collection of the data etc make the data set complex and challenging.

A set of interesting works on automatic classification is published using the ADHD-200 data set ([6], [7], [16], [17], [19], [24], [26], [30], [56], [63], [68]). Many of these works used some combination of rs-fMRI, sMRI and phenotypic data. Some of the common sMRI features used for the classification are cortical thickness, gray matter probability and texture of structural brain images. Regional homogeneity and Fourier transformation are some of the features calculated from fMRI data and used for the classification in the studies. Many of the studies computed functional networks from fMRI data and used different network statistics as the features. Brown et al., 2012 [7] used only phenotypic features for their experiments and still got impressive classification accuracies. All of these works achieved classification accuracy higher than the chance.

1.4 Proposed Approach

The purpose of this dissertation is to analyze the importance of the brain imaging data for developing an efficient method for automatic diagnosis of the ADHD affected subjects. We used rs-fMRI and gray matter (GM) structural MRI data released for the ADHD-200 competition for the experimental validation of our proposed method. We also identified the key brain regions which show significant differences of feature values between the ADHD and the control groups of

subjects. We believe that our study will not only help to build an efficient diagnostic system, but also will provide important pathophysiological findings which will help to better understand the root cause of the disorder.

As already indicated, though the root cause of the ADHD is still unknown, there are some hypothesis regarding the problem. One of the strong notions is the lack of neurotransmitters in the ADHD affected subjects that prevent the normal communications among the different brain regions. In our dissertation, we attempt to verify this hypothesis. The core step of our approach is the construction of the network which can capture the functional connectivity among different brain regions. To construct the network, the brain volume is divided into small regions where each region is represented by a node of the network. The brain regions can be selected at different resolutions. In our approach we chose to use two different resolutions; in the finer resolution, we represent each voxel as a node while in the coarser resolution, clusters of functionally homogeneous voxels are represented as the nodes of the network. Any two nodes of the network are connected by an edge if the correlation of the average time series of the regions is sufficiently high. Once the networks are constructed for the subjects under study, the networks' topological differences are exploited for the classification of the ADHD subjects. We started with a simple method which uses the Bag of Words (BoW) framework to encode the network topological features. In this method, each node of the functional connectivity network is expressed by a 4 tuples: the degree of connectivity and the physical 3D coordinates. The 4 tuples representations of all the nodes of the training data are then grouped into clusters using the K-mean clustering algorithm. These clusters are referred to as the words. The BoW framework represents each subject as a histogram of such words. The histograms are then fed into the SVM classifier for the automatic classification of the ADHD subjects. We achieved 64% classification accuracy rate using this method on the ADHD-200 hold out set.

While BoW framework provides us an automatic system for the classification of the ADHD subjects, we look forward to address the shortcomings of the method and improve the classification

accuracy. First, the stated framework uses the features from the whole network to construct the histogram, but some of the brain regions may not contain any useful information for the ADHD diagnosis problem. Hence, using features from the whole network may unnecessarily increase the dimensions of the feature vector and add noise to the system. To address this issue, we hypothesize that only some nodes of the network contributed useful information for the classification problem. We developed an algorithm to identify the useful nodes of the network and construct the feature vector using the features from the selected nodes only. In each iteration of the algorithm, it selects a random subset of the network nodes, extracts features from these selected nodes, and performs classification. The subset selection step is performed several times each time recording the classification accuracy. The subsets are then ranked based on the classification accuracy and the occurrences of each node of the network in the top performing subsets are computed. The algorithm selects the highly occurring nodes as useful regions. Another problem of the BoW method is the loss of spatial information while constructing the histogram of the degree features. To address this problem, we compute the feature vector by concatenating the features from the selected nodes in a fixed order. This helps to preserve the relative spatial position of the nodes. Finally, we realize that along with the degree features, other complex network features may also be useful for this problem. Therefore, we compute three more network features such as the network cycles, the varying distance degree and the edge weight sum.Experimental validation shows that the improvements help to increase the classification accuracy. The improved method achieve a classification accuracy of 69.59% on the ADHD-200 hold out set.

The method described so far computes the network features for each node of the network. While these features can capture the local structure of the network they ignore the global topology. In order to address this issue, we propose a classification framework which refrains from using the network features. Instead it maps the networks onto a low dimensional spatial configuration and perform classifications on the projected space. The networks on their own are hard to use as feature points as they are part of a unknown high dimensional space. The projection method

helps us to use the entire network structures of the subjects as the features for the classification. Our method can be subdivided into three main parts. In the first part we construct the resting state functional connectivity networks of the brains of all the subjects under consideration. The nodes of the network are formed by the clusters of highly active and functionally homogeneous voxels which helped to significantly reduce the network dimension as well as network computation cost. The networks are modeled as attributed graphs where each node has a signature [44]. The signature of a node is a set of attributes which characterizes the node. The attribute set includes the degree of the node, the degree of the neighbour nodes, the power of the node, the power of the neighbour nodes and the physical location of the node. The power of a node is calculated by averaging the power of the fMRI time series of all the voxels comprising the node. In the second part we compute distances between all possible pairs of networks. The distance computation for a pair of networks is a two step process. In the first step all node pair distances are computed based on their signature values. In the next step, all nodes of one network are assigned to the nodes of the second network such that the total matching cost is minimized. The Munkres algorithm is used for the node assignment problem [52]. In the last part the networks are projected to a space of specified dimensions based on their distance measures. The Multidimensional Scaling (MDS) [77] method is used for this purpose. Finally, a Support Vector Machine (SVM) is used for the classification of ADHD subjects in the projected space. The main contribution of the work is a novel automatic classification framework of ADHD subjects based on the topological differences of the functional brain connectivity networks of the ADHD and control groups of subjects. We achieved impressive detection accuracies on the holdout sets (73.55%) of the ADHD-200 data set.

Finally, we try to find the answers to the following questions. First, is sMRI data useful for solving our proposed problem? Second, is it possible to improve the automatic classification method by combining structural and functional imaging data? To seek the answers we use two classification frameworks for structural and functional data modalities. Later we combine the two modalities in a late fusion framework. For structural data we use the 3-D Gray Matter (GM) image

of the brain. The GM image is presented as 2-D slices to a Convolutional Neural Network (CNN) to extract features. The features from all the slices are then merged using a novel late fusion framework. For functional data we use a distribution of average power of all of the brain voxels. The average power of all the voxels of a brain constitutes the power map which is a 3-D image. We found considerable differences of power distributions between the ADHD and control groups of subjects. To capture the differences, we compute the Local Binary Pattern (LBP) texture features in three orthogonal directions of the power map image. The average accuracy on ADHD-200 hold out data set using GM and LBP power map features are 74.23% and 77.30% respectively while combination of both modalities further improve the accuracy to 79.14%.

1.5 Contribution

In this dissertation we propose a hypothesis for the automatic detection of the ADHD subjects using their MR brain image data. Different brain regions need to functionally coordinate with each other to perform different cognitive tasks. We propose that the ADHD subjects lack these coordinations due to reduced levels of presence of some neurotransmitters in the brain. To verify this hypothesis, our proposed method tries to find out the topological differences of the brain functional connectivity networks between the ADHD and control groups of subjects and use those for the classification problem. Finally, we showed that the structural brain images also contain useful information related to the ADHD diagnosis problem and using it with functional images can provide additional information which helps to improve the classification accuracy.

The experimental results validate our proposed method as we achieve impressive classification accuracies. Especially, our results using attributed graphs and combination of structural and functional imaging data beat the current state of the art detection rate on the holdout sets of the ADHD-200 data set. Other than the diagnosis of the ADHD subjects, our method helps to identify the brain regions with most useful information for the classification task. We believe that this will

help the community to better understand the pathophysiology of the problem.

1.6 Organization of Dissertation

The rest of the dissertation is organized as follows: Chapter 2 describes the previous approaches on automatic ADHD detection, different brain imaging techniques especially fMRI, the details about the ADHD-200 data set and the data preprocessing steps. Chapter 3 provides our first approach for ADHD detection based on the BoW framework. Next, in chapter 4 we present the network features for the classification of the ADHD subjects. Chapter 5 describes our method for ADHD detection using the whole network structure and projecting them into a lower dimensional space based on inter-network distances. Chapter 6 describes the fusion framework of structural and functional brain imaging data. Finally, Chapter 7 summarizes the contributions and findings of this dissertation followed by a discussion of future directions to explore.

CHAPTER 2: BACKGROUND

In this chapter first we provide a brief description of different brain imaging techniques and introduce the main concepts of fMRI data. Next, we describe the previous works related to the problem of automatic detection of the ADHD subjects. In the final section we provide a detailed description of the ADHD-200 data set and the preprocessing steps performed to make the data useful for any further analysis.

2.1 Brain Functioning and Functional Imaging Techniques

Most of the brain cognitive activities are performed in terms of communications among the neurons through their synapses. The communication, also termed as neural signaling, is performed through transmission and reception of the neurotransmitter molecules which are essentially electrically charged particles or ions. The transmission process of these ions through the electrical potential field between a transmitter and receptor neurons is called conduction. This neural signaling is a high energy consuming process. Whenever a region of a brain is activated by a cognitive task, it increases the neural signaling process in the region which in turn amplifies the energy requirements in the locality.

The energy required for the functioning of the brain is produced through the oxidation of the glucose supplied by the blood vessels of the brain. It is observed that the activity in a region of the brain is highly correlated with the local blood flow, Oxygen and glucose consumption as the increase of brain activity level in a locality leads to the increase of the other events. Thus the brain metabolism process is highly informative about the activity level of the brain. Brain functional imaging techniques take advantage of this relation to map the activity level of the brain regions with measured local blood flow and glucose/oxygen consumptions.

During the last few decades there have been a lot of interest in analyzing brain function-

ing using brain functioning imaging techniques. The plethora of research papers in the area of neuroimaging indicates the same. Here a brief discussion about the common functional imaging techniques is provided.

2.1.1 Positron Emission Tomography (PET)

In PET the subject is first injected with a short lived radioactive tracer isotope. The tracer is introduced into the body through a biologically active molecule. After a waiting period the active molecules are concentrated in the desired tissue and the subject is placed under a scanner to record radioactive emission of the tracer. In the process of decay, the tracer molecule produces a positron which in turn generates two photons moving in opposite directions. The scanner can detect the photons to measure the location of the emission. PET can detect the blood flow or glucose intake rates, which are the indirect measures of the brain activity levels, by measuring the quantity of radiation from a location. PET data has high spatial resolution (approximately 1-10 mm) at the cost of low temporal resolution. For further details please refer to the document [67].

2.1.2 Multichannel Electroencephalography (EEG)

As it is already described, the neurons communicate with each other by exchanging ionized particles through the synapses. The communication process constitutes the main part of the brain activity which causes an electrical current in the brain. EEG is a recording technique of brains electrical current for a short period of time. EEG can record the neuronal activity in a very high temporal frequency (in the range of milliseconds) but the spatial resolution is compromised.

2.1.3 Magnetoencephalography (MEG)

The flow of ionized particles through neurons produces a weak magnetic field in the brain. MEG is a functional neuroimaging technique which can record the magnetic field produced by the electrical current due to neuronal activity. The brain activity level is then mapped with the recorded

magnetic field. As the brain's magnetic field is very weak it is recorded using extremely sensitive magnetometers which use an array of superconducting quantum interference devices (SQUIDs). Similar to EEG it has very high temporal resolution and low spatial resolution.

2.1.4 Near Infrared Spectroscopic Imaging (NIRSI)

NIRSI is a non-invasive optical imaging technique which can be used as a functional brain imaging method. NIRSI uses near infrared (from about 800 nm to 2500 nm) electromagnetic signal to measure blood oxygenation changes in blood vessels of the brain by measuring the absorption of the near infrared signal emitted by the source onto the brain surface. The advantage of NIRSI is it is inexpensive, portable and can be used even when the subject is moving. NIRSI and functional Magnetic Resonance Imaging (fMRI) produce similar data as some previous studies [74] have shown close spatial and temporal correlations when the data is recorded using the two methods.

2.1.5 Functional Magnetic Resonance Imaging (fMRI)

As we used fMRI data for solving the ADHD classification problem, we provide the basic principles behind the data capturing method. The core concept of fMRI is based on the idea of the Nuclear Magnetic Resonance (NMR) technology which has been around for a long time. NMR has a widespread application in the biomedical field for analyzing the characteristics of biomolecules. The basic principles of NMR are explained in the next few sections without going into the mathematical details. The interested readers are referred to the following document [41] for further details.

It is observed that the proton and neutron particles that constitute the nuclei of atoms, possess some angular momentum. A well-known fact of Physics is that a moving electric charge produces a magnetic field. Now, because a proton is a charged particle, the rotational motion produces a magnetic field whose direction is along the direction of the rotational axis.

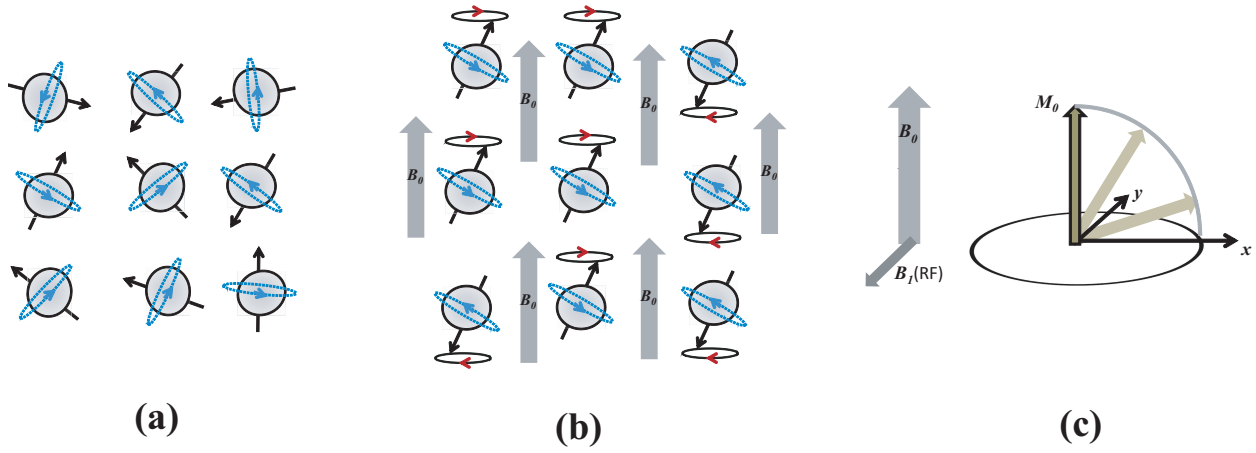


Figure 2.1: The figure demonstrates the main steps of NMR. (a) At the beginning the nuclei rotate around their axes where axes of rotation oriented in random directions. As a result the net magnetic effect is zero. (b) When an external magnetic field B_0 is applied, the axes of rotation are aligned along or against the direction of B_0. (c) When an radio wave in the Larmor frequency (B_1) is applied, nuclei absorb the energy to change their state from along the B_0 to against the B_0. As a result the net magnetization vector drops down to the $x - y$ plane.

On the contrary, a non-charged particle neutron does not show this property. Nuclei being constituted by protons and neutrons sometimes possess an angular momentum as a net effect. All the nuclei which have odd numbers of protons and/or neutrons have an angular momentum. This is also called nuclear spin. Because a nucleus is also a charged particle, it produces a magnetic field due to the rotational motion. Such nuclei with spins can be imagined as small bar magnets with north and south poles causing tiny magnetic fields. The concept is explained in Figure 2.1 (a). According to quantum mechanics, the nuclei with angular momentums are allowed to have only very specific quantized spin values. These quantized values are called spin numbers. In a magnetic field the energy of a nucleus with spin number I splits into $(2I + 1)$ discrete levels. For example the nucleus of a hydrogen atom has only one proton with spin number $I = \frac{1}{2}$ and $(2 \times \frac{1}{2} + 1) = 2$ discrete energy levels. There are other nuclei like ^{13}C, ^{15}N, ^{17}O which have non zero spin numbers but for the sake of the easiness of understanding we will describe the concept of

NMR using the example of ^1H (hydrogen nucleus) only. As it is stated ^1H can have two discrete energy levels under the influence of some external magnetic field. The energy levels correspond to the relative orientations of the nuclear magnetic moments. In the lower energy state the magnetic moment of a ^1H is aligned in the direction of the applied magnetic field where in the higher energy state the direction of magnetic moment is antiparallel to the applied field. Now consider a sample for example water that contains hydrogen atoms. Initially the magnetic moment of ^1H will be in the random direction producing a zero magnetic field in the net effect. Once an external magnetic field B_0 is applied on the sample, the ^1H will try to align themselves along the direction of the B_0. Actually, in the absolute zero temperature, all the ^1H should be in the lower energy state and hence should be aligned along the B_0. While in natural temperatures, due to the thermal agitation some of the ^1H will align along the B_0 and some against the B_0 cancelling each others' magnetic effect. In room temperatures, a slight excess of ^1H will align along the B_0 leading to a net magnetic moment along the B_0. The stronger the B_0 the more ^1H will align along the direction. Also, these alignments (along or against the B_0) of ^1H are not perfect. Instead, they wobble or precess about the axis of the B_0 with a frequency ω_0 (Figure 2.1 (b)). This is called the precessional, Larmor or resonance frequency, and is defined by the famous Larmor equation:

$$\omega_0 = \gamma B_0 \tag{2.1}$$

Where γ is the gyromagnetic ration and is unique for every types of atom. Now according to the quantum mechanics, the ^1H which are in the lower energy state can change their state to the higher energy if an external electromagnetic signal, which oscillates exactly in the Larmor frequency ω_0, is applied. For NMR this electromagnetic frequency lies in the range of the Radio Frequency (RF). Lets assume that the B_0 is in the direction of the z axis of a coordinate frame. Then the effect of applying the RF signal with frequency ω_0 can be viewed in the macro level as the M_0 spiral down towards the xy plane of the coordinate (Figure 2.1 (c)). Once the RF signal is

turned off three things begin to happen.

- In micro level the nuclear spins start returning from the higher energy state to the lower energy state. As a net effect the absorbed RF energy is retransmitted at the Larmor frequency. This retransmission can be detected as a signal whose amplitude decays away exponentially. The decaying of the signal is termed as the 'free induction decay'.

- The M_0 begins to return towards the initial direction along the z axis. The recovery rate of M_0 along the z axis can be mathematically described by an exponential curve. The time t needed to recover 63.2% of M_0 along the z axis is called the $T1$ relaxation time. This $T1$ value is unique for each sample under consideration.

- Initially in phase, the excited ${}^1H$ begin to dephase. This is because each ${}^1H$ experience a slightly different magnetic field due to the interaction of tiny magnetic fields created by neighbor ${}^1H$. As a result the ${}^1H$ start precessing at different frequencies which result in decaying of the amplitude of the released signal. The decay of signal amplitude is exponential and the time taken for the signal strength to reduce to the 36.8% of the original value is called $T2$ time. In real world the decay is faster ($T2*$) than the $T2$ due to the variables outside of controls.

From the above discussion it is easy to understand that the NMR can be used for the analysis of the chemical composition of the underlying sample because $T1$ and $T2$ relaxation times are uniquely dependent on the sample. The concept of the MRI lies in the realization that a spatially varying magnetic field results in a spatially varying Larmor frequency. To elaborate, we know from the Larmor equation that the Larmor frequency ω_0 is proportional to the strength of the applied magnetic field B_0. When a spatially varying external magnetic field is applied on the sample, the nuclei from different spatial locations start precessing in different frequencies. After the sample is excited using a RF signal, the nuclei start releasing signals in different frequencies which is

a function of their spatial locations. These signals are detected and when a Fourier transform is performed they reveal the whole spectrum of frequencies. Each frequency of the spectrum of the signals can then be mapped to the corresponding spatial location based on the function.

As described in Section 2.1 brain activity requires energy in the form of Adenosine Tri-Phosphate (ATP). The formation of the ATP requires glucose and oxygen transported to brain via blood vessels. This oxygen is carried by large iron-containing molecules called hemoglobin (Hb). When oxygen is bound, the molecule is represented as HbO_2. Now, Hb is paramagnetic (having significant magnetic effect on the environment) due to the presence of the iron atoms but HbO_2 is diamagnetic and therefore have very little magnetic effect. These changes in magnetic properties have an effect on $T2$ and $T2*$ relaxation times. Higher density of HbO_2 in the blood increases the $T2$ and $T2*$ relaxation time and as a result, also increases the contrast of the images. Because the brain regions with higher activity have higher density of HbO_2, high activity can be directly linked to the high intensity regions of the fMRI image. In summary, a frequency in the spectrum of signals retransmitted back by the blood sample, which is excited by an RF signal, indicates the spatial location of the transmission while the $T2$ relaxation time of the transmitted signal indicates the density of the HbO_2 as well as activity level of the particular spatial location.

2.2 Related Work

The fMRI data has been widely used in the studies of between-group statistics to identify the abnormal regions related to the ADHD subjects. While group level studies are definitely helpful for understanding the problem, they are not that useful for automatic diagnosis of the individual subjects. The use of the machine learning approaches on the brain imaging data for the prediction of functional diseases like Alzheimer's and Schizophrenia is very common [31, 36, 37], but automatic classification of the ADHD subjects is a relatively new field. Among the first few efforts, Zhu et al. [82] used rs-fMRI data to predict the ADHD labels of the subjects. Later, the release of

the ADHD-200 competition data set motivated a series of studies [6, 7, 16, 17, 19, 24, 30, 56, 63, 68] to be published related to the diagnosis of the ADHD subjects. We grouped these works based on the main approaches or features used to solve the problem. Many of the works fall under the multiple groups as they used different methods or features to compare the performances.

2.2.1 Regional Homogeneity (ReHo)

One of the earliest efforts made for the classification of the ADHD subjects using their rs-fMRI data used the regional homogeneity of brain activity as the feature for the classification process [82]. For each voxel of a brain volume, the regional homogeneity is measured with K nearest neighbor voxels using the Kendall's Coefficient of Concordance (KKC). This is a measure to determine how synchronous a voxel activity pattern is with its locality. Finally, the combination of the Principal Component Analysis (PCA) and Fisher Discriminative Analysis (FDA) are used for the classification. The result is inconclusive as the experiments are performed on a data set containing only 20 subjects. ReHo feature is also used in some of the studies performed on the ADHD-200 data set [17, 24, 63]. While ReHo can measure the similarity of the activity patterns in a local region, it completely ignores the similarity/dissimilarity of the activities of the regions which are spatially far from each others. Thus, it fails to capture a global picture.

2.2.2 Functional Connectivity Network (FC-Nw)

FC-Nw is produced by segmenting the brain volume into different Regions Of Interest (ROIs) and representing each ROI as a node of the network. Segmentation of brain into ROIs can be performed using different criteria such as the functional homogeneity or structural similarity. Intensity time series for each node of the network is then computed by averaging the intensity time series of all the voxels belonging to the node. Correlations of the time series of all pairs of nodes of the FC-Nw produce the edge weights. Different variations of correlation are used in different methods to compute the FC-Nw. Dai et al. [24] used Pearson's correlation coefficient to compute

correlations of average time series of 351 functionally homogeneous ROIs of CC400 map produced by Craddock et al. [22]. These correlation weights are used as the features for the classification. Bohland et al. [6] used AAL atlas of 116 ROIs to compute FC-Nws using three variations of correlation - Pearson's correlation coefficient, Sparse regularized Inverse Covariance and Patel's Kappa. Different local and global network features are computed for classification. Eloyan et al. [30] used 5 regions of motor network and 264 seed voxels to compute two different FC-Nws using Pearson's correlation coefficient. Network edge weights are used for final prediction. Colby et al. [19] used Harvard-Oxford atlas with 100 ROIs and CC400 map to compute two different FC-Nws. Classification is performed using network edge weights as the features. A similar FC-Nws formation technique is used by Cheng et al. [17]. The FC-Nw is a more sophisticated and efficient approach to model the brain functional activity patterns than the ReHo feature. This is because the FC-Nw can capture the functional similarities of the regions which are in close spatial proximity as well as far from each others. Often the networks cannot be used directly for the classification because of the very high dimensionality and a careful feature selection technique is needed on those cases.

2.2.3 Fractional Amplitude of Low-frequency Fluctuation (fALFF)

fALFF of a signal is defined as the power of the signal in a given low frequency range divided by the total power in the entire detectable frequency range. The low frequency fluctuation of the activity pattern is a basic characteristics of the resting state brain and can be used as a bio marker for the prediction of the ADHD label of the test subjects. Cheng et al. [17] computed fALFF score for each voxel of the brain volume in the frequency range of $0.009 0.08$Hz. Sato et al. [63] also computed voxel level fALFF score for the frequency range of $0.01 0.08$ Hz.

2.2.4 Structural Image Features

Structural images are also proved to be useful for the automatic classification of the ADHD subjects. Chang et al. [16] computed the Local Binary Patterns (LBP) texture feature from sMRI data provided with ADHD-200 data set. For each voxel of the brain volume three LBP scores are computed for three orthogonal plane directions. Next, each subject is represented by a combined histogram of LBP scores computed for the three plane directions. LBP scores in each plane direction can have 256 different values. Hence, the size of the combined histogram is $3 \times 256 = 768$ where each bin of the histogram represents the number of voxels with a particular LBP score. These histograms are used for the training of the classifier and ADHD label prediction of the test subjects. In some other papers, structural features from different cortical and non-cortical brain regions are computed. Dai et al. [24] used the cortical thickness and gray matter probability as the structural image features. Bohland et al. [6] used the average cortical thickness, surface area, volume, mean curvature and standard deviation of these measures for each cortical area of interest and subcortical gray and white matter structures. Colby et al. [19] computed the number of surface vertices, surface area, gray mater volume, average cortical thickness and standard deviation, cortical mean curvature, cortical folding index and cortical curvature index from 34 cortical regions and the regional volume, regional voxel intensity mean and standard deviation from 45 non-cortical regions. Structural images provide a different perspective to approach the ADHD subject classification problem. The data helps to verify if the brain structural deformities are related to the functional irregularities found in the ADHD subjects.

2.2.5 Phenotypic Information

Many of the studies used the phenotypic information provided for each subject in the data set to improve the prediction accuracies. Brown et al. [7] showed that the use of the phenotypic information only for the prediction of the ADHD label can outperform the imaging data. The

phenotypic information used for the classification includes the data collection site, gender, age, handedness, verbal IQ, performance IQ and full 4 IQ with a logistic classifier. Among other works Bohland et al. [6] used the age, gender, handedness, verbal IQ and Performance IQ, Sidhu et al. [68] used the age, gender, scanning site, verbal IQ, performance IQ and full IQ, Colby et al. [19] used the age, gender, full-scale IQ, handedness, ADHD index measurements, hyperactivityimpulsivityinattentive scores, secondary diagnosis and medication status. While the phenotypic information is an indirect measure and does not provides any insight about the brain functional or structural abnormalities, it can some times help boosting the classification accuracy when used with imaging data.

Table 2.1: Summary of the training and test sets from different data centers released for the ADHD-200 global competition.

Center	Sub Cnt	Age (yrs.)	Male	Female	Control	Combined	Hyperactive	Inattentive
Released data set								
KKI	78	8-13	42	36	57	16	1	4
NeuroIMAGE	39	11-22	25	14	22	11	6	0
NYU	176	7-18	111	65	87	57	1	31
OHSU	66	7-12	34	32	38	15	1	12
Peking	183	8-17	135	48	114	22	0	47
Pittsburg	89	10-20	46	43	89	0	0	0
Washington	61	7-22	33	28	61	0	0	0
Holdout data set								
KKI	11	8-12	10	1	8	3	0	0
NeuroIMAGE	25	13-26	12	13	14	11	0	0
NYU	41	7-17	28	13	12	22	0	7
OHSU	34	7-12	17	17	27	5	1	1
Peking	51	8-15	32	19	27	9	1	14
Pittsburg	9	14-17	7	2	5	0	0	4
Brown	26	8-18	9	17	-	-	-	-

2.3 Data Set and Preprocessing Steps

We used the ADHD-200 data set for all the experimental validations of our methods. The following sections describe the data set and the preprocessing steps needed to make the data useful for any further analysis.

Table 2.2: Table lists the summary of the scan parameters for all the data centers.

	TR/TE (ms)	Slices	Thickness (mm)	FoV Read (mm)	FoV Phase (%)	Flip Angel (degree)
KKI	2500/30	47	3.0	256	100	75
NeuroIMAGE	1960/40	37	3.0	224	100	80
NYU	2000/15	33	4.0	240	80	90
OHSU	2500/30	36	3.8	240	100	90
Peking	2000/30	33	3.5	200	100	90
Pittsburgh released	1500/29	29	4.0	200	100	70
Pittsburgh holdout	3000/30	46	3.5	240	100	90
Washington	2500/27	32	4.0	256	100	90
Brown	2000/25	35	3.0	192	100	90

2.3.1 Data Set

The ADHD-200 data set is prepared and publicly shared by the Neuro Bureau. Eight different centers contributed to the compilation of the whole data set, which makes it diverse as well as complex. The following abbreviations for the data centers are used throughout the dissertation: Kennedy Krieger Institute (KKI), Neuro Image Sample (NeuroImage), New York University (NYU), Oregon Health and Science University (OHSU), Peking University (Peking), University of Pittsburg (Pittsburgh), Washington University in St. Louis (Washington) and Brown university (Brown).

The data for the competition was released in two stages. In the first stage data from the seven data centers, containing in total 776 subjects, was released for the training of the classification model. Throughout the dissertation we refer to it as the released data set. Later, data for 197 subjects from the seven data centers was released without the label (ADHD or control) information for validation of the performance of the trained classification model. We refer to it as the holdout data set. After the competition, labels for the holdout data set were released for the research community. Mainly three different categories of data, including structural data, functional data and phenotypic information, are provided for each subject in the data set. Structural data contains 3D structural brain image of a subject. The voxel resolution ($1 \times 1 \times 1$ mm) of the structural data is four times higher than the functional data. Along with the whole brain images, Gray Matter

(GM), White Matter (WM) and CerebroSpinal Fluid (CSF) images are also provides. These are the segmented images contain only GM, WM, and CSF regions of the brain respectively. The voxel resolutions of these images are same as the whole brain structural images. Functional data contains rs-fMRI data of the brain where subjects are asked not to perform any conscious task while capturing the data. rs-fMRI data can be assumed as a 3D video of the brain function captured at a voxel resolution of $4 \times 4 \times 4$ mm. Different phenotypic information, such as the age, gender, handedness, IQ, is also provided for each subject. In our study, we used the rs-fMRI data, GM images and male-female phenotypic information.

Based on the information provided with the phenotypic data, we excluded all those subjects from our study which have questionable functional image quality ($QC_{Rest_1} = 0$ of the phenotypic data sheet). Consider Table 2.1 for an overview of the data used in our study. Different data centers used different scanners and scanning parameters for capturing data. For example KKI and NeuroIMAGE used the Siemens Trio 3-tesla scanner, OHSU used the Siemens Magnetom TrioTim syngo MR B17 scanner and Peking used the Siemens Magnetom TrioTim syngo MR B15 scanner. Some important scanning parameters used by the data centers are listed in Table 2.2. Also different data acquisition parameters are used by different data centers such as KKI and NeuroIMAGE captured data with subjects' eyes closed, OHSU and Peking asked their subjects to keep their eyes open. While OHSU showed a fixation cross at the screen, Peking didn't show anything. All research conducted by the ADHD-200 data contributing sites were performed with local IRB approval, and contributed in compliance with local IRB protocols. In compliance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rules, all the data used for the experiments of this dissertation are fully anonymized. The competition organizers made sure that the 18 patient identifiers as well as face information are removed.

2.3.2 Data Preprocessing

The recorded fMRI data need to be preprocessed before it can be useful for any analysis. For all our experiments we used the preprocessed resting state fMRI data released for the competition. The preprocessing is performed by the competition organizers using the AFNI [21] and FSL [40] tools and computed on the Athena computer clusters at the Virginia Tech advance research computing center. The main preprocessing steps performed on the fMRI data are described in the following paragraphs.

The first preprocessing step requires the slice timing correction. fMRI data can be assumed as a video of brain activity where in each time stamp a 3D image of the brain functioning is captured. These 3D images are formed by scanning the brain slices one after another. As a result each slice represents the brain activity at a different time point. To correct this problem a temporal interpolation method is used such that it appears that the data for all the slices of a brain volume is acquired at exactly the same time.

The next common preprocessing step is the head motion correction. During the scanning process the subjects might slightly move its head. As a result, the brain regions in different 3D images are not exactly superposed with each other. To fix this problem each of the 3D images is transformed using rotation and translation so that the different brain regions are aligned in the whole video.

The third main preprocessing step involves the registration of the brain volumes of the individual subjects onto a common template space. As the sizes and shapes of the brains may vary a lot for the subjects under consideration, the voxels with same coordinates in fMRI data may belong to the different brain regions for the different subjects. To solve this problem the data is registered on the $4 \times 4 \times 4$ mm voxel resolution Montreal Neurological Institute (MNI) space which is a common space on which further analysis can be performed.

Next the data is bandpass filtered (0.009 Hz $<$f $<$0.08 Hz) in the temporal domain to

exclude all the frequencies which are not relevant for the analysis of the resting state functional connectivity. Finally, to remove the noise, the data is blurred by convolving with a 6-mm Full Width at Half Maximum (FWHM) Gaussian filter. All fMRI data volumes are of size $49 \times 58 \times 47$ voxels, but the number of samples across time varies among the data capturing centers.

Structural images preprocessing involves removing skull images from the data, segmenting the images into GM, WM and CSF regions, and transforming the images to a template space. All the structural images have voxel resolution of $197 \times 233 \times 189$. For further information about the data and preprocessing steps and how to access the freely available data we refer the interested readers to the following web document [53].

2.4 Summary

In this chapter we provide a short description of how brain functions and explained the fundamental concept behind the brain imaging techniques. We listed the commonly used brain imaging techniques and provide a detailed description of the fMRI data capturing process. In the related work section we introduced the already existing techniques for the automatic detection of the ADHD subjects using the brain imaging data as well as the phenotypic information. Finally, we provide a detailed description of the ADHD-200 data set which is used in this dissertation for the experimental validation.

In the next four chapters we describe our method for solving the proposed problem. The first approach uses the BoW framework to compute the histogram of the brain functional network features. In the second approach we analyse the importance of the network features in more details for the classification of the ADHD subjects. The third approach modeled the functional brain networks as attributed graphs and uses the inter-network distances for projecting the networks in a low dimensional space for the efficient classification. In the fourth approach we combined the structural and functional imaging data to further improve our classification accuracy. Finally, in

the last section we provide a summary of the dissertation and the possible future works.

CHAPTER 3: BAG-OF-WORDS FRAMEWORK FOR THE DIAGNOSIS OF ADHD

The BoW approach, originated in the natural language processing, allows a dictionary-based modeling of the documents. The framework represents each document as a bag containing a subset of words from the dictionary where each word in the document can occur multiple times. This type of approach has also been popular in the Computer Vision area and has been applied to many problems such as the image or video representation [32, 47, 48]. In this chapter, we introduce the BoW approach to the biomedical imaging community, specifically for the processing of the functional brain networks for the automatic detection of the ADHD subjects. The following sections present an overview of BoW framework, our method of classification using the framework, experimental details and a discussion of the significance of the work.

3.1 BoW Overview

The idea of BoW framework originated in the area of document classification [23,43]. This is based on a simple idea which says the class of a document can be determined from the number of occurrences of the words in the document. Following the idea, a document is represented by a histogram where each bin of the histogram represents the number of occurrences of a distinct word of the document. A dictionary is constructed containing all the distinct words considered for the classification model. The number of bin count of the histogram represents the size of the dictionary. The framework is named BoW as a document is represented by the count of occurrences of all the distinct words only, ignoring the grammar and order of the words. Finally, a classifier can be trained based on the histogram representations of different examples of training documents. Given the histogram representation, the class of any unknown document can be determined using the trained classifier.

The same idea is adapted in the computer vision area for the representations of the images and videos as the bag of visual words. The main problem of incorporating the BoW framework in the computer vision is the construction of the visual word dictionary. This is because unlike in the case of words for document classification, visual words are not easily identifiable. For the purpose of constructing the visual word dictionary, first each image or video is represented as a set of local features. Next, all the local features from all the training samples are represented in the feature space where they are clustered using the K-mean clustering algorithm. Each of the clusters forms a codeword which can be considered as a set of similar patches. These codewords have similar functionality as words for a dictionary. Visual word dictionary, also referred as codebook, is constructed using the collection of all the different codewords generated. For a given image or video, its bag of visual words representation is constructed by computing the local features, assigning the computed features to the most similar codewords, and forming the histogram of codewords. Once the histograms for training and test samples are constructed, classification can be performed in a similar way as in the case of the document classification.

3.2 Method

The overview of our approach is depicted in Figure 3.1. The first step of our approach is the brain functional connectivity network construction followed by the network feature extraction, representation of each subject as a histogram following BoW framework, and classification using the SVM.

3.2.1 Functional Connectivity Network Construction

We assume that the activity of a brain can be modeled as a functional connectivity network constructed by connecting different brain regions. To construct the network, each voxel of the brain volume is represented as a node and any two nodes of the network are connected with an

edge if they show high similarity of activity patterns over the time domain. In this chapter we have used the terms voxel and network node interchangeably with the similar meaning.

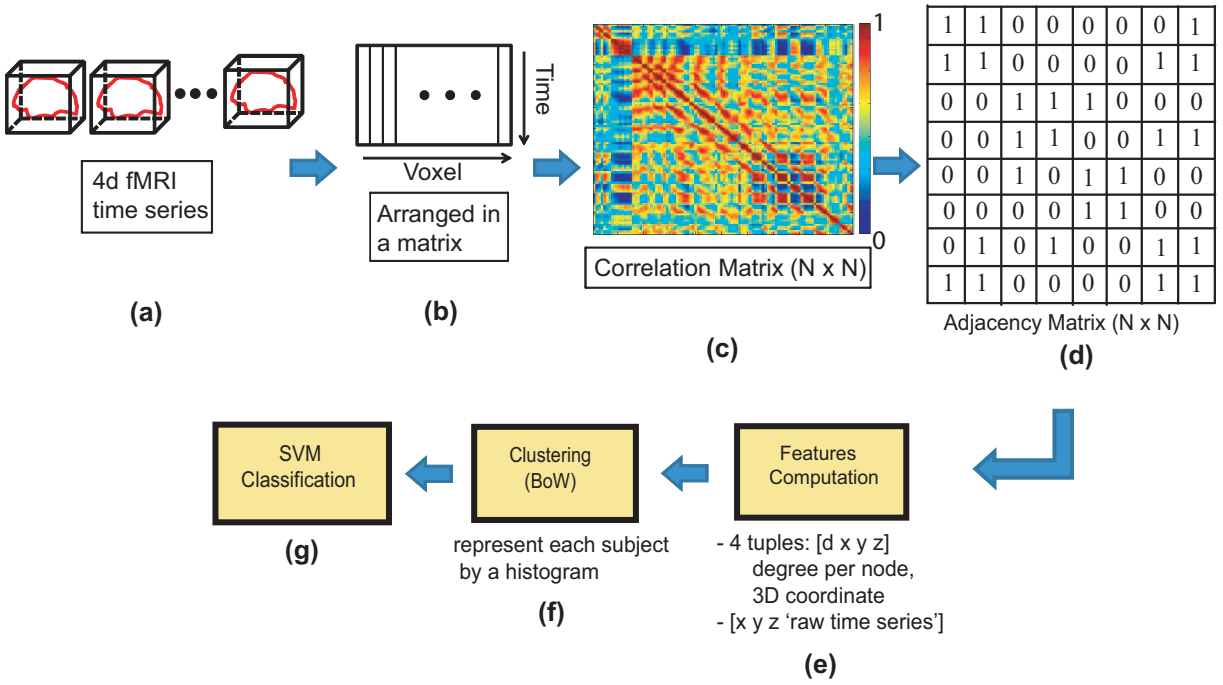


Figure 3.1: Overview of our approach: First (a) the 4D fMRI data is (b) reorganized in a matrix where each column of the matrix is the intensity time series of a voxel. (c) Next, we compute an $N \times N$ matrix which contains correlation values of pairs of voxel time series (N is the number of voxels inside the anatomical brain mask). (d) The adjacency matrix is formed by thresholding the entries of the correlation matrix. (e) The features such as the degree per node and raw intensity time series for each voxel are used for (f) BoW codebook generation. (g) Finally, classification is performed using an SVM.

As the first step of the algorithm, we extract the intensity time series for all the voxels of the brain volume and reorganize them in a 2-D matrix. Please note that the intensity time series of each voxel contains the information of its pattern of activities. This is illustrated in Figure 5.1 (b). Next, the correlations between all possible voxel pairs is computed as the measure of their similarity of activity patterns. If a subject contains N number of voxels, a correlation matrix of size $N \times N$ is constructed, where the i^{th} row of the matrix corresponds to the pairwise correlation values of the i^{th} voxel with all other voxels of the brain volume. The anatomical mask provided with the ADHD-200 data set is used to identify the voxels belonging to the brain volume.

For any two voxels u and v, if the time series are $\mathbf{u} = [u_1, u_2, ..., u_T]$ and $\mathbf{v} = [v_1, v_2, ..., v_T]$ respectively, the correlation can be computed as,

$$r = \frac{(T\sum_{i=1}^{T} u_i v_i) - (\sum_{i=1}^{T} u_i)(\sum_{i=1}^{T} v_i)}{\sqrt{[T\sum_{i=1}^{T} u_i^2 - (\sum_{i=1}^{T} u_i)^2][T\sum_{i=1}^{T} v_i^2 - (\sum_{i=1}^{T} v_i)^2]}}, \tag{3.1}$$

where T is the length of the time series.

Before we compute the correlations, the time series are normalized between $[-1, 1]$. Next, we threshold all the values of the correlation matrix to get a binary map of zeros and ones. We empirically choose the correlation threshold value as 0.80 and zeroed in all the absolute correlation values lower than that. This binary map can be considered as the adjacency matrix of the network where the i^{th} node is connected to all the nodes for which non-zero values are present in the corresponding column positions of the i^{th} row of the matrix. Note that we can consider two voxels to be connected by an edge when the correlation is high positive, high negative or simply the absolute value of the correlation is high. We have computed three different sets of networks considering high positive, high negative and high absolute correlation values respectively. As we consistently achieved higher detection accuracies using the networks with positive correlation val-

ues compared to the two other types of networks, all the experimental results reported are on the positive correlation networks only.

3.2.2 Network Feature Extraction

Once the functional connectivity networks for all the subjects are constructed, we extract degree feature from each node of the networks. As it is known, the degree of a node is the number edges incident on it. The degree for the i^{th} node of the network can easily be calculated by summing up the values of the i^{th} row of the adjacency matrix. Finally we represent each node as a 4-tuple $[d, x, y, z]$, where d is the degree and x, y, z are the 3D coordinates of the node. Adding the 3-D coordinates helps us to capture the spatial information of the node. Please note that the x, y, z and d are normalized to have values between 0 and 1.

3.2.3 BoW Histogram Representation

In the next step, following the BoW representation, we represent each subject by a histogram of codewords. The codewords are generated by extracting network features from each of the subjects under consideration and clustering the features in the feature space using the K-means clustering algorithm [2, 72]. As stated, our feature vector for each node of a network is a 4-tuple $[d, x, y, z]$. To be clear, a feature vector is constructed for each node of the networks corresponding to the subjects in the training data set and clustering is performed on all the feature vectors generated for the training set. The number of clusters used for the K-means clustering is the size of the codebook generated as well as it defines the bin count of the histogram.

For our experiments we empirically selected the cluster and histogram bin count ($K = 100$) where each bin is represented by the center of the corresponding cluster. Once the codebook is generated, any subject can be represented as the histogram of 4-tuple features by mapping the features to the nearest cluster centers in the 4-D feature space. Thus, the histogram representation for each subject captures the occurrences of each code words.

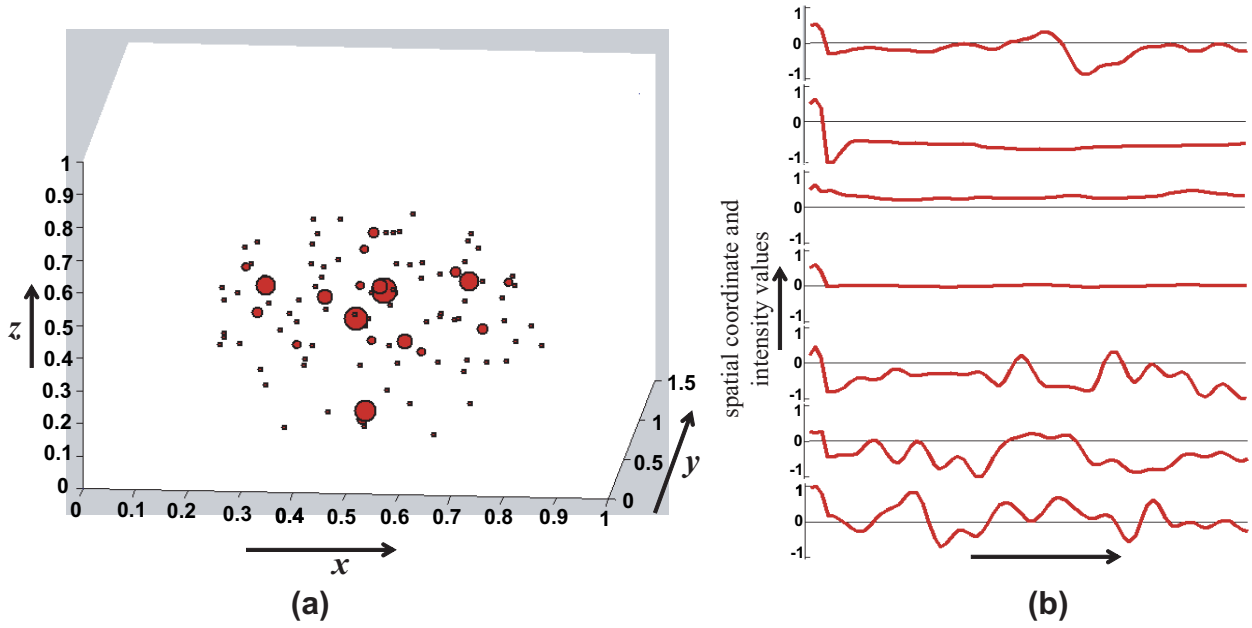


Figure 3.2: The figure shows the clusters formed using the K-mean clustering on the features computed from the training examples. (a) The $[d, x, y, z]$ 4-tuple clusters are plotted on the x, y, z space while the size of the clusters are proportional to the degree d. (b) Few of the raw intensity time series clusters are plotted among 75 different clusters due to space constraint.

To show the importance of network feature we used a different approach to compute histograms. Instead of the degree feature, we represent each voxel with their intensity time series. Formally, the feature vector for any voxel u is constructed as $[x, y, z, \mathbf{u}]$ where $\mathbf{u} = [u_1, u_2, ..., u_T]$ is the intensity time series of the voxel u. Please note that the different data centers of ADHD-200 data set have different time length for the fMRI data. To keep the length of the intensity time series equal, we consider only the first 72 time stamps which is the smallest length of the fMRI data for any of the subject of the data set. Hence, all of our time series feature vectors are of length $3 + 72 = 75$. Following the same steps as in the network features, we generate a codebook of 75 codewords and represent each subject by a histogram of 75 bins. Again the bin count is empirically selected.

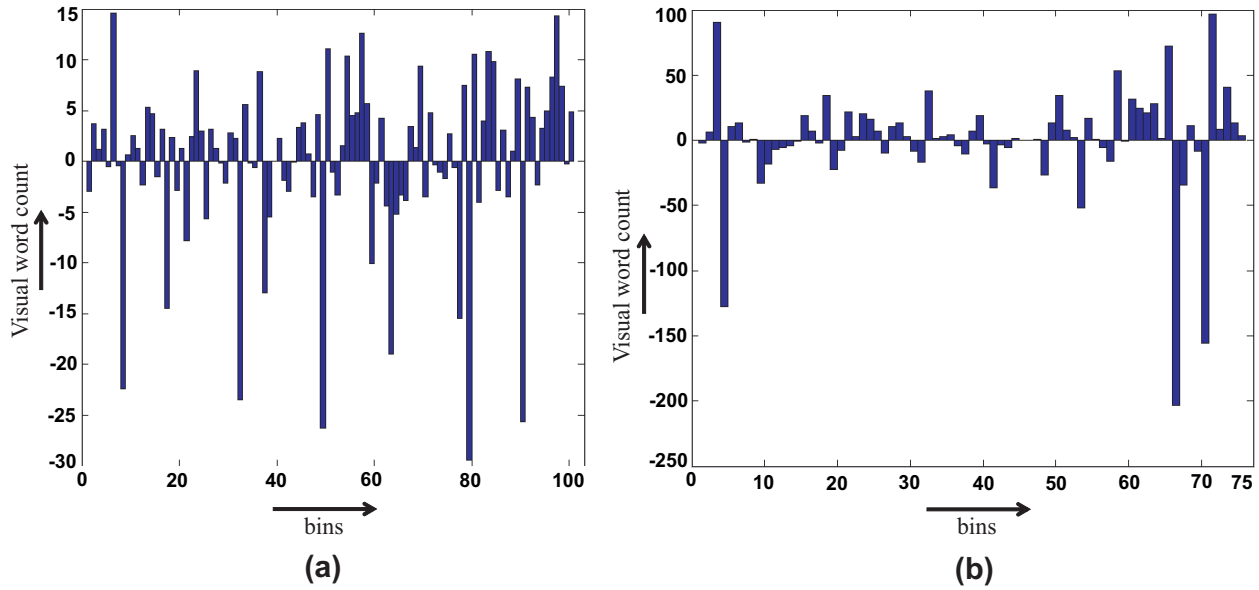


Figure 3.3: The figure shows the differences of average histograms of the control and ADHD group of subjects for (a) 4-tuple degree features and (b) raw intensity time series features.

A third approach is used to combine network and raw intensity features to generate the third type of histograms. For this purpose, we concatenate the normalized network feature and raw intensity feature histograms to represent each subject by a 175 dimensional histogram. Figure 3.4 explains the histogram generation process.

Figure 3.2 shows the examples of the clusters formed on the 4-tuple ($[d, x, y, z]$) and raw intensity time series features. 4-tuple clusters are plotted in the x, y, z space where the size of the clusters are proportional to the degrees d. The intensity time series clusters are plotted as $xyz + time\ stamps$ vs $spatial\ coordinate\ and\ intensity\ values$. Due to space constraint only a few of the 75 clusters are shown in the figure.

To find out if the histograms can capture the differences of the ADHD and control groups of subjects we construct Figure 3.3. The figure shows the average differences of the histograms corresponding to the Control and ADHD groups of subjects. All the subjects of the ADHD-200

released sets are used to construct the subjects. The positive bin counts represent a higher average codewords counts for the control subjects while the negative bin counts represent the opposite.

3.2.4 Classification

Finally, the SVM [15] with histogram intersection kernel is used for the classification. First, the SVM is trained using the histograms generated for the subjects in the training set. Given the histogram of a test subject, the trained SVM is used to classify the subject into the ADHD or control group. Three different sets of classification experiments are performed using the network feature histogram, raw intensity feature histogram and combined histogram.

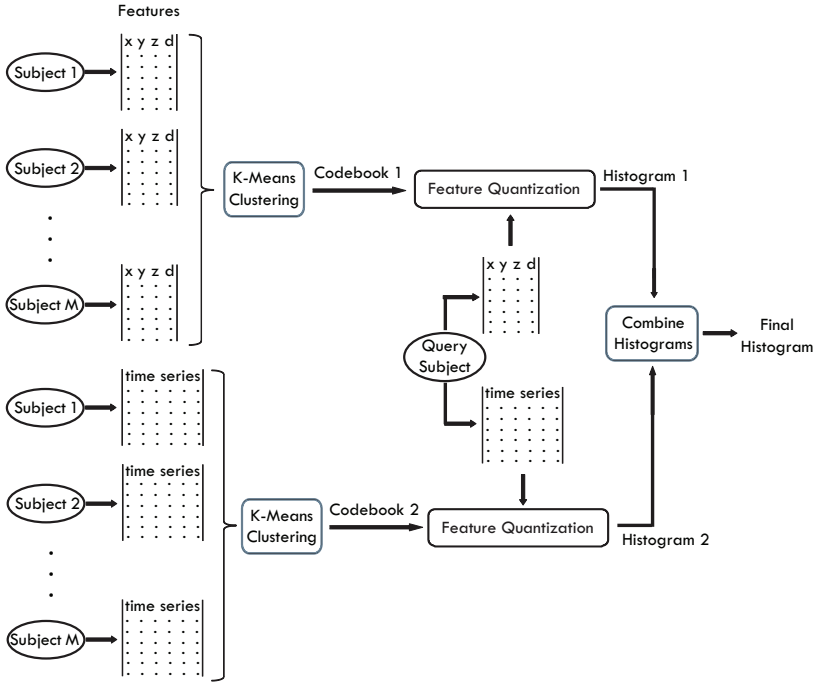


Figure 3.4: Overview of our BoW approach.

3.3 Experiments and Results

For the experimental validation we selected 506 subjects from across the released sets of the seven data centers. A brief description of all the subjects used in the experiment is included in Table 3.1. The classification is performed in a leave one out cross validation fashion, i.e. in each iteration a single subject is used for the test while rest of the subjects are used for the training of the SVM classifier. Hence, the training and testing are performed 506 times, each time choosing a separate subject for testing and using the rest of the subjects for the training of the classifier. Also, note that we performed three sets of experiments for the histograms using the raw intensity feature, the network feature and concatenation of the intensity and network features. The Receiver Operating Characteristic (ROC) curve, which is obtained by varying the confidence of detection, is shown in Figure 3.5 for all three sets of experiments. The best classification accuracies for all three of the experiments are included in Table 3.2. As it can be seen the network features perform better than the raw intensity features but the combined features perform the best. The best detection rate obtained is 64% at the cost of 0.50 sensitivity and 0.72 specificity.

Table 3.1: Description of the test subjects of the larger data set.

Test Center	Number of Subjects	Number of ADHD conditioned subjects	Number of control subjects	Female	Male
KKI	83	22	61	37	46
Neuro Image	48	25	23	17	31
NYU part 1	55	31	24	19	36
NYU part 2	67	32	35	22	45
OHSU	79	37	42	36	43
Peking 1	85	24	61	49	36
Pittsburgh	89	0	89	43	46

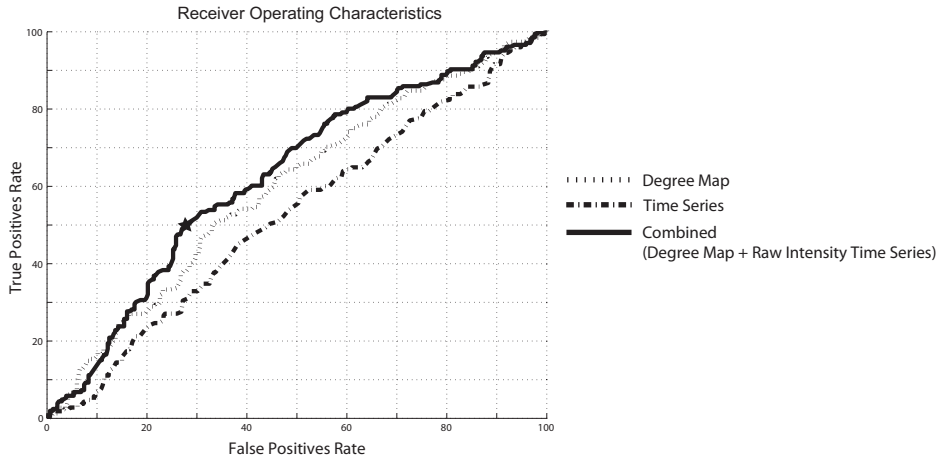


Figure 3.5: Receiver Operating Characteristics curves for different combinations of features on 506 subjects.

To verify the performance of our algorithm on the holdout sets, we performed another set of experiments. The experiments are performed on the five holdout sets which are reported in Table 3.3 along with their detection accuracies. To conduct this set of experiments, five SVM classifiers are trained separately on the corresponding released sets and tested on the holdout sets. Similar to the first set of experiments, we achieve overall highest detection accuracy (64.81% with 0.5341 sensitivity and 0.7416 specificity) when combined features are used to construct the 175 dimensional histograms.

Table 3.2: Summarize the detection rates of the ADHD classification results using three different types of histograms.

Used Feature	**Number of Subjects**	**Accuracy**
Degree Map	506	61%
Raw Intensity Time Series	506	56%
Degree Map+Raw Intensity Time Series	506	64%

3.4 Discussion

In this chapter we show that the brain function can be modeled as a connectivity network and the network topological differences of the ADHD and control group of subjects can be utilized for the prediction of their ADHD label. To capture the network topological information we express each node by its degree and 3D spatial coordinate and represent each subject as a 100 dimensional histogram of network features. We also represent each subject as a 75 dimensional histogram of intensity time series and a 175 dimensional combined histogram of network + intensity time series to compare the classification performance. As it can be seen, the detection accuracy using the network feature histograms is better than the intensity time series histograms. This shows the effectiveness of modeling the brain function as a network. It also indicates the presence of topological differences in functional connectivity networks between the ADHD and control group of subjects. Finally, the combined histogram performs best, which suggests that the network and time series representation captures complimentary information.

Table 3.3: Shows the detection rates of the classification experiments on the holdout sets released for the ADHD-200 competition.

	Accuracy (%)		
	Degree Map	**Intensity Time Series**	**Deg. + Intensity Time Serie**
KKI	81.82	72.73	81.82
Neuro Image	60.00	60.00	68.00
NYU	68.29	31.71	56.10
OHSU	61.76	82.35	70.59
Peking	54.90	52.94	62.75
Overall	62.35	56.17	64.81

One of the shortcomings of the method is the loss of spatial information while constructing the codewords using K-mean clustering. This is because each cluster is represented by their center which is the average of the cluster volume in the feature space. Also, in this framework we gave equal importance to all the nodes of the network even though some nodes may not be active during the resting state of the brain. Including features from all the nodes in the classification framework

can unnecessarily increase the feature dimensions which might negatively impact the classification accuracy. Finally, we analyze only the degree features for the classification of the ADHD subjects while there might be other features which are useful for the proposed problem. We address these issues in the next chapter.

CHAPTER 4: NETWORK FEATURES FOR THE ADHD DETECTION

In the previous chapter we showed that the brain functional activity can be modeled as a network where the network features, such as the degree of each node of the network, can be useful for the classification of the ADHD subject. In this chapter we further investigate the usefulness of the network features and therefore compute more complex features such as the cycles, the varying distance degree and the edge weight sum of the nodes along with the node degree. Moreover, we propose that the voxels from the whole brain are not useful but only some specific brain regions (group of voxels) contain information to distinguish the ADHD and control groups of subjects. For this purpose we developed an algorithm to identify the useful brain regions and we demonstrate that using the features only from the regions identified by our algorithm help to improve the classification accuracy. Throughout this chapter we refer to the useful regions identified by our algorithm as the useful region mask. Finally, we show that our finding is consistent with the other studies which are aimed to find the brain regions responsible for ADHD.

4.1 Method

Network motifs such as the distribution of node-degree, cycles etc. are analyzed in different disciplines of science including neuroscience [73], [51], [49]. We propose to use different graph theoretic concepts for our study. We assume that the different brain regions need to cooperate with each other for the proper functioning of the brain. These cooperations of the regions manifest in the fMRI data in the form of the correlations of their activity patterns. We modeled the correlations of the brain regions as a network with the belief that the network structures of the ADHD and control groups of subjects have sufficient differences to be used by the machine learning approaches for the automatic classification.

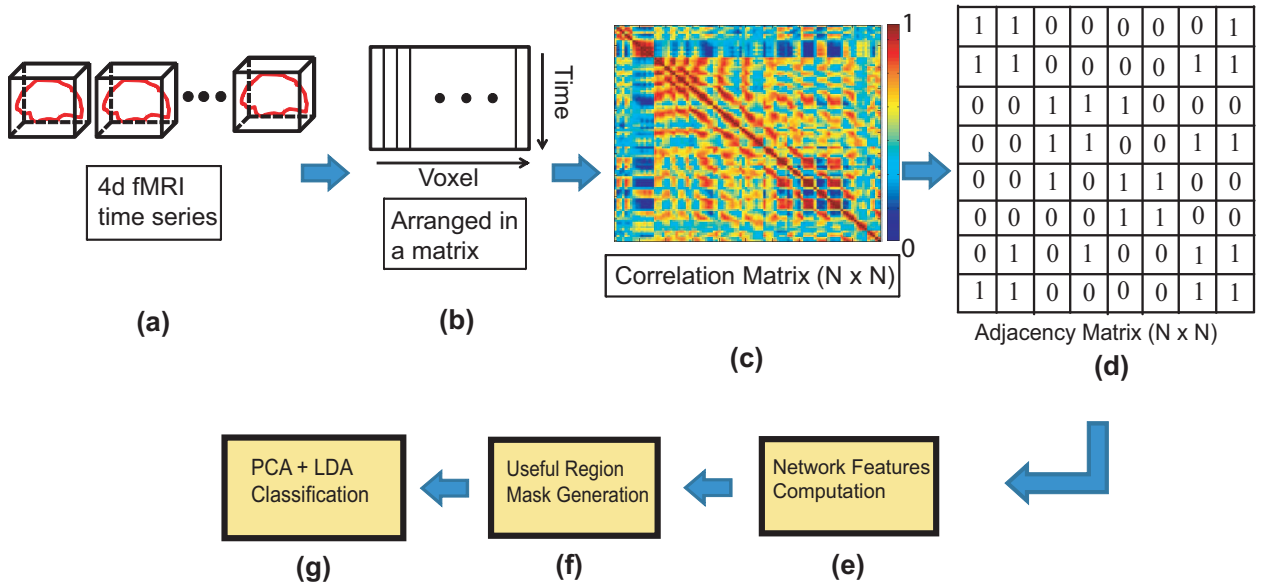


Figure 4.1: Overview of our approach: (a) given the 4-d fMRI data for a subject, (b) first we rearrange it as a matrix. (c) Next, a correlation matrix of size $N \times N$ (N is the number of voxels) is computed. (d) An adjacency matrix is generated after thresholding the correlation values into binary numbers. The adjacency matrix represents a network. (e) Network features such as the node degree and cycle count for each node of the network are computed. (f) Next, we generate the useful region mask. (g) Feature values from the nodes, identified by the useful region mask, are used to form the feature vector and a PCA-LDA classifier is used for the classification.

Figure 4.1 shows the flow chart of our classification model. The first step of our method is the computation of the functional connectivity network which is exactly the same as described in Section 3.2.1. The rest of the steps are described in the next few sections.

4.1.1 Network Feature Computation

Once the functional networks for each of the subjects in the data set is constructed, we compute different network features. The network features are expected to capture the structural information of the networks and exploit the network topological differences to segment the ADHD subjects from the control subjects. The features computed from all the nodes of a network are

referred to as the feature map, such as the degree Map, cycle Map etc. The descriptions of the different network features computed are given below.

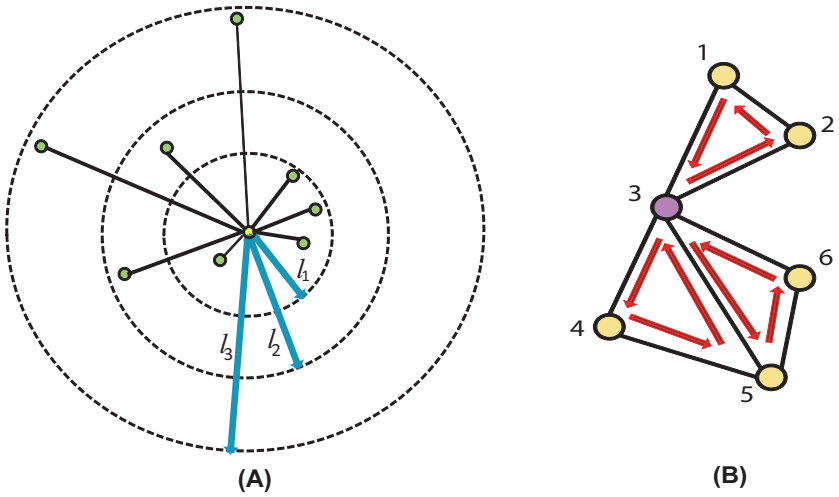


Figure 4.2: **(A)** The degree of the node, highlighted in yellow, is the count of all the green nodes connected to it (i.e. 8), while the varying distance degree is the counts of all the connected nodes in each of the bins defined by the three edge length thresholds (l_1, l_2, l_3) showed by the blue arrows. In this example the varying distance degrees of the yellow node are $\{4, 2, 2\}$. **(B)** Shows all the distinct 3-cycles that contain node 3.

Degree: The degree of a node in a network is the number of other nodes connected to the node. In other words, the degree of a node is the number of edges incident on it.

Varying Distance Degree: Instead of considering the count of all the edges of a node as its degree, we group the edges based on their physical length and compute a separate degree for each of the groups. So, if we have n threshold values for edge length, say $\{l_1, l_2, ..., l_n\}$, we can compute n degrees, $\{d_1, d_2, ..., d_n\}$, of a node v, where d_i is the count of all the edges connected to v with length between l_{i-1} to l_i. Refer to Figure 4.2 for details. We use the Euclidian distance for the computation of the edge length. For our experiments, we used threshold values of 20, 40, and 80 mm., where the average brain volume is approximately of size $172 \times 140 \times 140$ mm. Therefore, we get 4 degrees per node which are the counts of the edges of length 0-20, 20-40, 40-80 and greater

than 80 mm. respectively. The thresholds are selected in an intuitive fashion such that the different degrees capture local to global connectivity patterns. The percentage of average edge counts in the length range of 0-20, 20-40, 40-80 and above 80 mm are computed as 70.44%, 16.54%, 8.40% and 4.62% respectively.

L-cycle Count: A path in a network is a sequence of distinct nodes which can be traversed in a given order using the connecting edges. A cycle, on the other hand, is a closed path in the network where the starting and ending node is the same and all other nodes are distinct. The L-cycle count of a node is the number of all possible distinct L length cycles containing the node. Figure 4.2 illustrates this idea. L-cycle count for a node is computed by traversing through all the L-length paths starting from the node and counting the paths which lead to the starting node. The traversing can be performed using the breadth first search algorithm. We used the 3-cycle and 4-cycle count features for our experiments.

Weight Sum: Instead of binarizing the values of the adjacency matrix, we use the actual correlation values, if it is greater than a threshold, of voxel pairs as the edge weights. As the correlation values can be positive or negative, we separately add up all the positive, negative and absolute edge weights of a node to get its sum of positive, negative and absolute weights.

4.1.2 PCA-LDA Classification

Once we complete computation of the network features, we extract the features from all of the nodes within the useful region mask. The mask generation algorithm is described in the next subsection. Concatenation of the feature values extracted from all the nodes generates a feature vector per subject. A PCA-LDA based classifier is trained separately using different sets of the feature vectors computed for different types of the network features. Finally, the trained classifier is used for the automatic classification of the ADHD subjects.

It is expected that the topological characteristics of the computed networks are represented by their feature vectors. A feature vector of a network is represented by a point in the feature space

where the dimensionality of the space is the same as the length of the vector. If the feature vectors of the ADHD and control subjects are separable in the feature space, then their corresponding point representations should be clustered at different locations of the feature space. When a classifier is trained, it learns to partition the feature space in such a way that the feature vectors from the separate groups ideally fall under the separate segments of the space. Given the feature vector of a test example, the classifier can identify the specific segment of the feature space it belongs to and classify the test subject accordingly. Linear Discriminant Analysis (LDA) is a widely used data classification technique which maximizes the ratio of between-class variance to the within-class variance to produce maximal separability. Mathematically, the objective is to maximize the following function :

$$J(w) = \frac{w^T S_B w}{w^T S_W w}, \tag{4.1}$$

where S_B and S_W are between class and within class scatter matrix, and can be formulated as follows:

$$S_B = \sum_{i=1}^{n_A} ({x_i}^{(A)} - \mu^{(A)})({x_i}^{(A)} - \mu^{(A)})^T + \sum_{i=1}^{n_C} ({x_i}^{(C)} - \mu^{(C)})({x_i}^{(C)} - \mu^{(C)})^T, \tag{4.2}$$

$$S_W = (\mu^{(A)} - \mu^{(C)})(\mu^{(A)} - \mu^{(C)})^T, \tag{4.3}$$

n_A and n_C are the number of subjects, $\mu^{(A)}$ and $\mu^{(C)}$ are the mean feature vectors, ${x_i}^A$ and ${x_i}^C$ are the ith feature vectors of the ADHD and control group respectively. For all our experiments we used Matlab implementation of the LDA classifier ($classify$ function with $linear$ type of discriminant function).

In many cases, the dimension of the feature space becomes so high that the proper parti-

tioning of the space is difficult. For example, in our case, the dimensions of the feature space is equal to the number of voxels within the useful region mask which is several thousands. Again, most of the dimensions do not contain any significant data variance. The Principal Component Analysis (PCA) is a procedure to find out a set of orthogonal directions, called the principal components, along which the variance of the data is maximum. It then projects the data into the smaller dimensional subspace composed of the principal components. The classifier can work efficiently on the subspace which is significantly smaller in dimension than the original feature space. We use the first 40 and first 100 principal components for the experiments on the KKI and full data set respectively as they cover more than 98% of the data variance. We have included a plot of principal component vs. percent of data variance in Figure 4.7. Refer to [1] for details about PCA.

4.1.3 Useful Region Mask

Different research studies have proposed several Regions Of Interests (ROI) for the brain fMRI data analysis. These different ROIs vary in size and number. In some studies, ROIs are identified based on the anatomical structure of the brain while in some other studies they are segmented based on the homogeneity of the functional activities. Tzourio-Mazoyer et al. [78] identified the ROIs based on the similar functional responses in the brain. Craddock et al. [22] generated a homogenous functional connectivity map from the rs-fMRI data. Smith et al. [70] identified several co-varying functional subnetworks in the resting state brain. However, it is still unclear which ROIs are the best for the resting state functional connectivity network analysis. Also it is not known if all the ROIs detected by one method are required for the ADHD classification or the use of a subset of ROIs would be more efficient. To find out these answers we propose a novel method to identify the useful region mask for the classification of the ADHD and control subjects. The algorithm for the useful region mask generation is as follows:

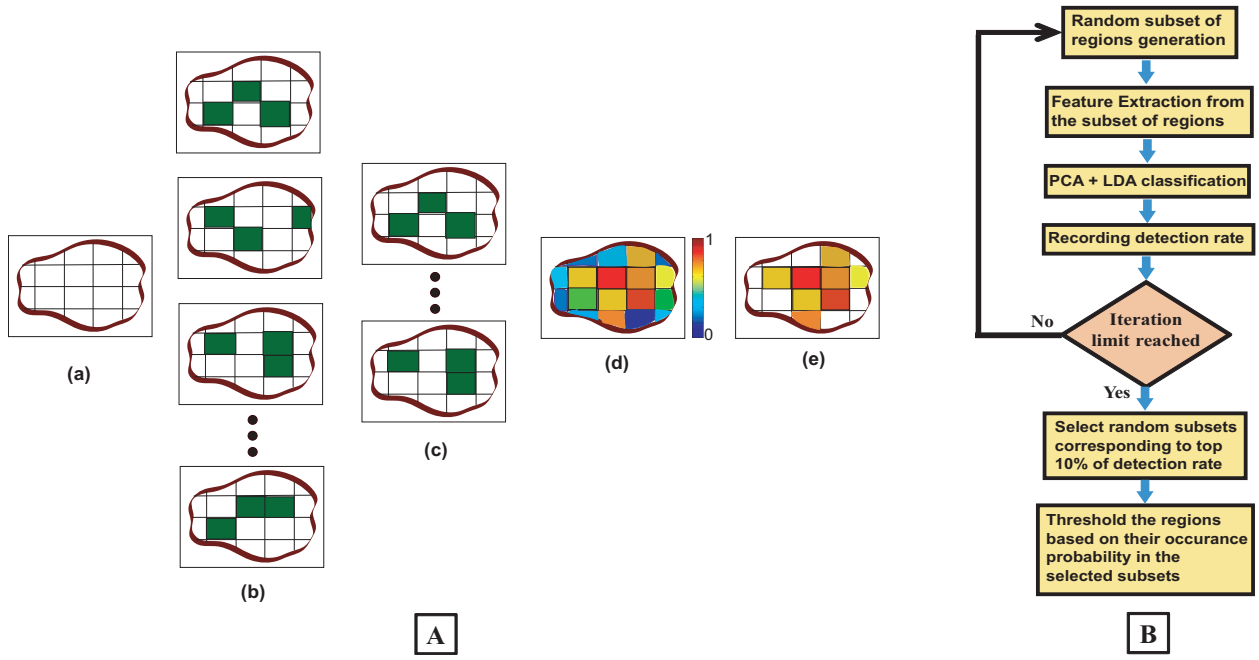


Figure 4.3: **(A)**This part of the figure explains the useful region mask generation algorithm on a single brain slice. The figure is just a graphical example, not the real data. In the actual experiments the brain volumes are used instead of slices and volumetric regions are used instead of square subdivision areas. **(a)** Divide the slice into square regions. **(b)** Select random sub sets of the regions marked in dark green. **(c)** Select the sub sets with top 10% of detection rate. **(d)** Generate a probability map based on the regions occurrence in top 10% subset. **(e)** Threshold the probability map to produce the useful region mask. **(B)** This part shows the flowchart for the mask generation algorithm.

step 1 For each of the subjects used for the mask generation algorithm we do the following:

- Divide the whole brain into small cubicle volumes. Each of the volumes is typically $5 \times 5 \times 5$ voxels except the volumes at the boundary of the brain.
- Select a random subset of the volumes. We include each volume in the subset with probability p.
- Generate a degree map by extracting the degrees for all the voxels within the selected subset of volumes.

step 2 Train the PCA-LDA based classifier and calculate the detection accuracy on the test data set.

step 3 Perform step 1 and step 2 for *m* number of times, each time generating a different random subset, and computing the detection accuracy.

step 4 Choose the random sub sets corresponding to the top 10% of the detection accuracy as the candidates for generating the useful region mask. We count the occurrences of each of the volumes in all of the candidate sub sets and normalize the counts between 0 to 1 after dividing it by the number of candidate sub sets. This gives us the probability of inclusion of each of the volumes in the mask.

step 5 Generate the useful region mask using a threshold *th* to prune the regions with low probability.

Final Threshold *th*

Region Selection Probability *p*	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
0.2	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	66.6667
0.25	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	71.7949	71.7949	66.6667	69.2308
0.3	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	71.7949	71.7949	71.7949	66.6667	66.6667
0.35	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	71.7949	71.7949	74.359	66.6667	66.6667
0.4	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	71.7949	**76.929**	74.359	66.6667	66.6667	64.1026
0.45	69.2308	69.2308	69.2308	69.2308	71.7949	69.2308	69.2308	69.2308	74.359	69.2308	71.7949	69.2308	66.6667
0.5	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	69.2308	64.1026	71.7949	71.7949	69.2308	66.6667	64.1026
0.55	69.2308	69.2308	69.2308	71.7949	69.2308	66.6667	69.2308	71.7949	74.359	71.7949	69.2308	66.6667	64.1026
0.6	69.2308	71.7949	71.7949	69.2308	71.7949	66.6667	69.2308	71.7949	69.2308	69.2308	66.6667	69.2308	66.6667
0.65	69.2308	71.7949	69.2308	69.2308	69.2308	66.6667	64.1026	71.7949	71.7949	69.2308	69.2308	64.1026	61.5385
0.7	69.2308	69.2308	69.2308	66.6667	69.2308	74.359	66.6667	69.2308	66.6667	71.7949	66.6667	66.6667	64.1026
0.75	71.7949	69.2308	69.2308	66.6667	69.2308	71.7949	69.2308	66.6667	69.2308	69.2308	71.7949	64.1026	61.5385
0.8	69.2308	69.2308	66.6667	71.7949	69.2308	69.2308	71.7949	69.2308	69.2308	66.6667	69.2308	66.6667	64.1026

Figure 4.4: Different detection results on KKI data set based on different set of values of p and th.

We experimentally verified that the highest detection rate is achieved when *p* is 0.40 and *th* is 0.60. The details of the experiment is included in Section 4.1.4. The value of *m* was kept at 500 so that the number of iterations should be large enough but computationally feasible. Figure

4.3 (A) is an illustration of the proposed algorithm on a cartoon 2-D slice of a brain while Figure 4.3 (B) is the flowchart for the mask generation algorithm. Note that the other network features may also be used in the algorithm but we only use the degree map feature. We assume that the regions, which are useful for identifying ADHD conditioned brains, should not vary depending on the feature type used for the detection of the mask. We verified the idea by computing the useful region mask using the 3-cycle map features also. We found that the final detection rates are very similar (refer to Section 5.3) which supports our hypothesis.

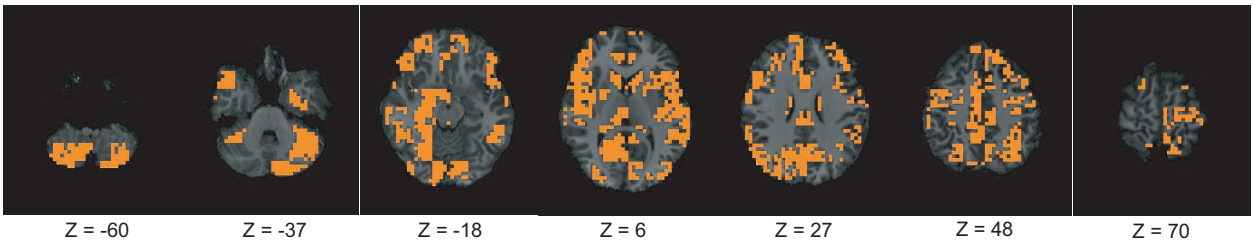


Figure 4.5: The figure shows different brain slices to demonstrate the computed useful region mask. The masked regions are highlighted in orange color and overlaid on the slices of the structural image of a sample subject.

Table 4.1: Shows list of the clusters and their approximate centers, sizes and standard deviations found using the most useful region mask algorithm. The coordinates are calculated on the HarvardOxford-cort-maxprob-thr0-1mm standard atlas provided with the FSL 4.1. We list the ROIs of Harvard-Oxford Cortical and Subcortical Structural Atlases for which more than 50% of the volumes are selected in the useful region mask. Atlas tool of FSL view is used for this purpose.

ROIs	**[x, y, z] centers in mm.**	**size in mm^3**	**standard deviation in mm.**		
			x	**y**	**z**
Precuneus Cortex	[0, -66, 42]	7872	5.4894	6.6435	10.3592
Cingulate Gyrus	[0, -36, 52]; [0, 6, 42]	13056	4.5593	11.3751	10.9128
Temporal Pole	[56, 14, -18]	5312	4.7728	5.5878	5.7664
Superior Temporal Gyrus	[60, -18, -8]; [-60, -20, -4]	3392; 6400	7.1938; 6.6817	9.4413; 11.6393	4.0790; 5.7075
Inferior Temporal Gyrus	[54, -30, -20]; [-60, -48, -10]	1856; 2816	7.6293; 5.4892	6.7262; 8.2390	8.2617; 5.3582
Pre-central Gyrus	[-6, -22, 62]	8000	16.7226	8.5099	5.2886
Lingual Gyrus	[6, -64, 4]	19072	12.5240	11.4946	5.8835
Right Amygdala	[24, -2, -18]	2176	9.6639	7.3186	7.1020

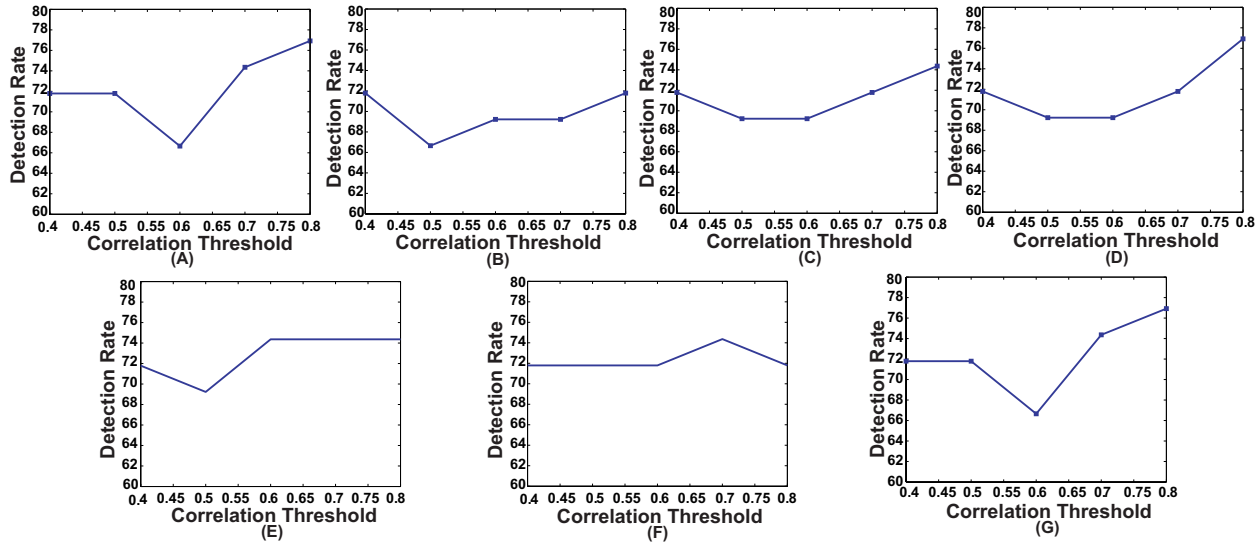


Figure 4.6: The plots show how detection rates for different network features change with correlation threshold. **(A)** Degree map positive correlations, **(B)** degree map negative correlations, **(C)** degree map absolute correlations, **(D)** varying distance degree map positive correlation, **(E)** 3 cycle map positive correlation, **(F)** 4 cycle map positive correlation, **(G)** weight map positive correlation.

4.1.4 Experimental Setup

In this section we describe all the experiments performed to validate our method. The results are reported in the following section.

First, we verified the performance of each of the network features computed on the released set of one of the data centers. We used fMRI data of 83 subjects from the KKI data set. Among the 83 subjects, the first 44 subjects are used for the training and the remaining 39 for the testing. The performances of each of the network features is computed with or without using the useful region mask. The mask is generated on the KKI training set comprising the first 44 subjects of the KKI subset and using the algorithm described in 4.1.3. Each time a random subset of regions is selected, the classification performance is measured by leave-one-out cross validation, i.e. take 43 subjects for the training and test on the one remaining subject; repeat the process 44 times, testing

each of the 44 subjects one at a time and averaging the correct detection counts.

As it is mentioned in Section 4.1.3, we experimentally determined the values of *p* and *th* used in the useful region mask computation algorithm. For this purpose, we varied the probability *p* of including a region in the random subset and the final threshold *th* used on the probability map of the regions to produce different useful region mask. For each pair of values of the *p* and *th*, we compute a different useful region mask which is used to generate different detection rates on the KKI data set. The detection rates are reported in the Figure 4.4. The best performance is achieved when $p = 0.4$ and $th = 0.6$. We used these values to generate the final useful region mask.

To remove the unnecessary connections in a network, we used a correlation threshold to remove all the edges whose correlation values are lower than the threshold. To empirically select the correlation threshold to be used for our experiments, we varied it from 0.4 to 0.8 with an increment of 0.1 in every step. In each step, a different set of networks is computed using different threshold values, network features are extracted and the detection rates of the classification process are computed on the remaining 39 subjects of the KKI released set.

We also perform a thorough experimental validation of our method on the full data set using the positive degree map and positive 3-cycle map features. We trained our classifier with the full released data, which has 776 subjects from 7 different centers, and tested on the holdout sets containing 171 subjects from 6 centers of the ADHD-200 data set. Again, we compared the performance with and without using the useful region mask. We reused the same mask generated using first 44 subjects of KKI. It is worth mentioning that the mask selects 6916 voxels from which features are extracted.

We assume that the regions, which are useful for identifying ADHD conditioned brains, should not vary depending on the feature type used for the detection of the mask. To justify our assumption we generate another useful region mask on the KKI released set using the 3-cycle map features. As in the case of generating useful region map using the degree map features, we use $p = 0.4$ and $th = 0.6$ for the map computation. The mask generated is used to verify the detection

rates of the degree map features on the ADHD-200 holdout sets.

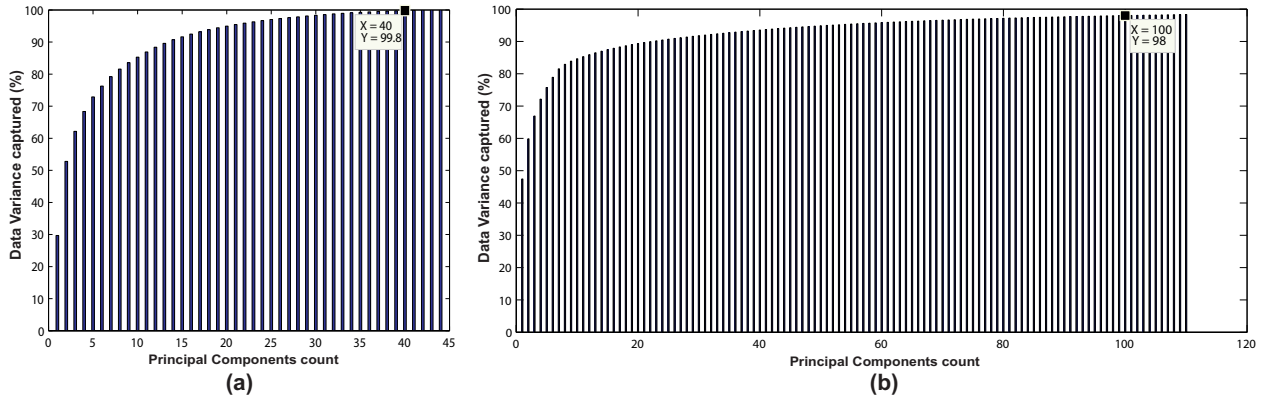


Figure 4.7: The figure shows the plots of principal component count vs percentage of data variance for (a) KKI released set (b) full released data of 776 subjects.

We use the first 40 and first 100 of the principal components for the experiments on the KKI released set and full data set respectively as they cover more than 98% of the data variance. Figure 4.7 shows the plots for the number of principal components vs. the percentage of the total data variance captured. For the KKI released set, the first 40 principal components are able to capture 99.8% of the total data variance while the first 100 principal components of the full released data set are able to capture 98% of the total data variance.

4.2 Results

As it is said in Section 4.1.4, we compute the useful region mask on the first 44 subjects of the KKI released set. Figure 4.5 shows the computed mask on the different slices of the brain. Table 4.1 lists the information of the different clusters found in the useful region mask and the ROIs they are overlapped with. The computed useful region mask is proved to be helpful in terms of improving the classification rates when the features are extracted only from the regions selected by the mask.

Table 4.2: Initial test results shows the performance of all the network features computed on the KKI released set. The Positive, negative and absolute keywords are used to indicate that the positive, negative and absolute correlation values are considered for the network construction. If any keyword is not specifically mentioned, then the positive correlation values are used.

Feature	**Correlation Threshold**	**Performance (%) using useful region mask**	**Performance (%) without useful region mask**
Degree Map positive	0.80	76.92	69.23
Degree Map negative	0.80	71.79	69.23
Degree Map absolute	0.80	74.36	71.79
Varying Distance Degree Map	0.80	76.92	69.23
3-cycle-map	0.80	74.36	71.79
4-cycle-map	0.70	74.36	69.23
Weight Map positive	0.80	76.92	69.23
BOW time series histogram	-	69.23	66.67
BOW Degree Map histogram	0.80	69.23	66.67
BOW time series and Degree Map histogram	0.80	69.23	66.67

We computed the detection rates while different correlation threshold values are used to construct the networks. This helped us to find out the relation between the detection rates and the correlation threshold values. The plots of correlation threshold vs. detection rate for different network features are shown in Figure 4.6. Note that the detection rates for each feature type are measured for the positive, negative and absolute correlation values. However, the features computed from the positive correlation values have always outperformed the other two cases. Hence, we have not reported the results for the other two cases. Since, for all the network features, other than the 4-cycle map, the best performance is consistently achieved when correlation threshold is 0.80, we choose to use this value for all the experiments on the full data set.

Table 4.2 summarizes the best performances obtained for each of the network feature types and the corresponding correlation threshold values. The performance in the table signifies the percentage of total number of correct detection (control and ADHD) among total number of test subjects. Note that for all the features, the performance without using useful regions mask is lower compared to when we use the mask. This demonstrates the importance of the voxel selection step through the generated mask.

Table 4.3: Shows the detection rates (Dt. Rt.), specificities (Spc.) and sensitivities (Sens.) of the classification experiments on the ADHD-200 holdout sets. Comparison of the performances are shown when useful region mask is used and not used for the degree map and 3-cycle map features.

	Deg. Map (mask)			Deg. Map (no mask)			3-cycle Map (mask)			3-cycle Map (no mask)		
	Dt. Rt.%	**Spc.**	**Sens.**	**Dt. Rt.**%	**Spc.**	**Sens.**	**Dt. Rt.**%	**Spc.**	**Sens.**	**Dt. Rt.**%	**Spc.**	**Sens.**
KKI	72.72	1	0	72.72	1	0	72.72	1	0	72.72	1	0
Neuro Image	68	.7857	.5454	64	.7143	.5454	72	.7857	.6364	68	.8572	.4545
NYU	70.73	.9167	.6207	65.85	.7500	.6207	70.73	.8333	.6552	63.41	.8333	.5517
OHSU	70.59	.7778	.4286	64.70	.7037	.4286	73.52	.8148	.4286	70.59	.7407	.5714
Peking	64.71	.8889	.3750	60.78	.8889	.2917	62.74	.9259	.2917	56.86	.9630	.1250
Pittsburgh	77.78	1	.5000	66.67	.8000	.5000	77.78	1	.5000	66.67	1	.2500
Overall	69.05	.8602	.4872	64.32	.7957	.4615	69.59	.8710	.4872	64.33	.8710	.3718

Table 4.4: Shows the detection rates (Dt. Rt.), specificities (Spec.) and sensitivities (Sens.) of the classification experiments on the ADHD-200 holdout sets. A PCA-SVM classifier with a quadratic kernel is used to generate the results. Useful region mask is used to extract the features from the selected voxels.

	Deg. Map			3-cycle Map		
	Dt. Rt. %	**Spec.**	**Sens.**	**Dt. Rt.** %	**Spec.**	**Sens.**
KKI	72.73	1	0	81.82	1	0.3333
Neuro Image	80	0.7143	0.9091	76	0.8571	0.6364
NYU	58.54	0.25	0.7241	58.54	0.25	0.7241
OHSU	73.53	0.7407	0.7143	79.41	0.8889	0.4286
Peking	64.71	0.8148	0.4583	64.71	0.8148	0.4583
Pittsburgh	88.89	1	0.75	77.78	0.8	0.75
Overall	69.01	0.7312	0.641	69.59	0.7849	0.5897

We compare the performance of our method with the BoW method introduced in the last chapter. Following the experimental setup of the BoW method each subject is represented by 75 and 100 dimensional histograms when the raw time series and degree map features are used respectively. A third kind of experiment is performed by representing each of the subjects as a concatenation of the two types of histograms resulting in a 175 bin histogram. These results are also included in Table 4.2.

The results on the full data set are reported in Table 4.3. The table includes the detection rate, specificity and sensitivity for each of the holdout sets along with the average measures for all the holdout sets. Since the subject labels of the Brown University holdout set have not yet

been released, we cannot compute the performance measures on that. To compare the result, we performed the same experiments using the PCA-SVM classifier with a quadratic kernel. The results are reported in Table 4.4. As it can be seen, the performance is very similar to the PCA-LDA classifier.

Finally, we compute a useful region mask using the 3-cycle features and use it to perform classification on the holdout sets. Figure 4.8 shows the useful region mask generated using the 3-cycle features and computed on the 44 subjects of the KKI released set. The mask is plotted on the different slices of the brain image of a sample subject. The experiment results on the full data set are reported in Table 4.5 where features are extracted from the regions selected in the new mask. The detection rates we got using the masks generated by the 3-cycle and positive degree map features are almost same. This matching results supports our initial assumption that the computed useful regions mask is invariant to the feature used to compute it.

Table 4.5: Shows the detection rates of the degree features on the ADHD-200 holdout sets while a useful region mask is used to select the features. The useful region mask is generated using the 3-cycle features computed on the first 44 subjects of the KKI released set.

	Detection Rate (%)	**Specificity**	**Sensitivity**
KKI	72.72	1	0
Neuro Image	72	.5714	.9091
NYU	70.73	.8333	.6552
OHSU	73.52	.8889	.1429
Peking	60.78	.9630	.1667
Pittsburgh	77.78	1	.5000
Overall	69.01	.8710	.4675

4.3 Discussion

We modeled the brain as a functional network which is expected to represent the interaction of the different active regions of the brain. We assumed that the ADHD is a problem caused due to the partial failure of the brain's communication network and the affected subjects can be

distinguished from the control subjects using the topological differences of their respective functional networks. To verify the idea, we extracted different network features to train a PCA-LDA based automatic classifier. Figure 4.9 shows that the average degree map, computed for the ADHD and control subjects of the KKI released set, is able to capture the differences of connectivity in the Cingulate Gyrus and the Paracingulate Gyrus regions of the brain. We also proposed that the features from the whole brain are not required for the classification, but some key areas hold useful information. Our results shows that the inclusion of the features from the whole brain can negatively impact the classification accuracy. This resulted in a novel algorithm to compute the useful region mask which helped to improve the classification performance.

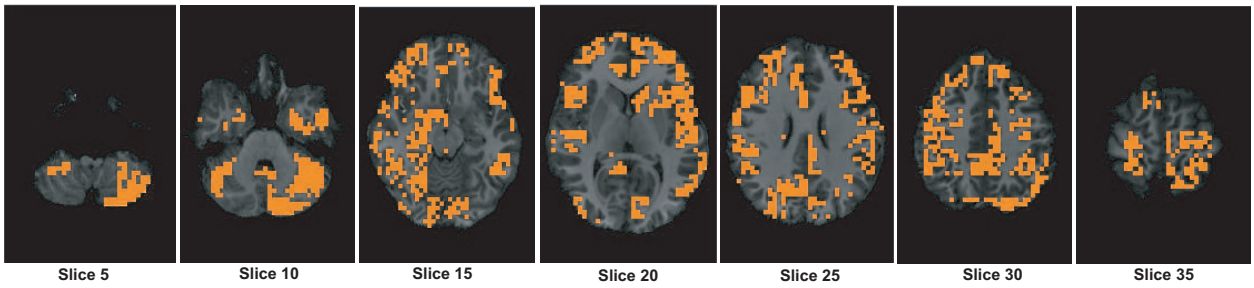


Figure 4.8: The figure shows different slices to demonstrate the computed useful region mask using the 3-cycle map features. The masked regions are highlighted in orange color and overlaid on different slices of the structural image of a sample subject.

For our analysis, we only selected node based features to capture the local structures for the network. The features we used are easy to compute, simple in concept, and expected to capture different topological characteristics of the functional network. As we hypothesize that the cause of ADHD is the presence of abnormalities in the brain functional connections, we selected the features such a way that they capture different connectivity pattern of the network. The degree map and the weight map can capture how densely the nodes of the network is connected. These give us measures of how synchronous different brain regions are. The varying distance degree

map, on the other hand, can also reveal how the synchronous regions are distributed over the brain. While the degree map only captures the pairwise interactions of the voxels, it ignores higher-order interactions, such as among three voxels simultaneously. We know from the brain anatomy that there are such multiply connected brain regions. Hence, cycle maps offer a different perspective from which a given network may be viewed. The utility of using network motifs such as the cycles is described in [51].

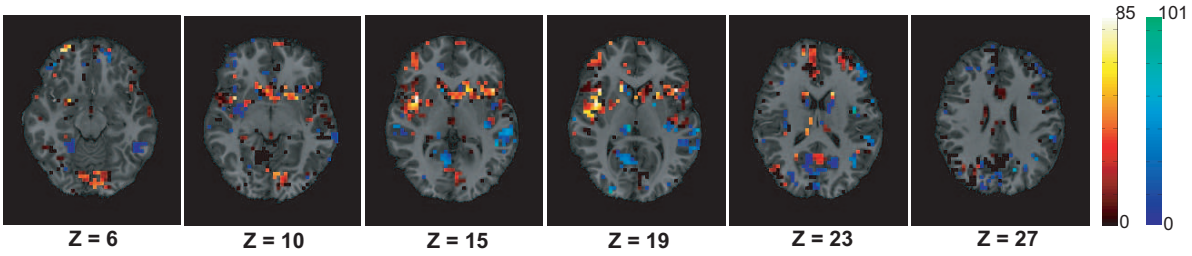


Figure 4.9: The figure shows the average differences of the degrees between the control and ADHD groups in the voxels belonging to the useful region mask. The average differences are calculated for the 83 subjects of the KKI released set. The dark red to white color map is used to represent the regions with higher degrees in control subjects and blue to green color map is used to show the opposite. The control group shows higher connectivity in the Cingulate Gyrus ($Z = 10, 15$) and Paracingulate Gyrus regions ($Z = 19, 23$).

The useful region mask selection algorithm has three parameters such as the probability of inclusion (p) of a region in an iteration of the algorithm, the threshold th to prune the low occurring regions, and the number of iterations the algorithm should run. The first two parameters are decided empirically (4.4). Let us assume that in each iteration of the algorithm $\vartheta.p = \nu$ numbers of regions are selected, where ϑ is the total number of brain regions considered. Then the algorithm can select $\binom{\vartheta}{\nu}$ number of possible distinct subsets of regions. The value of $\binom{\vartheta}{\nu}$ is the upper limit of the number of iteration parameter. Unless ϑ is a very small the number $\binom{\vartheta}{\nu}$ is very large which is impractical for the algorithm. We used the number of iteration as 500 in our algorithm as we observed not much changes in detection accuracies after the number of iteration crosses few hundreds. Further

thorough analysis can be performed to decide the best value for this parameter.

Figure 4.5 and Table 4.1 present the ROIs found through our adaptive labeling technique described in Section 4.1.3. These ROIs were used in the classification including regions such as the cingulate and precuneus which is consistent with the findings of Castellanos et al. [13]. The cingulate and precuneus regions are known to be part of the default-mode network [25]. Many regions in the Table 4.1 have also been identified by Assaf et al. [3], such as the precuneus, temporal pole, superior temporal gyrus, and pre-central gyrus. Regions in Table 4.1 that are consistent with those reported by Uddin et al. [79] include the inferior temporal gyrus and lingual gyrus. Interestingly, Table 4.1 identifies the right amygdala, which did not show up in the analysis of Castellanos et al. [13] or Assaf et al. [3] or Uddin et al. [79]. The limbic system is known to play a role in ADHD, and a study by Plessen et al. [58] reported disrupted connectivity between the amygdala and OFC in the children with ADHD. Hence the value of our technique is that it provides an independent and automatic source of hypotheses about the brain regions that are implicated in the diagnosis and classification of ADHD. In this sense, our technique for ROI identification can be considered to be a model-free method. Furthermore, our classifier is agnostic to any particular theory of ADHD, and works strictly on a machine-learning approach to separate the ADHD patients from the controls by utilizing labeled data. Therefore, the technique described in this chapter is applicable to other types of brain disorders where one can create labeled data for the accompanying brain scans.

The plots in Figure 4.6 show that for all the network features, high performance values are achieved when correlation threshold 0.80 is used for the network construction. In four out of seven cases the performances are the highest, in other two cases they are one of the highest and in one case it is slightly lower than the highest. The results are not surprising since they indicate that the differences of connection structures of the highly correlated voxels matter the most for the ADHD classification problem.

Considering the results in Table 4.3, we observe that the 3-cycle features performed slightly better than the the degree features. To the best of our knowledge, this is the first time that the utility

of the cycle-related features has been demonstrated in the fMRI imaging literature. The study in [49] showed that the cycle-related features are useful in discriminating biological networks from man-made networks, but did not investigate various types of fMRI-derived networks.

We found that the construction of the cycle-related features is more computationally intensive than the degree map, and the computation cost increases exponentially with the cycle length. The use of GPUs can reduce the cost of computation, as earlier studies with fMRI images have shown [60]. If standardized libraries for the cycle computation become available on GPU platforms, it will promote the use of such features in fMRI research. The use of the degree map provides a good compromise between the classification performance and computational cost. It is easy to compute, and provides classification performances that are only marginally worse than that of the 3-cycle maps in most cases.

In summary, the results clearly suggest that the use of the fMRI data for the analysis of ADHD can be helpful in terms of identifying the root cause of the problem as well as developing a system for the automatic detection of affected subjects. One of the shortcomings of this approach is that the features are computed on the nodes of the networks which can only capture the local, structures ignoring the global topology of the networks. Second, each selected voxel is represented by a node of the network which increase the size of the network as well as the computation cost. We address these problems in the next chapter.

CHAPTER 5: ATTRIBUTED GRAPH DISTANCE MEASURE FOR THE ADHD DETECTION

In the last two chapters we represented each voxel of a brain volume as a node of the brain functional network. While this representation gives us the mean to model the brain dynamics, the cost of the network computation becomes too high. This is because, in the fMRI data, the brain volume of each subject is represented by approximately 28,000 voxels which makes the size of the correlation matrix very big ($28,000 \times 28,000$). In this chapter, we propose an efficient representation of the network such that the maximum information is preserved with minimum redundancy. To achieve this goal, first we select only the highly active voxels for the construction of the network. We hypothesize that these highly active voxels contain the most useful information for the classification of the ADHD subjects. Next, we notice that the voxels in the spatial proximities contain redundant information as their activity patterns in the fMRI intensity time series are very similar. Therefore, we group the selected highly active voxels, belonging to the different functionally homogeneous regions, into different clusters. The functionally homogeneous regions are identified using the CC200 map [22], which segments the whole brain into 190 spatially contiguous and functionally correlated regions. Each cluster of voxels is then represented as a node of the network. These steps help us to significantly reduce the network computation cost.

The second main difference from the last two chapters is that we approach the classification problem in a different way. Instead of computing the network features, we map the networks onto a low dimensional spatial configuration and perform classification in the projected space. While the network features are computed for each node and can capture only the local network structure, the projection of the networks helped us to utilize the global topology in our classification framework. The Multi-Dimensional Scaling (MDS) technique is used for the projection of the networks using the inter-network distance measures. Our method shows promising results as we

achieve impressive classification accuracies on the released(70.49%) and holdout(73.55%) sets. Our results reveal that the detection rates are higher when classification is performed separately on the male and female groups of subjects.

5.1 Multidimensional Scaling

In this section, we provide a general overview of the MDS for the sake of the completeness of the chapter. The MDS is a set of data analysis techniques that enables one to understand the key dimensions of the objects under investigation. The method and the term were first introduced by Torgerson [77]. Given a set of objects and the proximities of each possible pairs of objects, MDS techniques can find a spatial configuration of the objects based on their proximities. Here, proximities suggest the overall dissimilarities or similarities of the objects being considered. Hence, MDS can be understood as a method to project the objects from a space of unknown dimensions to a space of specified dimensions in such a way that the original proximities of the objects are preserved as closely as possible. To state it formally, given *N* numbers of objects and a dissimilarity (or similarity) matrix D_{NxN}, MDS projects the objects on a space of given dimensions in such a way that $D - D_p$ is minimized. D_p is the distance matrix in the projected space.

Depending on how a dissimilarity (or similarity) matrix is computed, MDS can be subdivided into direct and indirect methods. While for the direct methods numerical dissimilarity value of each pair of objects can be directly computed, for the indirect methods dissimilarity values need to be derived from other values like confusion data. Again, MDS can be divided into classical and nonmetric classes depending on how the problem is solved. While the classical methods assume that the dissimilarity matrix contains exact distances of the objects, the nonmetric methods consider only the ordinal information of the object proximities. For more details on the MDS, we refer the interested readers to [45]. For our experiments, we used a direct classical MDS technique.

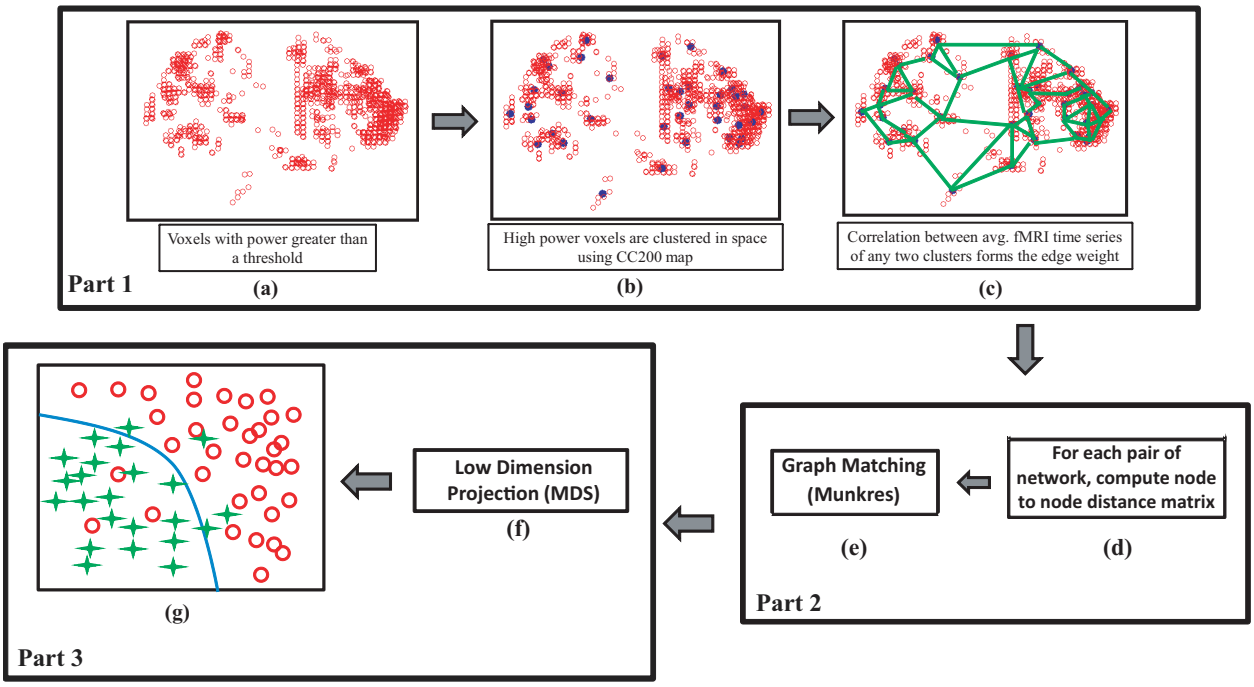


Figure 5.1: Flowchart of our proposed method. (a) High power voxels are selected. (b) High power voxels belong to each region of interest of the CC200 map are clustered together and represented by their cluster centers. Each of the clusters represents a node of the network. (c) Edges of the network are formed based on the correlations of average fMRI signals of the clusters. (d)-(e) Inter-network distances are computed in two steps. First, for a pair of networks a node to node distance matrix is computed. Next, each node of the network with a fewer node count is assigned to a node of the second network using Munkres algorithm such that the total matching distance is minimized. (f) The MDS is used to form a spatial configuration of the subjects on a low dimensional space based on the inter-network distance measures. (g) Classification is performed in the projected space.

5.2 Method

The proposed method can be divided into three main parts: network construction, graph distance computation and ADHD subject classification. The following sections describe the parts in details.

5.2.1 Network Construction

For each subject of the data set, the resting state brain functional connectivity network is computed. The following steps describe the network construction method. The concept is graphically explained in Figure 5.1 $(a) - (c)$.

The first step of the network construction method is the selection of the candidate voxels which constitute the network. We observe that all the brain voxels do not contain valuable information and including irrelevant voxels can degrade the classification performance. We hypothesize that the voxels with high activity levels contain the most useful information for the ADHD classification problem and therefore selected to construct the functional connectivity network. We substantiate our hypothesis by examining the experimental data in Section 5.3, where we show that the inclusion of all the brain voxels in the construction of the network degrades the classification performance. We consider the power of the fMRI time series of a voxel as the measure of its activity level. Higher the power of a voxel, higher is its activity level. For a discrete time series $T = \{t_1, t_2, ..., t_n\}$, the power can be computed as,

$$P(T) = \frac{1}{n} \sum_{i=1}^{n} {t_i}^2 \tag{5.1}$$

We then normalize the power values of all the voxels between $[0, 1]$. The voxels are then ranked based on their power values. Finally, we selected the voxels ranked with 98 percentile or more for the network construction.

In the second step of the network construction method we used an efficient way to represent the nodes of the network such that the node count is reduced without sacrificing any relevant information. In the last two chapters we represent each selected brain voxel as a node of the network. There are two problems in doing this. First, it makes the size of the network very large, which is inefficient for further computational analysis. Second, we observed that the voxels in the close spatial proximities have very similar functional activity patterns. Hence, including all

these voxels for the network construction makes the network full of redundant information. For these reasons we use an ROI map, (CC200) proposed by Craddock et al. [22], to cluster the highly active voxels to form the nodes of the network. The map is generated by parcellating the whole brain resting state fMRI data into 190 spatially contiguous regions of homogeneous functional connectivity (FC). The selected highly active voxels belonging to each of the ROIs form the cluster. The issue concerning the best resolution of ROIs which contain the maximum information with minimum redundancy for the functional study of the brain is not addressed in this work.

In the third step, we construct the edges of the network and compute the weights of the edges. We represent each of the nodes by the average fMRI time series of all the voxels comprising the node. Then, a correlation matrix is computed which contains the correlation values of the fMRI time series of all possible pairs of the nodes in the network. For two nodes m and n with fMRI time series $m_T = \{m_1, m_2, ..., m_t\}$ and $n_T = \{n_1, n_2, ..., n_t\}$ respectively, the correlation value is computed as:

$$corr(m_T, n_T) = \frac{(t\sum_{i=1}^{t} m_i n_i) - (\sum_{i=1}^{t} m_i)(\sum_{i=1}^{t} n_i)}{\sqrt{[t\sum_{i=1}^{t} m_i^2 - (\sum_{i=1}^{t} m_i)^2][t\sum_{i=1}^{T} n_i^2 - (\sum_{i=1}^{t} n_i)^2]}}, \tag{5.2}$$

Note that the correlation values have range $[-1, 1]$. We empirically verified that the networks constructed with only the positive correlation values provide better classification accuracies compared to the networks constructed with only the negative correlation values or absolute correlation values. Hence, the experimental results reported on the networks constructed with positive correlation values only. Also, we use a correlation threshold *corrTh* to remove all the edges from the network which have correlation values less than the threshold.

In the final step, we represent the network as an attributed graph where each node of the network is represented by a set of attributes. We call it the signature of a node. Given a node *n*, its

signature is defined as:

$$Signature(n) = \langle deg(n), deg(ngh(n)), pow(n), pow(ngh(n)), coord(n) \rangle, \tag{5.3}$$

where the functions, $deg(.)$, $ngh(.)$, $pow(.)$, return the sum of weights of all the connected edges, the nodes connected by edges and the power respectively corresponding to the input nodes in the functions. $coord(.)$ is the mean physical coordinates of all the voxels comprising the node.

5.2.2 Graph Distance

Once the functional networks are constructed for all of the subjects in the data set, we compute the distances of all possible pairs of networks as shown in Figure 5.1 (d). For a pair of networks, the distance computation is a two step process. In the first step we compute the distances of all the node pairs formed by selecting one node from each of the networks. Given two networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ and two nodes $v_1 \in V_1$ and $v_2 \in V_2$, the distance between v_1 and v_2 is computed as the difference of their signatures:

$$dist(v_1, v_2) = \mathbf{W} \cdot [d_1, d_2, d_3, d_4, d_5]^T, \tag{5.4}$$

where $\mathbf{W} = [0.2, 0.1, 0.2, 0.1, 0.4]$ is the weight vector and d_1, d_2, d_3, d_4, d_5 are the differences of the node degrees, the neighboring node degrees, the node powers, the neighboring node powers, and the physical locations of v_1 and v_2. All the difference values are normalized between $[0, 1]$ to enable proper comparison. The values of d_1 and d_3 are simply calculated by computing the degree and power differences of v_1 and v_2 and dividing them respectively by the maximum degree and power encountered for any of the nodes in the training set. To compute d_2, first we sort the neighbor degrees in descending orders. The node with less number of neighbor nodes is zero

padded at the end to make the size of the degree arrays the same. Finally, we sum up the absolute differences of the array elements and divide it by $(maximumdegree * size(degreearray))$. d_4 is computed in a similar fashion while power values are used instead of degrees. d_5 is calculated as follows:

$$d_5 = \frac{1}{1 + 300e^{-(|c_1 - c_2|)/4}}, \tag{5.5}$$

where c_1 and c_2 are the physical coordinates of v_1 and v_2 respectively. This is a sigmoid curve which restricts the value of d_5 in the range of $[0, 1]$. The parameters of the equation are heuristically determined in such a manner that the value of d_5 is close to zero when $|c_1 - c_2| = 0$, low for the nodes in a spatial locality and steeply increasing for the nodes which are further apart. The components of the weight vector W are also determined heuristically considering the following criteria. First, we want to make sure that the nodes which are physically far apart should not match and hence set the highest weight corresponding to the nodes' physical distance. Next, we want to give the same importance to the degree and power distances of the nodes. Hence, the weights corresponding to the node degree and power distances are assigned the same value. Similar condition is applied for the weights of the neighboring node degree and power distances. Finally, we assume that the importance of the node feature distances should be higher than the importance of the neighboring nodes' feature distances. Hence, weight for the neighboring nodes' distances are lower than the node distances. In general the distance of a pair of graphs should be calculated in such a way that the nodes from the nearby regions with similar degrees and powers and with similar neighboring nodes' degree and power distributions should match.

In the next step, we use the Munkres assignment algorithm [52] to assign all the nodes of one network to the nodes of the second network in such a way that the total assignment cost is minimized. This assignment cost is considered as the distance of the network pair. Note that the numbers of nodes for all the networks are not the same. This is because when we select the high

power voxels there are some ROIs from which no voxels are selected. But this does not cause any problem in our case as the Munkres algorithm can find the assignment cost even when the node counts of the two networks are not the same.

5.2.3 Classification

When the subjects are modeled as the networks, they cannot be directly used for the classification but first need to be mapped onto a feature space. A common way to deal with this is to compute different network features which can be used for the classification [6], [82]. We took a different approach to solve this problem. As shown in Figure 5.1 $(e) - (f)$, we use the direct classical MDS technique to project the networks in a space with specified dimensions. The MDS technique takes the network distance matrix, as discussed in the last Section 5.2.2, as input and produces a spatial configuration of the networks in the projected space. The number of dimensions of the projected space can also be specified in the MDS method. We achieve the best classification accuracy when we use the number of dimensions as 2. All the results of our proposed method are generated on the 2 dimensional projected space.

The classification is performed in the projected space using the SVM [20] with a polynomial kernel. We choose to use the SVM classifiers for the following reasons. First, SVM can classify the data points from two classes even when they are not easily separable in the original feature space. SVM use a technique called kernel trick to project the data points into a hyperspace where the separation is easy. Second, SVM regresses the feature space without over fitting on the data by allowing miss-classification with a penalty. Our experimental results also show that the classifiers perform better when trained separately on the male and female subjects. This indicates that there may be considerable differences in the functional connectivity networks of the male and female subject groups. Our result is consistent to the work of Balint et al. [5] who showed that the male and female ADHD subjects have differences in the brain functions.

5.2.4 Experimental Setup

The setups for all the different experiments performed are described in this section. Experiment results are listed in section 5.3.

For all our experiments we used MATLAB (version R2008b) implementations of MDS and SVM. For MDS, we used the function named *mdsscale* with the *criterion metricstress* and $MaxIter = 100,000$. For SVM, we used the functions named *svmtrain* (with polynomial kernel) and *svmpredict* to train the classifiers and test the detection accuracies respectively.

In Section 2.3.1 it is stated that the different data centers used different experimental protocols for the data capturing. Also, in Table 2.2 it is shown that the scanners and scan parameters also vary a lot across the data centers. These motivate us to train our classifiers separately on the subjects corresponding to the different data centers to avoid possible data variance due to the change of the experimental protocols. All the experiments are performed on the subjects of released and holdout sets of 4 data centers; KKI, NeuroIMAGE, OHSU and Peking.

For all the released and holdout sets of all the data centers, three different sets of experiments are performed. While first set of experiments is performed on all the subjects, second and third sets of experiments are performed on the male and female groups separately. Hence, in total $(4releasedsets + 4holdoutsets) * 3 = 24$ different sets of experiments are performed. For the released sets detection accuracies are achieved by the leave one out cross validation method. For the holdout sets the classifiers are trained on the subjects of the corresponding released sets and the validations are performed on the holdout sets. For each of these sets of experiments, we construct the networks by varying the *corrTh* from 0.30 to 0.90 with a step size of 0.10. The *corrTh* is explained in Section 5.2.1 while describing the network construction steps.

We compared the performances of our method with a SVM graph kernel based approach [9] which can be a natural choice to try on our problem. Graph kernel is a function to compute the inter graph distance for any given pair of graphs. As we know, SVM can use the kernel trick to

project the input data into the kernel space and perform the classification in the projected space for the better separations of the input classes. Similarly, a graph kernel can be used to project a set of networks from an unknown space to a network distance matrix which contains the inter-network distances for all possible network pairs. Hence, the networks themselves become the dimensions of the projected space and each coordinate signifies the distance from the network representing the particular dimension. We used our graph distance computation approach as the graph kernel and the computed inter-network distance matrix as the input of the SVM. The feature vector for any given network becomes the distances of the network from all the networks in the training set.

For the purpose of comparing our results, we perform the same classification experiments using some standard graph features computed on the brain functional connectivity networks. The features are computed using the Brain Connectivity Toolbox (BCT) [62], which contains a large selection of complex network measures commonly used for characterizing structural and functional brain connectivity data sets. The features we used are the degree, the topological overlap, the clustering coefficient, the local efficiency and the rich club coefficient. The following are the brief descriptions of the network features used:

- Degree of a node is the number of edges incident on it.
- The m_{th} step generalized topological overlap measure quantifies the extent to which a pair of nodes have similar m_{th} step neighbors. Where m_{th} step neighbors are the nodes that are reachable by a path of at most length m. We got best results for $m = 5$.
- The clustering coefficient is the fraction of the triangles around a node. In other words, it is the ratio of the neighboring nodes count which are connected to each other to the total number of neighboring nodes of the node.
- The local efficiency is the global efficiency computed on the node neighborhood. Where the global efficiency is the average of the inverse shortest path lengths in the network.

- The rich club coefficient at level k is the fraction of edges that connect the nodes of degree k or higher out of the maximum number of edges that such nodes might share. We compute the coefficients for all the k values where $0 \leq k \leq K$. Here, k is an integer and K is the maximum degree found for any node of the training data.

Since each of the network features returns a feature vector whose size depends on the node count of the network, we had to make the node counts same for all the subjects to make the feature sizes same. For this reason we construct the networks in a little different way. Instead of using one power threshold value for selecting the highly active voxels of the whole brain, we use separate power thresholds for each of the ROIs of the CC200 map. For each of the ROIs, we select the voxels ranked 98 percentile or higher based on their power values. The rest of the network construction process is the same as before. The experiments are also set up in the similar fashion as described for our proposed method.

To better understand the physical interpretations of each of the dimensions of the MDS projected space, we performed some analysis. First we compute some global feature values for each of the networks of the KKI released set. A brief description of the computed features is as follows:

- **Density:** it is the fraction of the present connections to all possible connections of the network.
- **Global efficiency:** it is the average of inverse shortest path lengths of the network.
- **Rich club coefficient:** it is as described before in Section 5.3. The correlation values reported with x coordinates of the male and female groups are achieved when $k = 11$ and $k = 1$ respectively.
- **High power node fraction:** it is the fraction of the nodes with power greater than a threshold to the total number of nodes of the network. The correlation value reported with x

coordinates of female group is achieved when $powTH = 0.85$.

For each of the computed global features, two separate feature vectors for the male and female group of subjects are formed. Please note here each feature vector represents a group of subjects (e.g. the male and female groups) but not the individual subjects. Then the correlations of the feature vectors are computed with the x and y coordinates of the networks when projected as points on the 2 dimensional space.

To show the importance of the high power voxel selection step we perform a set of experiments using our method but without the voxel selection step. Finally, we experimentally validate the effectiveness of the node attribute set used in our method. For this purpose, we compute the inter-graph distances using different subsets of the attributes used for the original framework. For each of the subsets, the inter-graph distances are computed separately followed by the projection of the subjects to a low dimensional space using the MDS and classification using the SVM. It is not possible for us to compute the results for all possible subsets as there can be 31 different subsets for 5 attributes. Instead we start with one attribute and keep on adding attributes in the subsets. The results show that the classification accuracies steadily increase as we keep adding attributes in the subset. Finally, we performed the experiments using all combinations of 4 attributes to show that even missing one of the attributes from our attribute set decreases the classification accuracy.

5.3 Results

The detection rates of our method, when classification is performed separately on the male and female subjects, are plotted in Figure 5.2. The plots show how the detection rates vary for the different data centers and with respect to different $corrTh$ values. In Table 5.1 we reported the best detection rates of our method along with the specificity and sensitivity values for all the released and holdout sets. The $corrTh$ values corresponding to the best detection rates on the released sets are selected and used to get the detection rates for the holdout sets. One interesting fact is

that in most of the cases we get better classification accuracies when experiments are performed on the male and female subjects separately. We achieve an average detection rate of 64.48% on the released sets and 62.81% on the holdout sets when the classification is performed on all the subjects and 70.49% on the released sets and 73.55% on the holdout sets when the classification is performed separately on the male and female subjects.

Table 5.1: Summary of the results: table shows the best detection rates achieved (along with their specificities and sensitivities) on all the released and holdout sets using the proposed approach. The $corrTh$ values are selected from the released sets where we achieve best detection rates. The rates on the holdout sets for the corresponding $corrTh$ values are reported. The values under the heading 'Male Female Separate' are computed by averaging the accuracies on the male and female groups.

All Subjects							
Data Centers	**Released sets**			**Holdout sets**			$corrTh$
	Detection Rate $\%$	**Specificity**	**Sensitivity**	**Detection Rate** $\%$	**Specificity**	**Sensitivity**	
KKI	75.64	1	0.952	54.55	0.6250	0.3333	0.8
NeuroIMAGE	64.10	0.6818	0.5882	48.00	0.6429	0.2727	0.5
OHSU	60.61	0.6579	0.5353	82.35	0.8929	0.5000	0.9
Peking	61.20	0.8661	0.2113	58.82	0.9259	0.2083	0.6
Average	64.48	0.8471	0.3066	62.81	0.8312	0.2727	
Male Female Separate							
KKI	76.92	0.9048	0.3684	54.55	0.6250	0.3333	0.5
NeuroIMAGE	76.92	0.8182	0.7059	100	1	1	0.5
OHSU	68.18	0.7895	0.5357	61.76	0.6071	0.6667	0.3
Peking	67.21	0.8393	0.4085	72.55	0.7407	0.7083	0.3
Average	70.49	0.8453	0.4672	73.55	0.7273	0.7500	

Table 5.2 summarized the results of the graph kernel based approach described before. As it can be seen, the classification accuracies are much lower compared to our method. The possible reason for the low classification accuracy can be the following. In the graph kernel space, the projected inter networks distances may not be the same as the original distances. This is easy to understand with an example. Let us assume three networks A, B, C with inter-network distances computed as $A - B = 4$, $B - C = 2$ and $C - A = 4$. Then their representations in the kernel space are $A = \{0, 4, 4\}$, $B = \{4, 0, 2\}$, and $C = \{4, 2, 0\}$. Hence, the Euclidian distance between A and B in the kernel space becomes 6, B and C becomes around 2.83, and C and A becomes 6 which are different from the original distances. MDS on the other hand tries to preserve the

original distances in the projected space.

Table 5.2: Summary of the results: table shows the best detection rates achieved (along with their specificities and sensitivities) on all holdout sets using the SVM graph kernel method. The $corrTh$ values are selected from the released sets where we achieve best detection rates using our proposed approach. The values under the heading 'Male Female Separate' are computed by averaging the accuracies on the male and female groups.

All Subjects				
Data Centers	**Holdout sets**			$corrTh$
	Detection Rate %	**Specificity**	**Sensitivity**	
KKI	63.64	0.625	0.6667	0.8
NeuroIMAGE	32	0.1429	0.5455	0.5
OHSU	70.59	0.8571	0	0.9
Peking	54.90	0.9259	0.125	0.6
Average	55.37	0.7162	0.2444	
Male Female Separate				
KKI	27.27	0.25	0.3333	0.5
NeuroIMAGE	96	0.9286	1	0.5
OHSU	61.76	0.7143	0.1667	0.3
Peking	58.82	0.8889	0.25	0.3
Average	64.46	0.7662	0.4318	

The detection rates of the classification experiments performed using the standard network features are shown in Figure 5.3 along with the results of our method. The results are reported separately for each of the data canters as well as the average detection rates. As it can be seen, in almost all of the cases our method performs better than the network features. Also, on average, none of the features performs better than our method when used separately on the male and female subjects. This justifies the need of a specialized method for the analysis of the brain functional problems like ADHD. Please note that we ignored the classification results if any of the specificity or sensitivity is zero. This implies that either all the subjects are classified as ADHD or control. This is why for some of the network features the detection accuracies are zeros in Figure 5.3. Figure 5.3 also shows the best detection rates of our method when no power threshold is applied for the voxel selection during the network construction step. The lower detection accuracies of these experiments compared to our results demonstrate the importance of the voxel selection step.

Figure 5.4 reports the results when different subsets of node attributes are used for the cal-

culation of the inter-graph distances. For each of the subsets, the average classification accuracies on all the data centers are plotted in the Figure. The reported results are achieved when the classifications are performed separately on the male and female subject groups. As it can be seen the best detection rates are achieved when we use all the attributes in the set. This demonstrates the importance of using all the attributes for the calculation of the inter-graph distance.

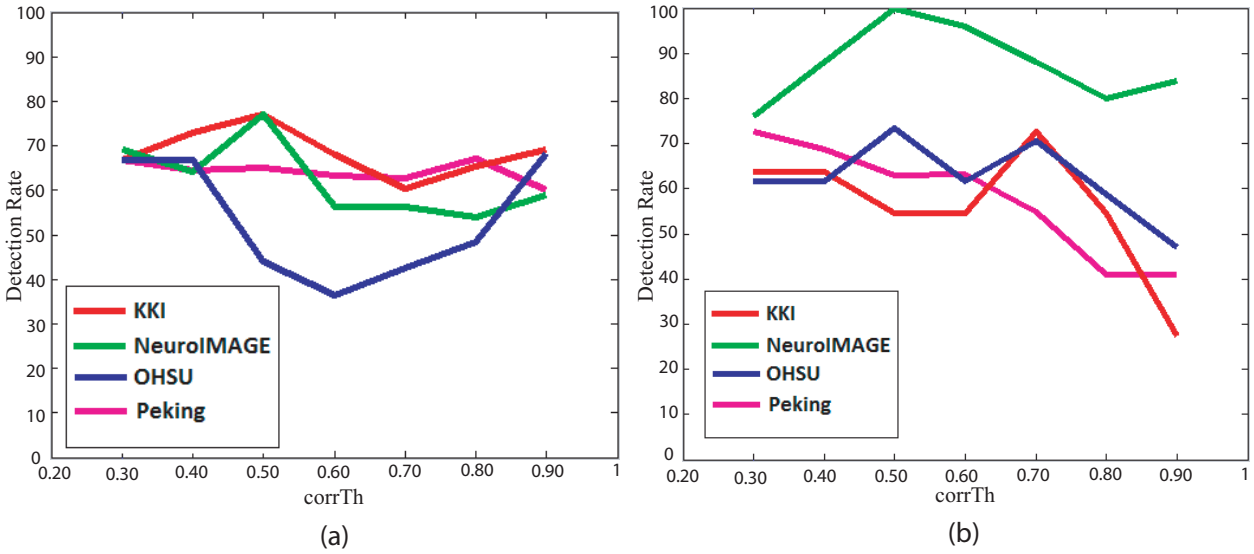


Figure 5.2: Figure plots *corrTh* vs detection rates of our method on the (a) released sets and (b) holdout sets.

5.4 Discussion

In this work we proposed a novel framework for the automatic detection of the ADHD subjects using rs-fMRI data of the brain. For this purpose we construct the functional connectivity network of the brain where each node of the network is represented by a set of attributes. The first step of the network construction method is the efficient selection of the voxels which relate to the functionally active regions of the brain. These highly active voxels are used for the networks

construction where the voxels activity levels are measured based on the power of their fMRI time series. Often signal to noise ratio of the low active voxel time series is very high. Also, these noisy time series can have considerable correlations with each other which lead to the adding of spurious edges in the network or changing of the edge weights of the network. The intuition behind the selection of the highly active voxels is to reduce this noise which can affect the correlation weights of the network edges. As shown in the plots of Figure 5.3 (a) and (b), the voxel selection process in general helps to improve the classification scores. However, we have not experimentally verified what is the ideal power threshold value for this. Further we used a functional ROI map (CC200) to construct the network nodes by clustering the selected voxels belonging to the same ROIs. The active voxel selection step along with the use of the CC200 map to cluster the voxels helped us to reduce the computational cost of our algorithm by a great amount. Compared to around 28000 voxels per brain volume, the average node count of the constructed networks is around 60.

Next, we model the network as an attributed graph where each node of the networks has its signature. The signatures of the nodes contain information about the local structures of the networks. Next, at the time of inter-graph distance computation step, the Munkres algorithm is used to match these local descriptors in a globally optimized fashion. To discourage the algorithm from matching two nodes which are spatially apart, we use the Euclidian distance of their coordinates as a parameter of the matching cost computation.

The inter-graph distance measures allow us to use the MDS technique to map the networks from an unknown space to a 2 dimensional projected space. Figure 5.5 shows the spacial configuration of the subjects of the KKI released set when they are mapped to the 2-D projected space. As it can be seen, ADHD subjects can be better segmented when the male and female groups are plotted separately compared to when all the subjects are plotted together. This fact is reflected in the experimental validations where we consistently get better results when classifications are performed separately on the male and female groups.

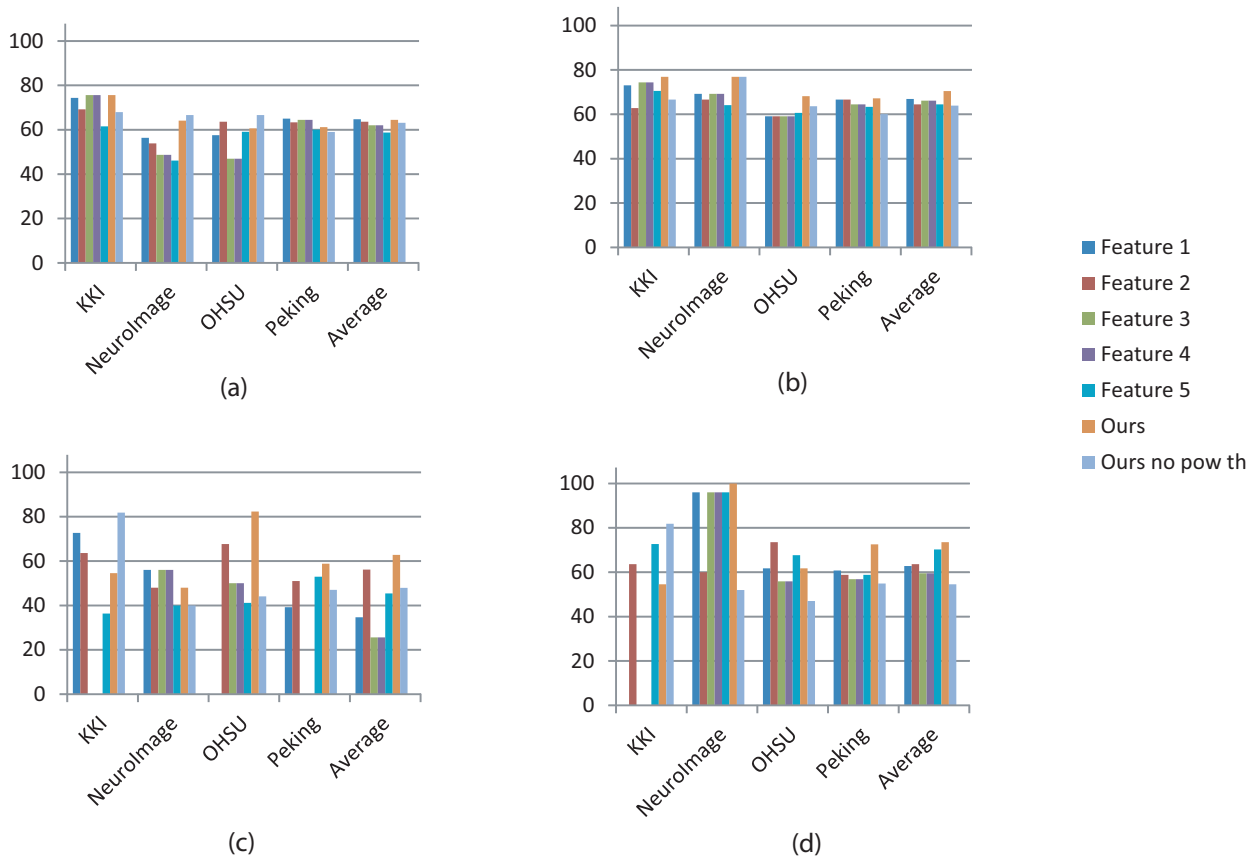


Figure 5.3: Summary of the results: figure plots the best detection rates achieved on all the released and holdout sets using five commonly used network features implemented in the BCT, our method and our method without the high power voxel selection step. Features 1 to 5 are the degree, topological overlap, clustering coefficient, local efficiency and rich club coefficient respectively. (a) and (b) show the results on the released sets when the classification is performed on all the subjects and on the male and female subjects separately. (c) and (d)show the similar results on the holdout sets. The detection rates of (b) and (d) are computed by averaging the detection rates on the male and female groups.

We perform an analysis to understand the physical interpretation of the different dimensions of the MDS projected space. For this purpose we compute the correlations of the different global features of the networks with their coordinates in the projected space. The correlation values are reported in Table 5.3. It can be seen that the x coordinates of the projected spaces of the

male and female groups are highly correlated with the density and rich club coefficient features and moderately correlated with the global efficiency. It should be noted that these three features capture different aspects of network edge structures. The last feature shows some correlation with the y coordinate of the female group.

To justify the importance of a specialized method for analysis of the ADHD, we compared our results with some of the standard brain connectivity measures heavily used for functional analysis of the brain. As shown in Figure 5.3 our method outperforms the standard network features by a large margin. Only the topological overlap feature performs similar to our method on the released data sets.

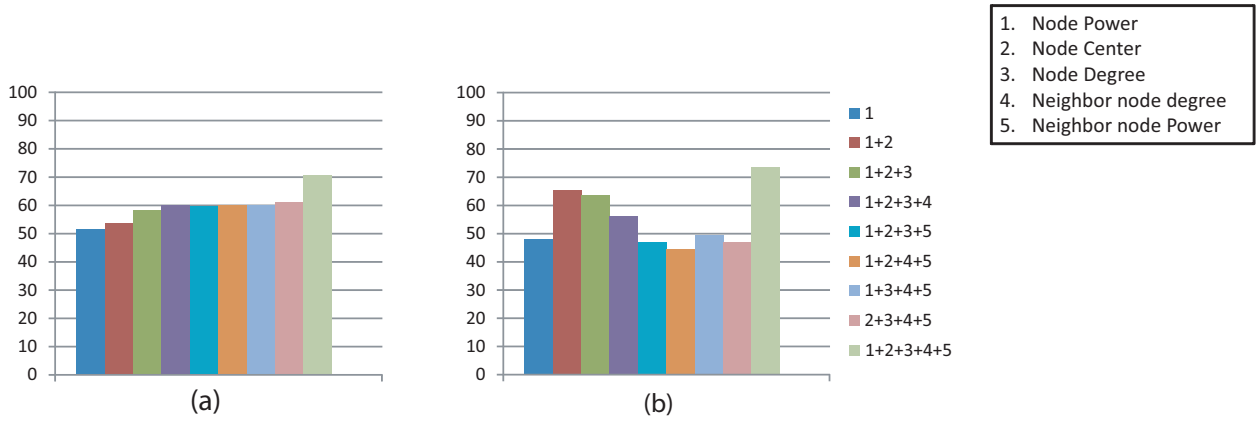


Figure 5.4: Figure plots the average detection accuracies on all the data centers when the inter-graph distances are computed using different subsets of the node attributes. The classification is performed on the male and female groups of subjects separately to achieve the reported results on (a) the released sets and (b) holdout sets.

Figure 5.2 shows how detection rates vary with different correlation thresholds used for the network computation. It can be seen that the peaks of the detection rates are not the same for the different data centers. There are two main potential reasons for this variation. First, there are variations in experimental protocols followed by the different data centers. Also, to capture the

data, different data centers used different scanner models and scanning parameters. Second, the subjects, participated in the different centers, have different age distributions. Mehnert et al. [50] found changes of functional connectivity measures with age in human brain. The variation of detection rate patterns across the centers indicates that there is a need to follow a more standardized experimental procedure for the future studies.

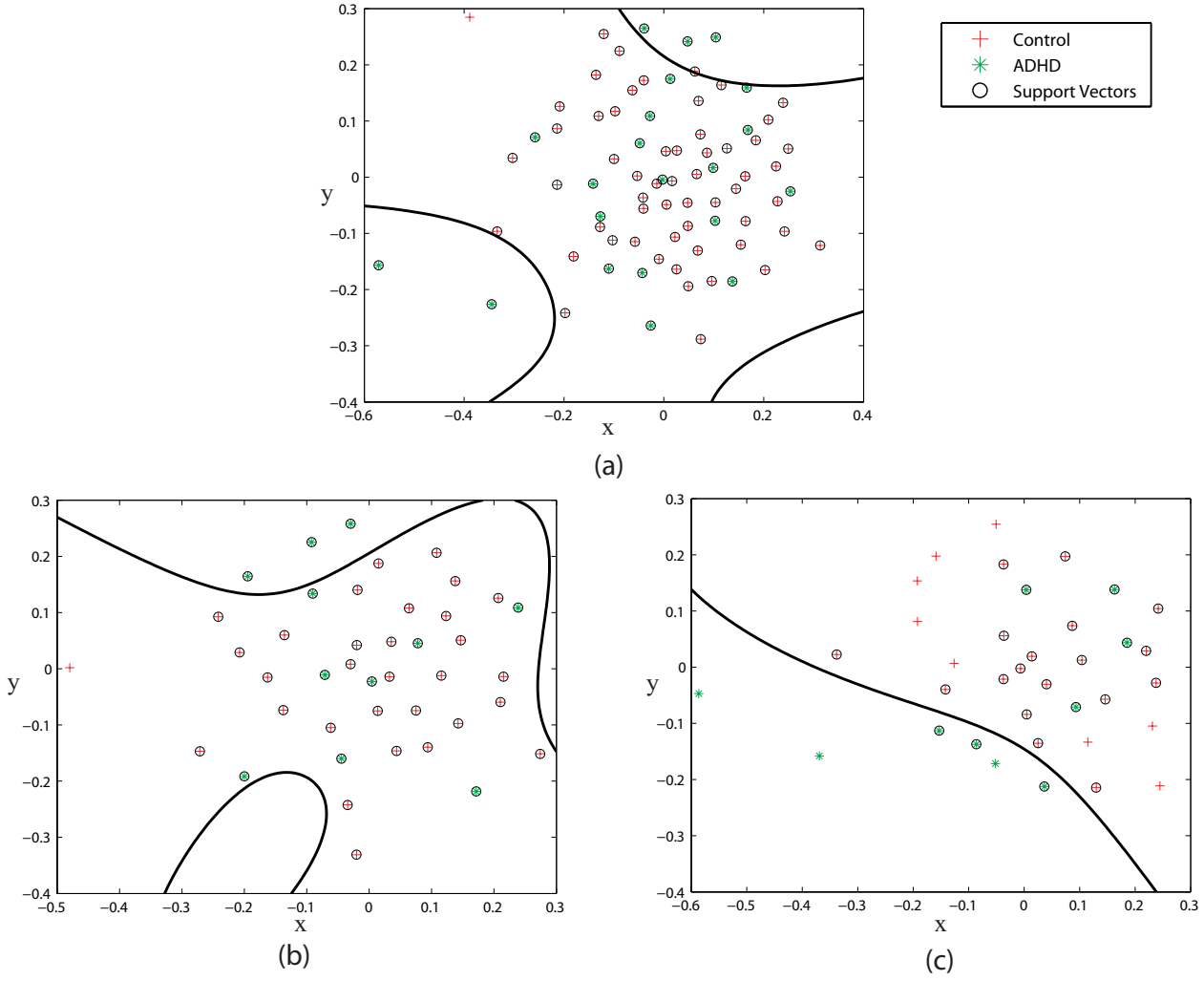


Figure 5.5: Subjects from the KKI released set are plotted on the MDS projected space. (a) All subjects, (b) subjects of the male group, (c) subjects of the female group. The spaces are segmented during the SVM training phase.

5.5 Conclusion

To summarize, we develop a novel classification framework which is modeled in a computationally efficient fashion as we are able to drastically reduce the size of the functional connectivity network by efficiently selecting the voxels and clustering them to form the nodes of the network. Also, our approach is able to produce impressive classification accuracies (70.49% on released data sets and 73.55% on holdout sets) especially on the holdout sets where we get the better detection accuracies than any of the previously reported results (67% by Bohland et al. [6] was the previous best). Our approach utilizes the global structure of the networks as we use the inter-network distances to project the networks in a 2 dimensional spatial configuration where the classification is performed. We provide physical interpretations of the dimensions of the projected space in our analysis. Also, we show the superior performance of our method over the standard network measures.

Table 5.3: Summarize the correlation values of the global features of the networks with the x and y dimensions of the projected spaces of the male and female groups.

Global features	x_{male}	y_{male}	x_{female}	y_{female}
Density	0.6906	0.3248	0.8310	0.1070
Global efficiency	0.4594	0.1924	0.5391	0.2578
Rich club coefficient	0.6367	0.4228	0.6482	0.4146
High power node fraction	0.3055	0.1984	0.1338	0.4942

CHAPTER 6: MULTIMODAL DATA FUSION TO IMPROVE ADHD DETECTION ACCURACY

In the last chapter of this dissertation we aim to address two aspects of the proposed classification problem. First, are structural brain images useful for the automatic diagnosis of the ADHD subjects? Second, can we further improve the classification accuracy when combining information from the functional and structural brain images?

To answer the first question, we used the Gray Matter (GM) brain image for our analysis. The GM image is the segmented sMRI image which contains only the GM regions of the brain. The GM regions are very important for brain cognitive tasks as they contain most of the neuronal cell bodies of the brain. The GM image for each subject is also provided with the ADHD-200 data-set. We used a Convolutional Neural Network (CNN) to extract the features from the GM images. Finally, the SVM is used for the classification.

To answer the second question, we use a separate classification framework using the 3-D power map image and fuse the detections obtained using the two modalities to deduce the final classification label. The power map concept is introduced in Section 5.2.1. A brain power map is constructed by computing the average power of the fMRI time series for each voxels of the brain volume. We compute the Local Binary Pattern (LBP) texture feature in the three orthogonal directions of the power map. The final representation of the LBP is a histogram for each subject of the data-set. The PCA-LDA classifiers are used for the final classification.

The experimental validation showed impressive classification accuracies using the GM (74.23%) and power map (77.30%) features. We use the late fusion to combine information from both of the data modalities which further improves the classification accuracy to 79.14%.

6.1 Method

The method is divided into mainly two parts. Section 6.1.1 describes the classification framework using the GM images and Section 6.1.2 describes the classification framework using the power map images. Lastly, the multi-modal data fusion is described in Section 6.1.3.

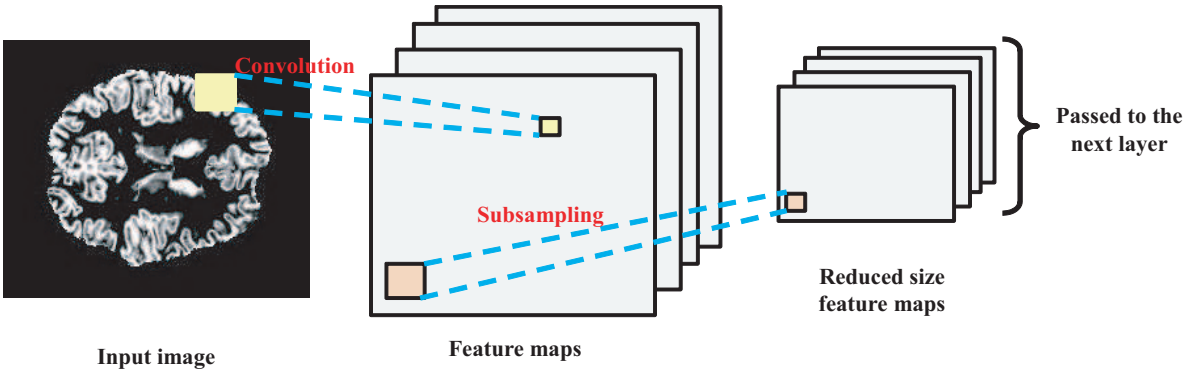


Figure 6.1: Figure shows the functionality of a CNN layer. First the input is convolved by a set of filters to produce the feature maps. Next the subsampling of the feature maps helps to reduce the map dimension. The reduced feature maps are then passed to the next layer for processing.

6.1.1 Classification Framework using GM Images

We first provide a short overview of the CNN for the better understanding of our method followed by a detailed description of the GM feature extraction and the classification framework.

6.1.1.1 CNN Overview

CNN is a variant of multilayer perceptron (MLP) which is a feed forward artificial neural network. The architecture of CNN is inspired by neurobiology, especially the neuron organization in the visual cortex of a cat. CNN was first introduced by K. Fukushima [33] and later improved by LeCun et al. [46].

As it is specifically designed for the image processing, CNN has some architectural advantage over MLP. For example, MLP has difficulties in learning object shape with spatial invariance i.e. learning to recognize object present in one location of the image does not transfer to learning to recognize the same object when it is present at a different image location. The other advantages are scale invariance, lower number of parameters to train etc. Each layer of CNN performs two functions; convolution and subsampling. Convolution is performed on the input of the layer by several small filters. Convolution with each filter generates a feature map of the input. Subsampling is used to reduce the size of the feature map. It also helps to add the position invariance property of the network. The down-sampled feature maps are then passed to the next layer for the processing. The concept is explained in Figure 6.1.

For our experiments, we used an already implemented CNN model called Caffe [42]. The network accepts input images of size 256x256x3. The network has 5 convolution and subsampling layers followed by two fully connected layers called FC6 and FC7. The convolution layer 1 to 5 has $96, 256, 384, 384$ and 256 filters of sizes $11 \times 11 \times 3$, $5 \times 5 \times 96$, $3 \times 3 \times 256$, $3 \times 3 \times 384$, and $3 \times 3 \times 384$ respectively. The max pooling is used for the subsampling of the feature map. We used the output of FC6 and FC7 layers to form the feature vector.

6.1.1.2 GM Feature Extraction

The ADHD-200 data-set comes with the 3-D GM image for each of the subjects (Figure 6.2). All of these images are of size $197 \times 233 \times 189$. The details information of the GM image can be found in the data description section (2.3.1). The 3-D GM image can also be considered as a stack of 2-D images which we refer to as slices. Slices are constructed by considering all the voxels of the $x - y$ plane while fixing the z dimension. Our algorithm treats each slice of the 3-D GM images independent of other slices. For this purpose the features from each of the slices are extracted separately for the classification. Later, the pieces of information from all of the slices are combined in a late fusion framework. For our experiments we consider one out of every 5 slices

starting from $z = 40$ to $z = 140$. This gave us 21 slices in total. The range is selected in such a way because the slices outside the range do not contain any useful brain region for our problem. Also, slices with similar z-axis values have very similar structure, which is why we selected one in every 5 slices. Slices are saved as $256 \times 256 \times 3$ JPEG images to be used for further processing. As the original size of the slices is 197×233, appropriate zero padding is performed at the borders of the images. Also, the GM images are gray-scale images and we repeat the gray-scale values of the slice in red, green and blue channels to produce the image of size $256 \times 256 \times 3$.

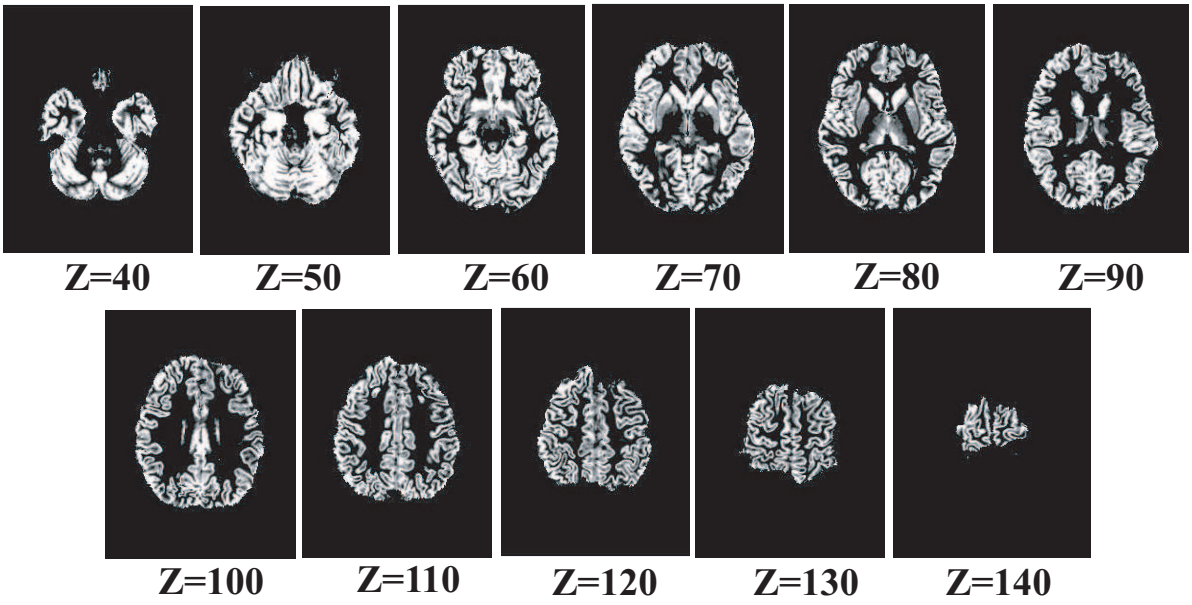


Figure 6.2: Figure shows different GM image slices of a subject.

We use a CNN implementation for extracting the features from the saved image slices. A CNN is believed to be able to automatically learn the feature representation useful for classifying any particular concept. The concept can be anything, for example objects, which can be linked to a pattern of data. We used a pre-trained model of CNN which is trained on a large image data-set released for the Large Scale Visual Recognition Challenge 2012 [38]. The data-set contains 1.2 million images with 1,000 categories. The data-set is so large that the network has learned a

generic representations of the filters which can extract the useful features for the image categories even if they don't belong to the training categories. This has proved to be useful in our case also as we obtained good classification accuracy using the features extracted from the pre-trained network. We only considered the features from the FC6 and FC7 layers. Each of the layers produces a feature vector of 4,095 dimensions. The final feature vector is formed by concatenating the feature vectors of the FC6 and FC7, resulting a 8,190 dimensional representation. Note, each of the dimensions of the feature vector is normalized in the range of $[0, 1]$. The feature extraction steps are described in Figure 6.3 $(a) - (c)$.

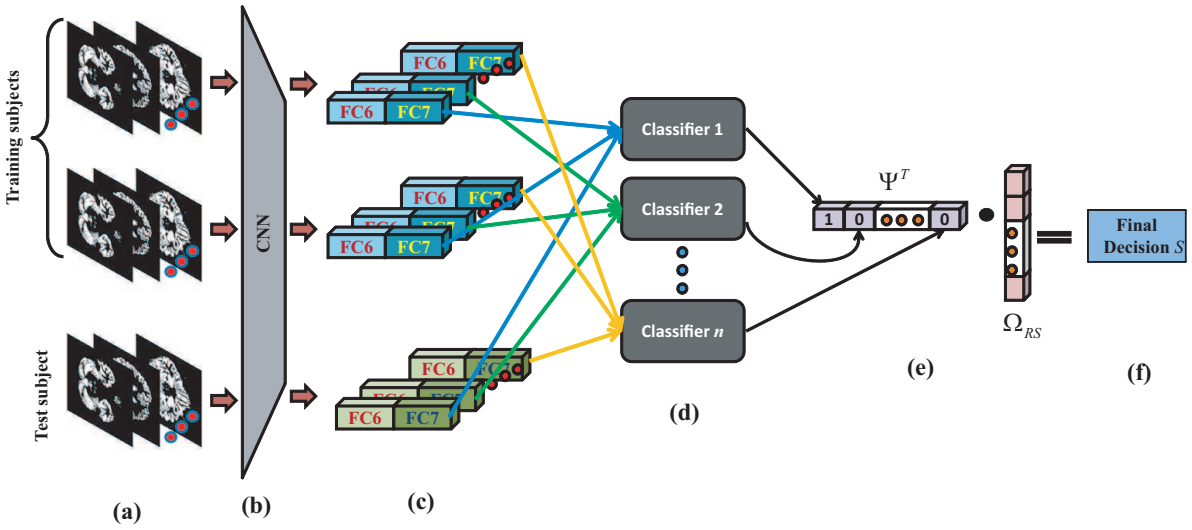


Figure 6.3: Flowchart of the GM classification framework: (a) GM images of the training and test subjects are provided to a pre-trained CNN (b) to extract features from FC6 and FC7 layers. (c) For separate slices, separate feature vectors are constructed concatenating the features from the FC6 and FC7 layers. (d) Separate classifier are trained for the separate slices to produce the decision vector Ψ. Dot product of Ψ and a weight vector Ω_{RS} generates the final decision score S.

Figure 6.4 shows some of the filters learned by the pre-trained CNN model in five convolution layers. The filters of the first layer are particularly intuitive as they learned to capture textures, color gradients, and edges in different orientation. The figure also shows the feature maps gener-

ated by the same layers for an example input image. Please note that we do not show all the filters and feature maps per layer for the ease of visualization. The filters of the first layer are colored because the size of the filters is $11 \times 11 \times 3$ which helped to plot them as color images. For the rest of the layers, we display the first 10 slices of the first 10 filters. The slices of the filters are arranged in the rows of the figure. For each layer of the CNN, convolving a filter with the input produces a feature map. Thus, total number of feature maps generated by a layer is equal to the number of filters of the layer. The figure shows the first 36 feature maps for each layer.

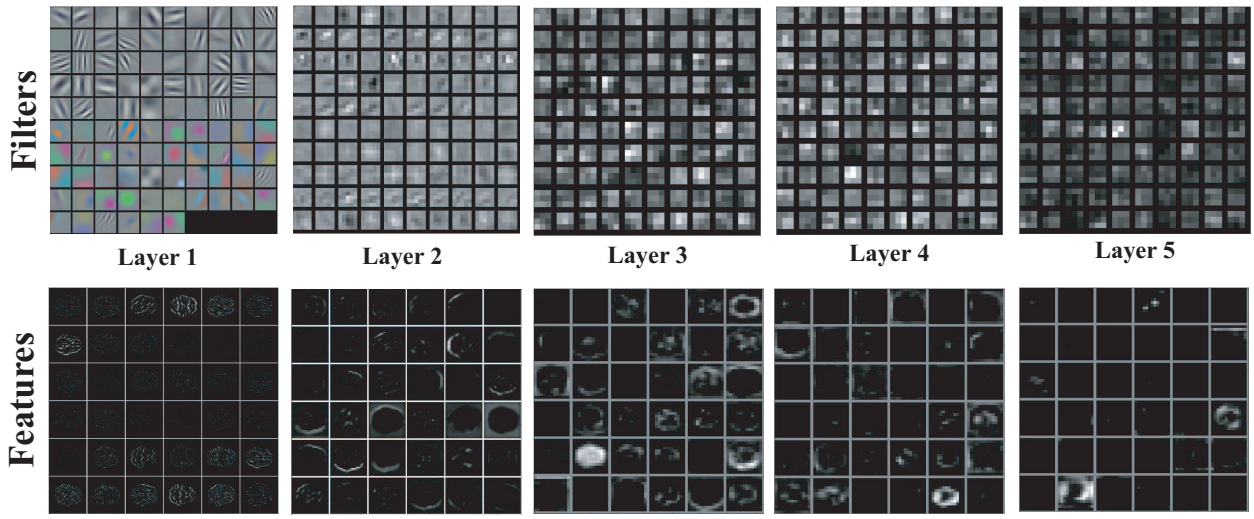


Figure 6.4: Figure shows some of the filters learned by the pre-trained CNN model for all five convolution layers and the corresponding feature maps generated for some input subject. Note that due to the space constraint, the figure is showing only a subset of the filters and features of each layer.

6.1.1.3 Classification

As stated, the features are extracted separately for the GM slices, and separates classifiers are trained using the extracted feature vectors. We use the Matlab implementation of the SVM classifier with the quadratic kernel. In total 21 classifiers are trained for 21 slices. Given a test

subject, each of the classifiers produces a diagnosis label for the subject. The concatenation of the diagnosis labels from all of the slices makes a decision vector Ψ. Elements of Ψ vector are combined in a late fusion framework to produce the final classification label. The fusion is performed in two stages. First, we compute a weight vector $\Omega = \{\omega_1, \omega_2, ..., \omega_\eta\}$ where ω_i represents the weight of the i_{th} slice and $\eta = 21$. For this purpose, we record the classification accuracy for the i_{th} slice by performing leave one out cross validation on the training set using the features from i_{th} slice only. Recording the accuracy values for all the slices forms the accuracy vector $AC = \{ac_1, ac_2, ..., ac_\eta\}$. Now each element of Ω can be computed as:

$$\omega_i = \frac{ac_i}{\sum_{j=1}^{\eta} ac_j}. \tag{6.1}$$

Ω is used to ranked the slices based on their weight values. Slices with higher weights get higher ranks. In the second step, a sigmoid function is used to further re-scale the weight vector so that the weights of higher ranked slices get a boost. This step produces a re-scaled weight vector Ω_{RS} as follows:

$$\omega_{RS_i} = \omega_i \times (1 - \frac{1}{1 + e^{-(rank(\omega_i) - \eta/2)}}), \tag{6.2}$$

where ω_{RS_i} is the i_{th} element of Ω_{RS}. Final decision score S is achieved as follows:

$$S = \Psi^T \cdot \Omega_{RS}. \tag{6.3}$$

A decision threshold is applied on S to detect the ADHD label.

For each of the five data centers (KKI, NeuroIMAGE, NYU, OHSU and Peking) the framework is used separately by considering the released set as the training data and the hold out set as the test data.

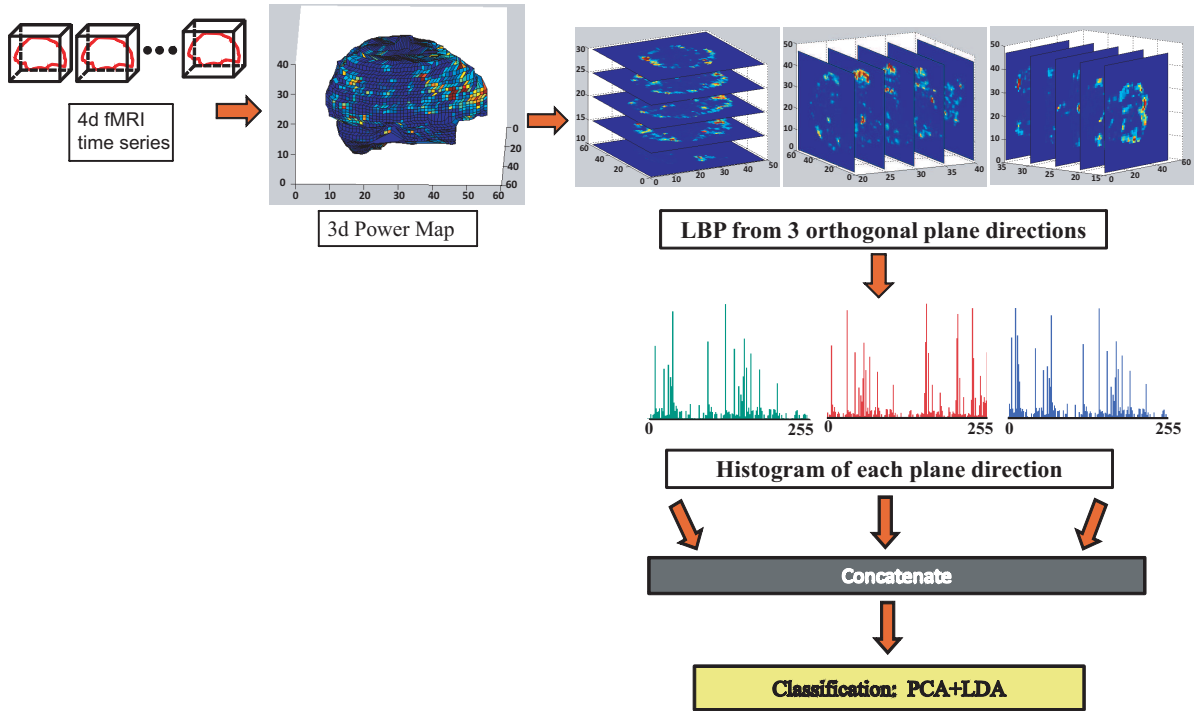


Figure 6.5: Flowchart of the power map classification framework: First, the 3-D power map image is generated from the 4-D fMRI data. Next, the LBP texture features are computed in three orthogonal plane directions of the power map image. The classification is performed using the PCA-LDA classifier.

6.1.2 Classification Framework using Power Map Images

The concept of power map is first introduced in Section 5.2.1. A power map is constructed by computing the average power of the fMRI time series of each voxel of the brain volume. In this section, we further analyze the role of the power map only for solving the ADHD diagnosis problem. Hence, we do not use any functional connectivity network which requires the fMRI time series to be constructed. The whole classification framework is explained in Figure 6.5. As it can be seen, the power map for each subject is a 3-D image. We compute the LBP texture feature of the power map in three orthogonal plane directions. A PCA-LDA classifier is used for the final classification. Feature extraction and classification steps are explained in details in the following

sections.

6.1.2.1 Power Map Feature Extraction

LBP is an image texture feature originally introduced by Ojala et al., 1996 [55] and Ojala et al., 2002 [54]. Recently, Chang et al., 2012 [16] used the LBP feature on the structural brain images for automatic ADHD detection but their best detection accuracy (69.95%) is much lower than what we achieved. The steps involve in LBP feature computation on a 2-D image are explained in Figure 6.6 . For our experiments, we compute the LBP features of the 3-D power map on three orthogonal plane directions. Finally, we concatenate the features from each of the plane directions to construct the 3-D image feature.

The LBP operator for a voxel v can simply be defined as follows:

$$LBP(v) = \sum_{p=1}^{P} sign(pow(v_p) - pow(v))2^{p-1}, \tag{6.4}$$

where

$$sign(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

P is the number of neighbor voxels, v_p is the p_{th} neighbor voxel, function $pow(.)$ returns the power of the input voxel. For our experiments, we only considered the immediate 8 neighbours of a voxel. Hence, the LBP score of any voxel is always in the range of $[0, 255]$. Again, the neighbour voxels are indexed in a particular order as shown in Figure 6.6. Once the LBP scores of all the voxels for each of the three plane directions is computed, a histogram of LBP scores is constructed per subject per plane direction. Each histogram consist of 256 bins which represent 256 possible LBP scores. As the final feature vector is computed by concatenating the histograms of all three plane directions, the total feature vector size becomes 768.

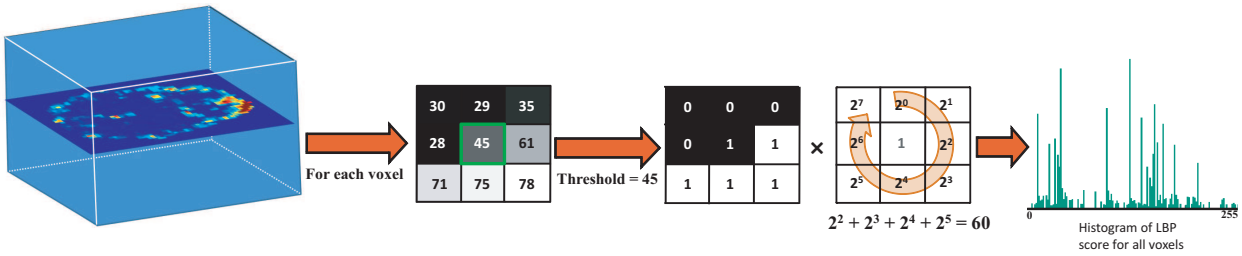


Figure 6.6: Figure describes the LBP feature computation on a 2-D image. First, for each voxel immediate 8 neighbour voxels in the plane direction are identified. Then, a neighbour voxel is assigned a value 0 if its power value is less the center voxel's value. Otherwise it is assigned a value 1. Next, the binary values of all the neighbour voxels are multiplied by different powers of 2 in a particular order and summed. This is the LBP score of the center voxel. Finally, the histogram of LBP scores is computed for all the voxels of the brain volume under consideration.

6.1.2.2 Power difference image formation

We identified the key regions with power differences between the ADHD and control groups which are shown in Figure 6.9. The figure shows the average difference of power between the ADHD and control groups as they are plotted on the different brain slices. The figure is generated using the power maps of the KKI released and hold out data sets' subjects. For any voxel $v_{DM_{i,j,k}}$ of the power difference map showing high power regions of the control group, the power value is computed as follows:

$$pow(v_{DM_{i,j,k}}) = \begin{cases} 0, & \delta \leq 0 \\ \delta, & \delta > 0 \end{cases}$$

where

$$\delta = \frac{1}{C}\sum_{c=1}^{C} pow(v_{c_{i,j,k}}) - \frac{1}{A}\sum_{a=1}^{A} pow(v_{a_{i,j,k}}). \tag{6.5}$$

C and A are the control and ADHD subject counts. The power values of the voxels of the power

difference map, showing the high power regions of the ADHD group, can be computed in a similar fashion. False Detection Rate (FDR) controlling technique, introduced by Benjamini and Hochberg, is applied on the image as it is described in [34] (Genovese et al.). The FDR controlling technique guarantees that the average false detection rate will be less than a parameter value q (0.01 in our case) specified in the algorithm. The algorithm works as follows. First, the voxels are sorted in the ascending order according to their P values. P value for each voxel is calculated for the null hypothesis that the voxel has no statistical power difference in the ADHD and control subject groups. Finally, all the voxels with P values lower than the P_i are selected where i is the largest index such that the following condition satisfies:

$$P_i \leq \frac{i}{V}\frac{q}{c(V)}, \tag{6.6}$$

where V is the total voxel count and $c(V)$ is a constant whose value is 1 in our case. The final selected voxels are plotted in the power difference images.

As it can be seen, the high power regions of the control group are more evenly spread across the brain slices while the high power regions of the ADHD group are distributed as isolated small clusters. We performed similar analysis on the subjects of the other data centers where we observed similar patterns for the NeuroIMAGE (Figure 6.10), NYU (Figure 6.11), and OHSU (Figure 6.12) data centers. Surprisingly, for the Peking data center (Figure 6.13) an opposite trend is observed where the average high power regions of the control group is spread out in the brain volume while the average high power regions of the ADHD group are small segmented regions.

6.1.2.3 Classification

Similar to the classification framework using GM image features, classification framework is used separately on each of the data centers. For each data center, the hold out set is used as the training data and the released set is used as the test data. For all our experiments on the power

map we used the Matlab implementation of the LDA classifier preceded by PCA. First we compute an average power map of the training subjects and select the voxels whose average power values are greater than a threshold. Only the selected voxels are used for the LBP feature computation. We varied the power threshold from 0.05 to 0.40 with an interval of 0.01. For each value of the threshold, a different set of feature vectors are constructed, classifiers trained and accuracies are recorded. The reason behind choosing the particular threshold range is because beyond either end of the range detection accuracies generally drop rapidly.

For the purpose of the comparison of the classification accuracies, we performed a set of experiments on the raw power map features. The raw power map feature vector is formed by selecting the voxels with average power value greater than the power threshold and arranging their power values in a vector. The power threshold range is the same as in the case of the LBP features.

6.1.3 Multi-modal Data Fusion

We use a simple late fusion model for combining the GM and power map information. We employ a voting using the final decisions of the GM and power map classification frameworks. As we are dealing with only two votes, a subject is classified as ADHD if any of the decisions is positive.

Table 6.1: Summary of the results: showing the best detection results for all different methods and their corresponding specificities and sensitivities.

	FC6-FC7			**LBP**			**FC6-FC7 & LBP**		
	Det	sens	spes	Det	sens	spes	Det	sens	spes
KKI	81.82	33.33	100	90.91	100	87.50	90.91	100	87.50
NeuroIMAGE	68.00	81.82	57.14	88.00	72.73	100	72.00	90.91	57.14
NYU	73.17	89.66	33.33	78.05	86.21	58.33	75.61	100	16.67
OHSU	88.24	50.00	96.43	85.29	16.67	100	91.18	66.67	96.43
Peking	66.67	41.67	88.89	62.75	25.00	96.30	74.51	62.50	85.19
Average	74.23	67.57	79.55	77.30	58.11	92.50	79.14	83.78	75.00

6.2 Results

As stated, for all our experiments classification is performed separately on the separate data centers. The released set of each data center is considered as the training data and the hold out set is considered as the test data.

Different LBP features are calculated by varying a power threshold and each time selecting the set of voxels whose average power value is greater than the threshold. For each set of LBP features, ADHD detection accuracy is recorded. Figure 6.7 (a) plots the power threshold vs classification accuracies for all the data centers. The average detection accuracies of all the data centers for different power threshold values are also plotted. As it can be seen the highest average detection accuracy value is achieved for the power threshold value of 0.21. Fusion of the FC6-FC7 and LBP features are also performed for the different power threshold values (Figure 6.7 (b)). The best average detection accuracy for the fusion feature is also achieved for the same power threshold value ie. when the FC6-FC7 features are combined with the LBP features computed for the power threshold value of 0.21.

For the comparison of the detection accuracies, classifications are performed using GM feature vectors of the FC6 layer only, the FC7 layer only and concatenation of the FC6 and FC7 layers. The average accuracy of the experiments for all five data centers are plotted in Figure 6.8. As it can be seen, we achieve the best classification accuracy (74.23%) when we concatenate the feature vectors from the two layers. Also, the classification experiments are performed using the features from the White Matter (WM) and normalized whole brain images. The WM and normalized whole brain images are also structural brain images containing segmented white matter regions and whole brain regions respectively. The feature extraction and classification frameworks on the WM and whole brain images are same as the GM classification framework. Finally, late fashion is used to combine the LBP features with GM FC6-FC7, WM FC6-FC7 and whole brain FC6-FC7 features respectively. Late fusion of the GM FC6-FC7 and LBP features gives us the

overall best classification accuracy which is 79.14%. Figure 6.8 plots the classification accuracy of all of the experiments.

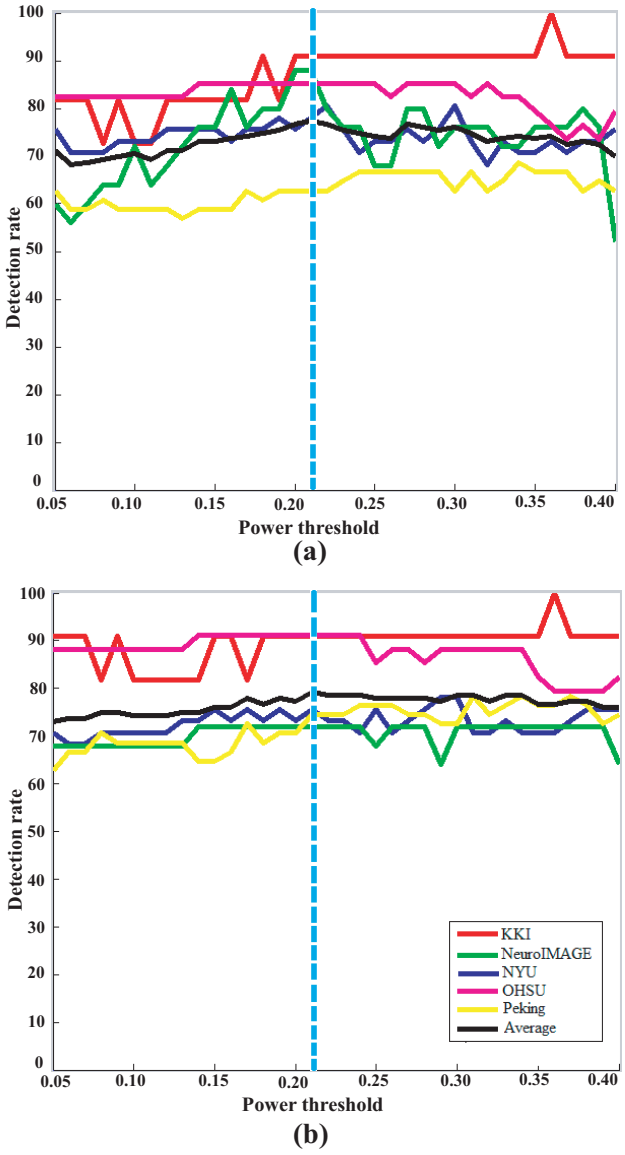


Figure 6.7: Figure plots the power threshold vs detection rates generated using the LBP features computed on different data centers. Average detection rates for the different power threshold values are plotted in black. Dotted blue line indicates the power threshold value for which the highest average detection rate of all the data centers is achieved.

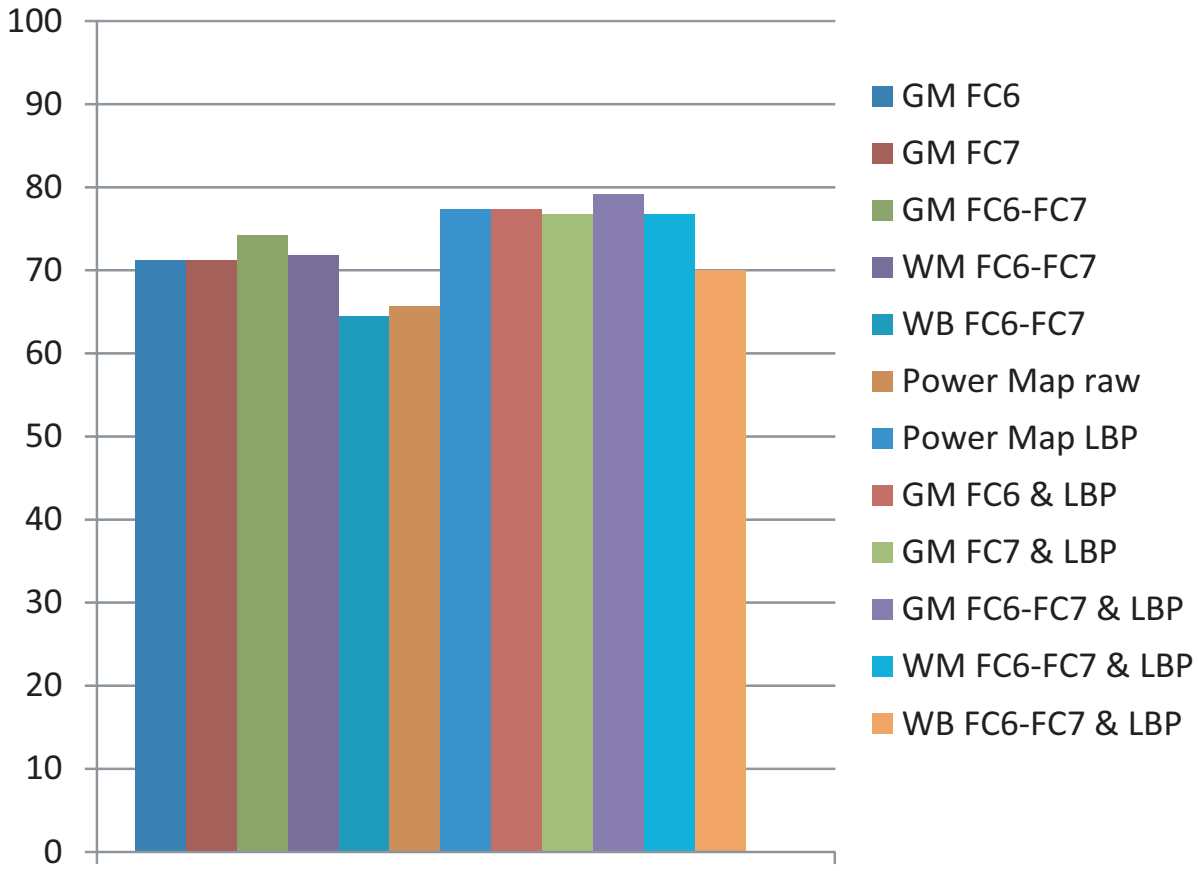


Figure 6.8: Figure plots the average detection rates on all the data centers using different feature combinations. GM stands for the gray matter, WM stands for the white matter and WB stands for the whole brain.

For each of the data centers, the classification accuracies along with the sensitivity and specificity values are listed in Table 6.1.

6.3 Discussion

In this chapter we argued that the brain structural images contain useful information for solving the ADHD diagnosis problem. To verify our claim we use an already implemented CNN model to extract the features from the GM images of the brain. Our experiments show that the extracted features can classify the ADHD and control subjects with an impressive accuracy. The CNN model we used is pre-trained using a very large data-set containing 1000 object categories and 1.2 million images. This helped the network to learn to extract features in a generic fashion such that the features can classify the objects categories even if they are outside of the training categories. We noticed that the GM is the most useful brain regions for solving the ADHD diagnosis problem as the other two structural image formats, i.e. the WM and whole brain images, didn't perform that well. At the end, we combined the output of the CNN for each of the GM slices in a novel late fusion framework to achieve a higher classification accuracy.

For our functional data based approach, we used the 3-D power map images which is derived from the fMRI data. The concept of the power map is introduced in the previous chapter (5.2.1) where it is used to select the highly active voxels for the functional network construction. In this chapter, we investigate if the distribution of the average voxel powers can reveal any difference of patterns between the ADHD and control groups of subjects. For this purpose we compute the LBP texture features on three orthogonal directions of the power map image. LBP is a global feature which can encode the texture pattern around a voxel into a number between $[0, 255]$. The histograms constructed from the LBP feature estimates the count of different texture features appearing in an image. We achieve the state of the art classification accuracy (77.14%) using the LBP features.

We notice that our findings of the power difference regions are consistent with the existing literature. Vincent et al. [80] and Castellanos and Proal [14] have investigated the role of the fronto-parietal network in performing executive control tasks. The frontal pole is known to be a part of this

network, and our method is able to identify this region as shown in Figure 6.9, panel with $z = 16$. Clark et al. [18] have reported right-frontal cortex abnormalities in ADHD. We identify the right frontal orbital cortex (Figure 6.9, panel with $z = 16$) as a region where the controls have higher power than ADHD subjects. Schachar et al. [64] studied response inhibition deficits in the context of ADHD subjects. Diane and Victoria [27] demonstrated the role played by the left inferior frontal gyrus in response inhibition tasks. From our analysis we also identify the same region (Figure 6.9, panel with $z = 20$). Several studies have shown diminished activity in the precuneus region of ADHD subjects vs. controls, such as Cao et al. [11] and Castellanos et al. [13]. We also obtain a similar result as we found high power in the precuneus region for the control subjects (Figure 6.9, panel with $z = 24$). Dickstein et al. [28], in their paper, compare regions in the brains of control subjects that are hyperactivated with respect to ADHD subjects. Many of the regions they identified are in agreement with the regions shown in Figure 6.9, such as the inferior frontal gyrus ($z = 20$) and the precentral gyrus ($z = 24$). These regions have been implicated in tasks involving executive function and inhibition. Sharp et al. [66] showed that the lateral occipital cortex, which shows up in our finding (Figure 6.9, panel with $z = 24$), is also implicated in inhibitory tasks that are studied using a stop signal task. Furthermore, the lateral occipital cortex is also involved in spatial attention tasks, as shown by Silk et al. [69].

Finally, we are able to further improve our classification accuracy by combining the GM and power map information in a late fusion framework. This indicates that the structural and functional data modalities might share complementing information.

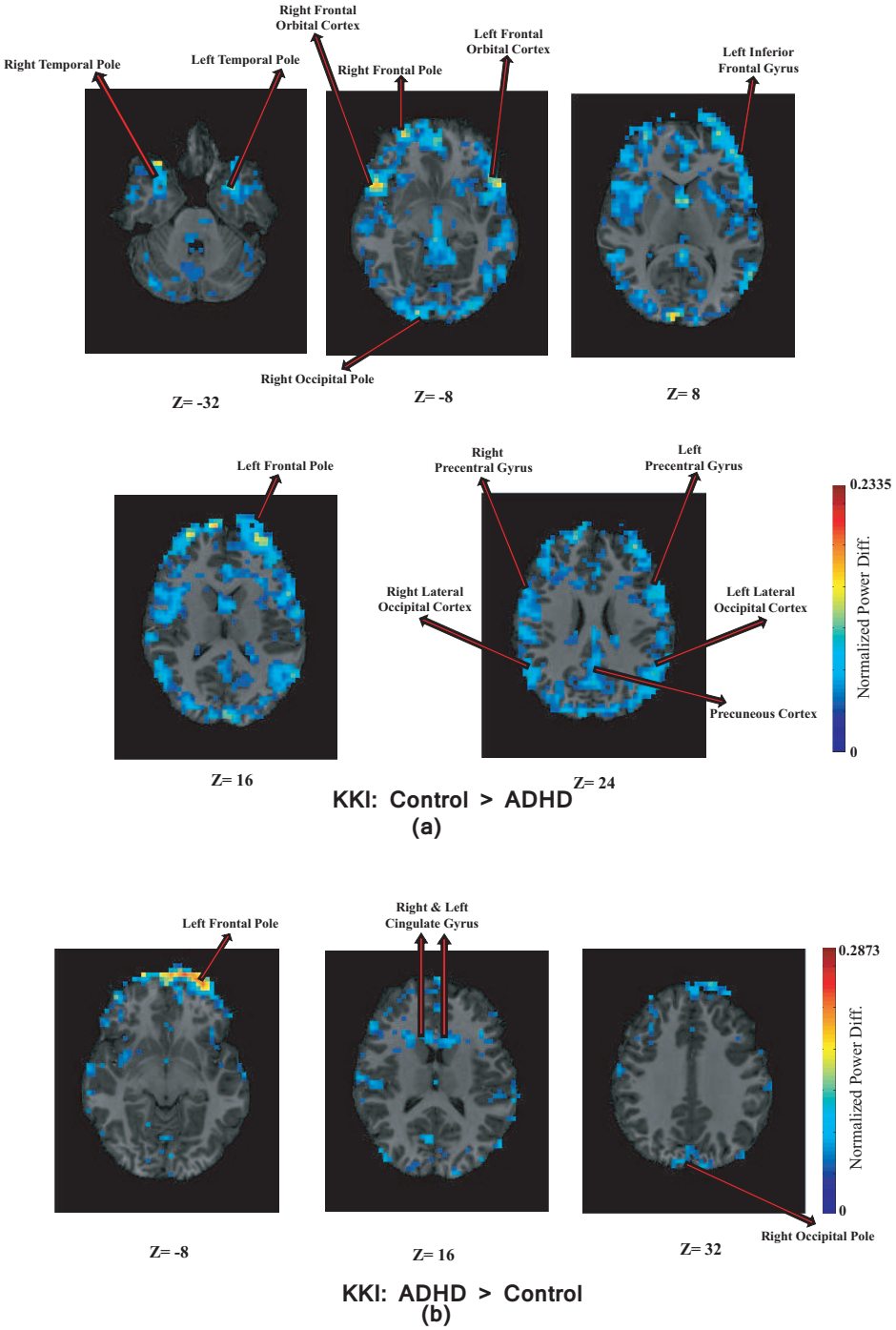


Figure 6.9: Plots of the average power differences of the control and ADHD groups of the KKI released data set. Power differences are plotted on the different brain slices. The top and middle rows are showing the regions where the control group has higher power while the bottom row is showing the regions where the ADHD group has higher power.

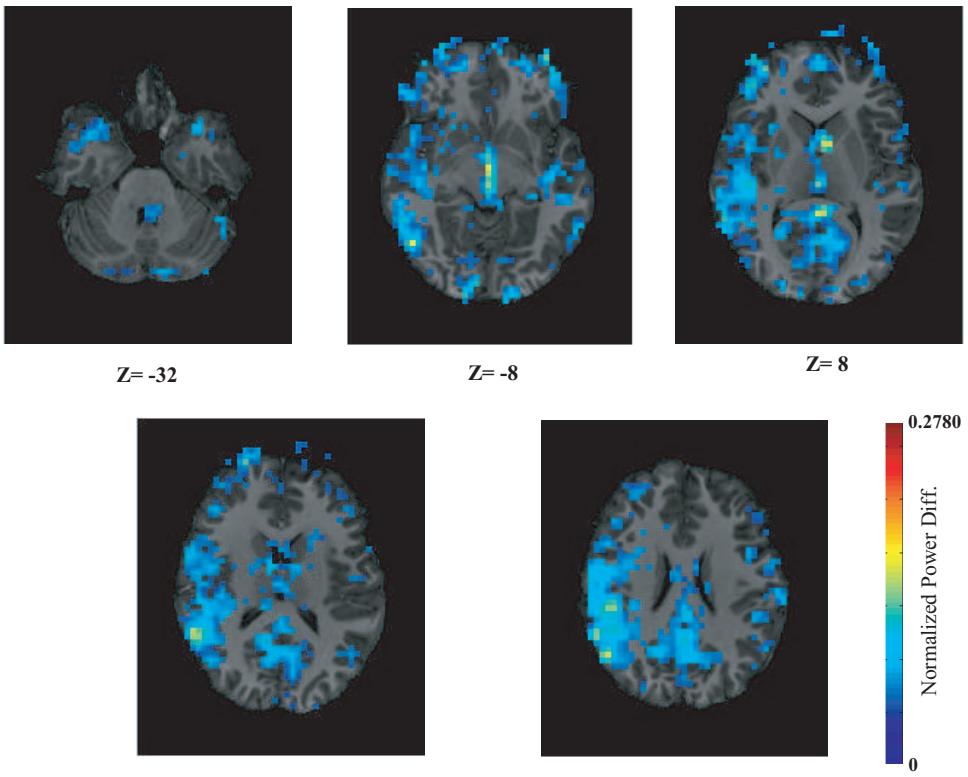


Figure 6.10: Plots of the average power differences of the control and ADHD groups on the subjects of NeuroIMAGE released and hold out set on different brain slices. (a) shows the regions where control group has higher power, (b) shows the regions where ADHD group has higher power.

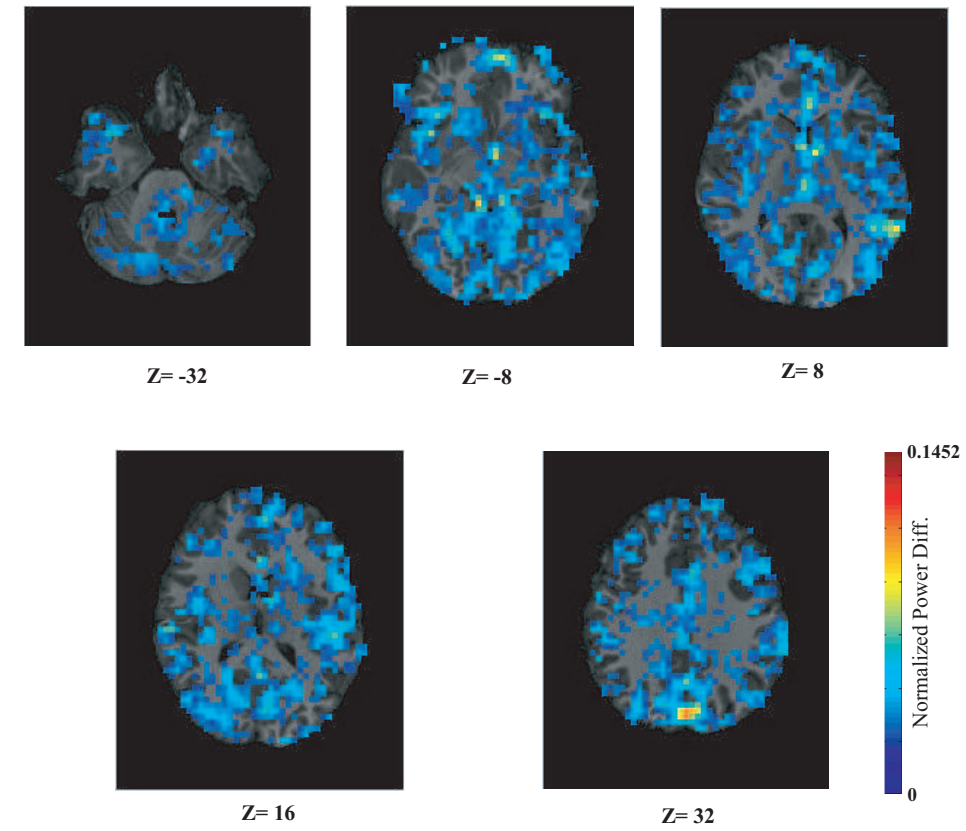


Figure 6.11: Plots of the average power differences of the control and ADHD groups on the subjects of NYU released and hold out set on different brain slices. (a)shows the regions where control group has higher power, (b) shows the regions where ADHD group has higher power.

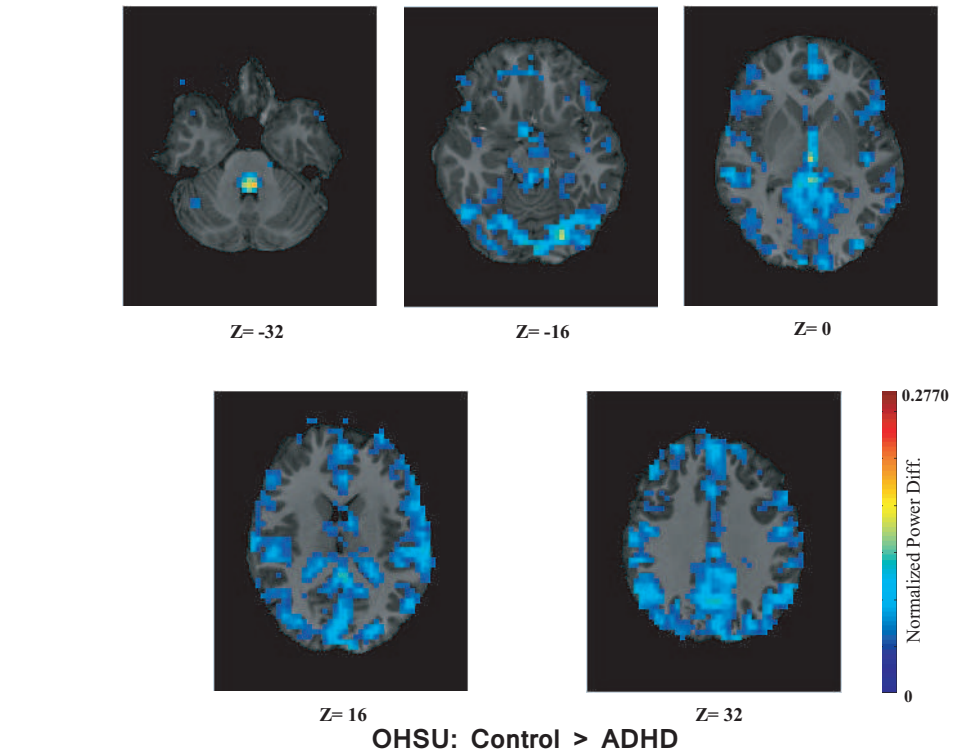
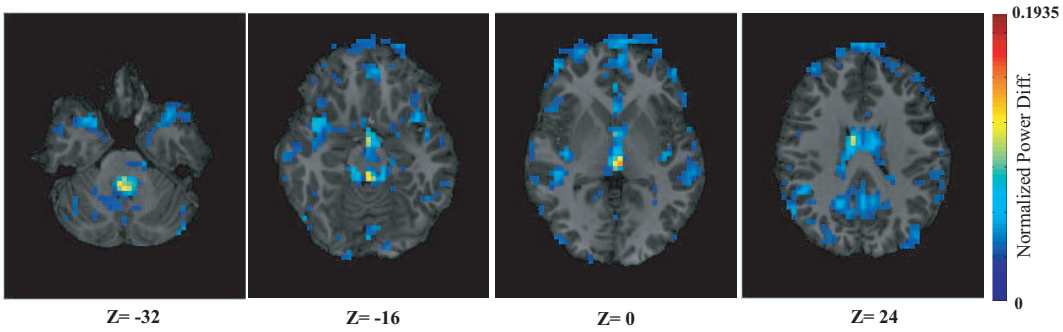


Figure 6.12: Plots of the average power differences of the control and ADHD groups on the subjects of OHSU released and hold out set on different brain slices. (a)shows the regions where control group has higher power, (b) shows the regions where ADHD group has higher power.



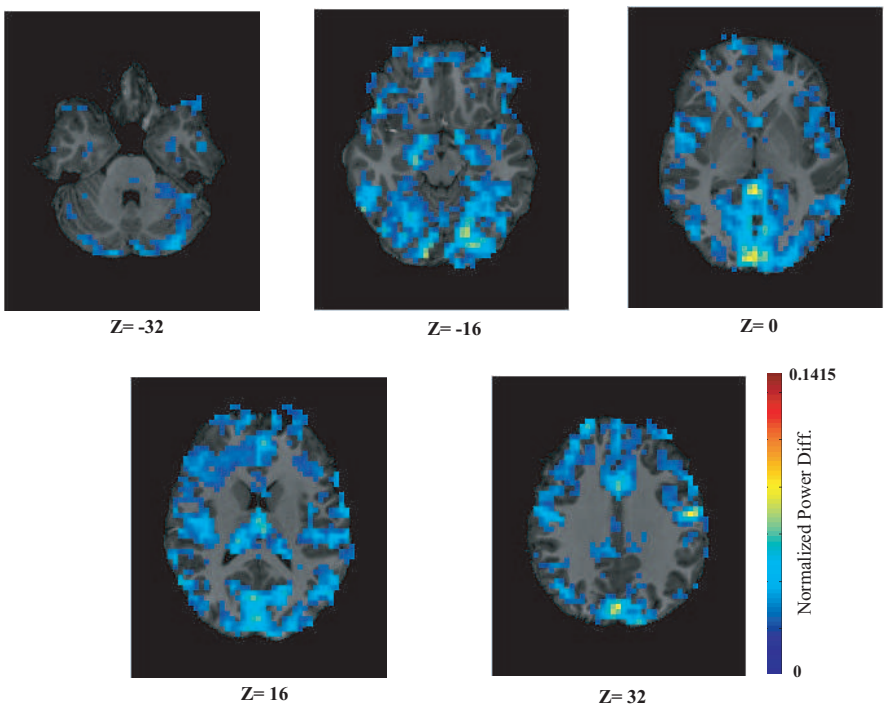


Figure 6.13: Plots of the average power differences of the control and ADHD groups on the subjects of Peking released and hold out set on different brain slices. (a)shows the regions where control group has higher power, (b) shows the regions where ADHD group has higher power.

6.4 Summary

In summary, we showed that the brain structural images contain useful information related to ADHD diagnosis problem as we received high classification accuracy using the GM features. We also analysed the 3-D power map images derived from the brain functional data. Our study showed differences in power map patterns between the ADHD and control groups of subject. The LBP features are able to encode the pattern differences as we achieve the state of the art classification accuracy on the ADHD-200 hold out sets. Finally, combination of the GM and power map features helped to further improve our classification accuracy.

CHAPTER 7: CONCLUSIONS AND FUTURE WORK

In this dissertation, we addressed the problem of automatic detection of the ADHD subjects using their brain rs-fMRI data. The problem is particularly of importance due to the widespread impact of the ADHD on the global child population and the lack of biological measures to diagnose it. Approximately $5 - 10\%$ of the children all over the world are diagnosed with ADHD. These motivate us to propose a solution for the automatic ADHD diagnosis problem. The central idea of our approach is to model the resting state brain activities as a network which we refer to as the functional connectivity network. We exploited the topological differences of the networks between the ADHD and control groups of subjects for the classification processes. Lastly, we showed that the functional and structural brain images share complementary information as the combination of information from both of these modalities helped to achieve a better classification accuracy than any of the modalities. In Table 7.1 we have listed the best classification accuracies of all our approaches along with the other top performing results in the literature.

Table 7.1: List of the best classification accuracies of our approaches (marked in bold) and other top performing approaches in the literature.

Dai et al. [24]	Bohland et al. [6]	Sidhu et al. [68]	BoW	Nw. feature	Attributed Nw.	Multi-modal data
61.54%	66.67%	71.35%	**64.81**%	**69.59**%	**73.55**%	**79.14**%

Our first approach for solving this problem used BoW framework to cluster the node degrees of the network. Final representation of the BoW is a histogram of degree features per subject which is treated as the feature vector to be used by the classifier. We achieved 64.81% accuracy using this approach. The BoW approach has few problems. First, it loses the spatial information of network nodes since the histogram does not contain any spatial information. Second, BoW approach extract features from the whole network whereas some brain regions may not contain any useful information. Thus, it may unnecessarily increase the feature dimension and noise of the

system. Third, the approach only employs the degree features where other network features may also be useful.

Towards addressing the shortcomings of the BoW approach, we first investigated if only some selected regions of the the brain volume contain the useful information for the ADHD diagnosis problem. Our proposed algorithm is able to successfully identify the important brain regions and experimental results suggest that the classification accuracy improves when we extract the features from the selected regions only. The regions selected by our algorithm are similar to the regions identified by many other independent studies in the existing literature on ADHD. Next, we construct the feature vector by concatenating the network features from the nodes of the selected regions only. As the concatenation is performed in a fixed order, it helped to preserve the relative spatial information of the nodes. Finally, along with the degree features, we evaluate three complex network features such as the network cycles, the varying distance degree and the edge weight sum. We are able to achieve 69.59% classification accuracy using this approach. However, as we represent each voxel of the brain volume as a node of the network, it makes the node count of the functional network several thousand which is computationally very expensive. Also, the network features, which are computed for each node, can only capture the local structures of the network ignoring the global network topology.

Next, in order to exploit the global structures of the networks in our classification framework, we use MDS technique to project the networks from an unknown network-space to a low dimensional space based on their inter-network distance measures. Also, we significantly reduce the computation cost for the construction of functional network as we propose an efficient representation of the nodes such that the network can preserve the maximum relevant information with minimum redundancy. For this purpose, we represent each node as the cluster of highly active voxels where the activity levels of the voxels are measured based on the average power of their corresponding fMRI time-series. As a result, the number of nodes per network is reduced to 60 on average compared to 28000 voxels in the brain volume. Our approach is able to achieve a classi-

fication accuracy of 73.55% on the ADH-200 hold out set. Our results show that the classification accuracies significantly improve when experiments are performed separately on the male and female groups. One possible reason is the differences of brain functioning of the male and female subjects.

Finally, we focused on answering two questions. First, is the structural brain image useful for solving the proposed problem? Second, if it is then can we improve the accuracy of the diagnosis system by fusing information of the structural and functional data? For the structural data modality, we use the GM brain images while for the second modality we use the power map images which are derived from the rs-fMRI data. Both of the modalities showed impressive classification accuracies as we received 74.23% accuracy using GM images and 77.30% using power map images on the ADHD-200 hold out data set. Combining information from the two modalities further improves the accuracy to 79.14%.

In summary, this dissertation showed enough evidence that the brain imaging data contains useful information for the diagnosis of ADHD subjects. At present the accuracy is not high enough to be used as the biological measure of the problem but it can be used as the supporting evidence with the manual diagnosis. Further investigation regarding standardization of data resolution and data capturing protocols are needed to increase the reliability of the automatic diagnosis process.

7.1 Future Work

The brain imaging based methods showed promise for solving the proposed problems as different independent studies reported ADHD detection accuracy higher than a chance factor. Still, there are many areas to improve on because none of the method is good enough to replace the current manual diagnosis process. Further investigations need to be performed regarding the data capturing protocols and the community needs to decide on a standard method as different protocols may lead to the variations of cognitive activities of brain which can reduce the performance of the

diagnosis method.

In our approach we model the brain functions as a network which connects different brain regions based on their correlations of activity patterns. The network constructed in this process is static as the weight of an edge connecting two regions is computed based on the correlation of the whole fMRI time series of the two regions. Therefore, does not change over time. One interesting idea to try is to compute the correlation on two local windows of the time series. Thus, if we slide the windows along the time series and each time computes a different correlation value, the edge weight will be a function of time. The analysis of patterns of the changing edge weights in the network can be useful for this problem. Also, to reduce the network computation cost, we used a particular ROI map to cluster the voxels to form the nodes of the network. But we didn't draw any conclusion as to which ROI map is the best for this problem. For future work, different ROI maps can be tried to get more insight on this.

To verify the usefulness of structural brain images, we use a CNN model to extract features from GM brain image slices. We treated each of the slices independently as we use separate classification framework for the features extracted from each slice. Later we used a late fusion framework to combine the information from different slices. One possible direction is, instead of treating the slices separately, the CNN network can be modified to extract features from the whole brain volume. Also, we used a CNN model which was pre-trained on a large image data set. There are two other possible approaches to explore in future to train the network. First, one can start with the pre-trained model and fine tune the network weights by further training using GM images. Second, a network can be trained from scratch. In either way, training a CNN requires lot of sample data so that the filters can learn to extract relevant features. Also, training from scratch can be tricky as it needs lot of parameters to decide on such as learning rate of the network, number of network layers, number of filters per layer, size of the filters in each layer etc.

LIST OF REFERENCES

[1] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[2] H. Ahrens. Seber, G. A. F.: Multivariate Observations. J. Wiley & Sons, New York 1984, xx, 686 s. 48, 50. *Biometrical Journal*, 28(6):766–767, 1986.

[3] M. Assaf, K. Jagannathan, V. D. Calhoun, L. Miller, M. C. Stevens, R. Sahl, J. G. O'Boyle, R. T. Schultz, and G. D. Pearlson. Abnormal functional connectivity of default mode subnetworks in autism spectrum disorder patients. *NeuroImage*, 53(1):247 – 256, 2010.

[4] J. Biederman. Attention-deficit/hyperactivity disorder: A selective overview. *Biological Psychiatry*, 57(11):1215 – 1220, 2005.

[5] S. Blint, P. Czobor, S. Komlsi, . Mszros, V. Simon, and I. Bitter. Attention deficit hyperactivity disorder (adhd): gender- and age-related differences in neurocognition. *Psychological Medicine*, 39:1337–1345, 8 2009.

[6] J. W. Bohland, S. Saperstein, F. Pereira, J. Rapin, and L. Grady. Network, anatomical, and non-imaging measures for the prediction of adhd diagnosis in individual subjects. *Frontiers in Systems Neuroscience*, 6(78), 2012.

[7] M. R. G. Brown, G. S. Sidhu, R. Greiner, N. Asgarian, M. Bastani, P. H. Silverstone, A. J. Greenshaw, and S. M. Dursun. Adhd-200 global competition: Diagnosing adhd using personal characteristic data can outperform resting state fmri measurements. *Frontiers in Systems Neuroscience*, 6(69), 2012.

[8] G. Bush, J. A. Frazier, S. L. Rauch, L. J. Seidman, P. J. Whalen, M. A. Jenike, B. R. Rosen, and J. Biederman. Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fmri and the counting stroop. *Biological Psychiatry*, 45(12):1542–1552, 1999.

[9] G. Camps-Valls, N. Shervashidze, and K. Borgwardt. Spatio-spectral remote sensing image classification with graph kernels. *Geoscience and Remote Sensing Letters, IEEE*, 7(4):741–745, Oct 2010.

[10] Q. Cao, Y. Zang, L. Sun, M. Sui, X. Long, Q. Zou, and Y. Wang. Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study. *NeuroReport*, 17(10):1033–1036, 2006.

[11] Q. Cao, Y. Zang, C. Zhu, X. Cao, L. Sun, X. Zhou, and Y. Wang. Alerting deficits in children with attention deficit/hyperactivity disorder: Event-related fmri evidence. *Brain Research*, 1219(0):159 – 168, 2008.

[12] F. X. Castellanos, J. N. Giedd, W. L. Marsh, S. D. Hamburger, A. C. Vaituzis, D. P. Dickstein, S. E. Sarfatti, Y. C. Vauss, J. W. Snell, N. Lange, and et al. Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. *Archives of General Psychiatry*, 53(7):607–616, 1996.

[13] F. X. Castellanos, D. S. Margulies, C. Kelly, L. Q. Uddin, M. Ghaffari, A. Kirsch, D. Shaw, Z. Shehzad, A. Di Martino, B. Biswal, and et al. Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 63(3):332–337, 2008.

[14] F. X. Castellanos and E. Proal. Large-scale brain systems in adhd: beyond the prefrontalstriatal model. *Trends in Cognitive Sciences*, 16(1):17 – 26, 2012. Special Issue: Cognition in Neuropsychiatric Disorders.

[15] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[16] C.-W. Chang, C.-C. Ho, and J.-H. Chen. Adhd classification by a texture analysis of anatomical brain mri data. *Frontiers in Systems Neuroscience*, 6(66), 2012.

[17] W. Cheng, X. Ji, J. Zhang, and J. Feng. Individual classification of adhd patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. *Frontiers in Systems Neuroscience*, 6(58), 2012.

[18] L. Clark, A. D. Blackwell, A. R. Aron, D. C. Turner, J. Dowson, T. W. Robbins, and B. J. Sahakian. Association between response inhibition and working memory in adult adhd: A link to right frontal cortex pathology? *Biological Psychiatry*, 61(12):1395 – 1401, 2007. Advances in the Neurobiology of ADHD.

[19] J. B. Colby, J. D. Rudie, J. A. Brown, P. K. Douglas, M. S. Cohen, and Z. Shehzad. Insights into multimodal imaging classification of adhd. *Frontiers in Systems Neuroscience*, 6(59), 2012.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[21] R. W. Cox. Afni: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162 – 173, 1996.

[22] R. C. Craddock, G. 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*, page 19141928, 2011.

[23] M. Craven, D. Dipasquo, D. Freitag, A. Mccallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. pages 509–516. AAAI Press, 1998.

[24] D. Dai, J. Wang, J. Hua, and H. He. Classification of adhd children through multimodal magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6(63), 2012.

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

[26] S. Dey, A. R. Rao, and M. Shah. Exploiting the brain's network structure in identifying adhd. *Frontiers in Systems Neuroscience*, 6(75), 2012.

[27] S. Diane and A. Victoria. Left inferior frontal gyrus is critical for response inhibition. 2008.

[28] S. G. Dickstein, K. Bannon, F. Xavier Castellanos, and M. P. Milham. The neural correlates of attention deficit hyperactivity disorder: an ale meta-analysis. *Journal of Child Psychology and Psychiatry*, 47(10):1051–1062, 2006.

[29] S. Durston. Differential patterns of striatal activation in young children with and without adhd. *Biological Psychiatry*, 53(10):871–878, 2003.

[30] A. Eloyan, J. Muschelli, M. B. Nebel, H. Liu, F. Han, T. Zhao, A. D. Barber, S. Joel, J. J. Pekar, S. H. Mostofsky, and B. Caffo. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6(61), 2012.

[31] Y. Fan, D. Shen, R. Gur, R. Gur, and C. Davatzikos. Compare: Classification of morphological patterns using adaptive regional elements. *Medical Imaging, IEEE Transactions on*, 26(1):93–105, Jan 2007.

[32] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524 – 531 vol. 2, June 2005.

[33] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[34] C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4):870 – 878, 2002.

[35] G. H. Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, 22(2):133 – 139, 2011. Functional Imaging.

[36] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, and S. C. Johnson. Spatially augmented {LPboosting} for {AD} classification with evaluations on the {ADNI} dataset. *NeuroImage*, 48(1):138 – 149, 2009.

[37] C. Hinrichs, V. Singh, G. Xu, and S. C. Johnson. Predictive markers for {AD} in a multi-modality framework: An analysis of {MCI} progression in the {ADNI} population. *NeuroImage*, 55(2):574 – 589, 2011.

[38] IMAGENET. Large scale visual recognition challenge 2012. http://www.image-net.org/challenges/LSVRC/2012/.

[39] T. Insel. Thomas insel on diagnostic and statistical manual of mental disorders. http://www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml.

[40] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. {FSL}. *NeuroImage*, 62(2):782 – 790, 2012.

[41] P. M. Jezzard, P. Matthews and S. M. Smith. *Functional MRI: An Introduction to Methods*. Oxford university press, Inc., New York, USA, 2001.

[42] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[43] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In C. Ndellec and C. Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998.

[44] S. Jouili and S. Tabbone. Graph matching based on node signatures. In A. Torsello, F. Escolano, and L. Brun, editors, *Graph-Based Representations in Pattern Recognition*, volume 5534 of *Lecture Notes in Computer Science*, pages 154–163. Springer Berlin Heidelberg, 2009.

[45] J. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[47] T. Leung and J. Malik. Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision*, 43:29–44, June 2001.

[48] J. Liu and M. Shah. Learning human actions via information maximization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, June 2008.

[49] A. Ma'ayan, G. A. Cecchi, J. Wagner, A. R. Rao, R. Iyengar, and G. Stolovitzky. Ordered cyclic motifs contribute to dynamic stability in biological and engineered networks. *Proceedings of the National Academy of Sciences*, 105(49):19235–19240, 2008.

[50] J. Mehnert, A. Akhrif, S. Telkemeyer, S. Rossi, C. H. Schmitz, J. Steinbrink, I. Wartenburger, H. Obrig, and S. Neufang. Developmental changes in brain activation and functional connectivity during response inhibition in the early childhood brain. *Brain and Development*, 35(10):894 – 904, 2013.

[51] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[52] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):pp. 32–38, 1957.

[53] NITRC. Adhd-200 data description, data processing and regions of interest list page. http://nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline.

[54] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[55] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996.

[56] E. Olivetti, S. Greiner, and P. Avesani. Adhd diagnosis from multiple data sources with batch effects. *Frontiers in Systems Neuroscience*, 6(70), 2012.

[57] S. Overmeyer, E. T. Bullmore, J. Suckling, A. Simmons, S. C. Williams, P. J. Santosh, and E. Taylor. Distributed grey and white matter deficits in hyperkinetic disorder: Mri evidence for anatomical abnormality in an attentional network. *Psychological Medicine*, 31(8):1425–1435, 2001.

[58] K. J. Plessen, R. Bansal, H. Zhu, R. Whiteman, J. Amat, G. A. Quackenbush, L. Martin, K. Durkin, C. Blair, J. Royal, K. Hugdahl, and B. S. Peterson. Hippocampus and amygdala morphology in attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry*, 63(7):795–807, 2006.

[59] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):676–682, 2001.

[60] A. Rao, R. Garg, and G. Cecchi. A spatio-temporal support vector machine searchlight for fmri analysis. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 1023–1026, 2011.

[61] S. A. Rombouts, J. S. Damoiseaux, R. Goekoop, F. Barkhof, P. Scheltens, S. M. Smith, and C. F. Beckmann. Model-free group analysis shows altered bold fmri networks in dementia. *Human Brain Mapping*, 30(1):256–266, 2009.

[62] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059 – 1069, 2010.

[63] J. R. Sato, M. Q. Hoexter, A. Fujita, and L. A. Rohde. Evaluation of pattern recognition and feature extraction methods in adhd prediction. *Frontiers in Systems Neuroscience*, 6(68), 2012.

[64] R. Schachar, G. Logan, P. Robaey, S. Chen, A. Ickowicz, and C. Barr. Restraint and cancellation: Multiple inhibition deficits in attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology*, 35(2):229–238, 2007.

[65] L. J. Seidman, E. M. Valera, N. Makris, M. C. Monuteaux, D. L. Boriel, K. Kelkar, D. N. Kennedy, V. S. Caviness, G. Bush, M. Aleardi, and et al. Dorsolateral prefrontal and anterior cingulate cortex volumetric abnormalities in adults with attention-deficit/hyperactivity disorder identified by magnetic resonance imaging. *Biological Psychiatry*, 60(10):1071–1080, 2006.

[66] D. J. Sharp, V. Bonnelle, X. De Boissezon, C. F. Beckmann, S. G. James, M. C. Patel, and M. A. Mehta. Distinct frontal systems for response inhibition, attentional capture, and error processing. *Proceedings of the National Academy of Sciences*, 107(13):6106–6111, 2010.

[67] A. Shukla and U. Kumar. Positron emission tomography: An overview. *Journal of Medical Physics*, 31(1):13 – 21, 2006.

[68] G. S. Sidhu, N. Asgarian, R. Greiner, and M. R. G. Brown. Kernel principal component analysis for dimensionality reduction in fmri-based diagnosis of adhd. *Frontiers in Systems Neuroscience*, 6(74), 2012.

[69] T. J. Silk, M. A. Bellgrove, P. Wrafter, J. B. Mattingley, and R. Cunnington. Spatial working memory and spatial attention rely on common neural processes in the intraparietal sulcus. *NeuroImage*, 53(2):718 – 724, 2010.

[70] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 2009.

[71] E. R. Sowell, P. M. Thompson, S. E. Welcome, A. L. Henkenius, A. W. Toga, and B. S. Peterson. Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder. *Lance*, 362(9397):1699–1707, 2003.

[72] H. Spath. *The Cluster Dissection and Analysis Theory FORTRAN Programs Examples*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1985.

[73] O. Sporns. Graph theory methods for the analysis of neural connectivity patterns.

[74] J. Steinbrink, A. Villringer, F. Kempf, D. Haux, S. Boden, and H. Obrig. Illuminating the bold signal: combined fmrifnirs studies. *Magnetic resonance imaging*, 24:495 – 505, 2006.

[75] M. H. Teicher, C. M. Anderson, A. Polcari, C. A. Glod, L. C. Maas, and P. F. Renshaw. Functional deficits in basal ganglia of children with attention-deficit/hyperactivity disorder shown with functional magnetic resonance imaging relaxometry. *Nature Medicine*, 6(4):470–473, 2000.

[76] L. Tian, T. Jiang, Y. Wang, Y. Zang, Y. He, M. Liang, M. Sui, Q. Cao, S. Hu, M. Peng, and et al. Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neuroscience Letters*, 400(1-2):39–43, 2006.

[77] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[78] 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, January 2002.

[79] L. Q. Uddin, A. Clare Kelly, B. B. Biswal, F. Xavier Castellanos, and M. P. Milham. Functional connectivity of default mode network components: Correlation, anticorrelation, and causality. *Human Brain Mapping*, 30(2):625–637, 2009.

[80] J. L. Vincent, I. Kahn, A. Z. Snyder, M. E. Raichle, and R. L. Buckner. Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *Journal of Neurophysiology*, 100(6):3328–3342, 2008.

[81] Y.-F. Zang, Y. He, C.-Z. Zhu, Q.-J. Cao, M.-Q. Sui, M. Liang, L.-X. Tian, T.-Z. Jiang, and Y.-F. Wang. Altered baseline brain activity in children with adhd revealed by resting-state functional mri. *Brain & Development*, 29(2):83–91, 2007.

[82] C.-Z. Zhu, Y.-F. Zang, Q.-J. Cao, C.-G. Yan, Y. He, T.-Z. Jiang, M.-Q. Sui, and Y.-F. Wang. Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *NeuroImage*, 40(1):110 – 120, 2008.