Using Functional or Structural Magnetic Resonance Images and Personal Characteristic Data to Diagnose ADHD and Autism

by

Sina Ghiassian

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

A clinical tool that can diagnose psychiatric illness using functional or structural magnetic resonance (MR) brain images would greatly assist physicians. Here, we propose a learning algorithm that uses the histogram of oriented gradients (HOG) features of MR brain images, as well as personal characteristic data, as features. We show that this learner can produce effective classifiers when run on two large public datasets. It is able to diagnose ADHD with hold-out accuracy of 0.696 (over baseline $= 0.550$) using personal characteristics and structural brain scan features when trained on the ADHD-200 global competition dataset and is also able to diagnose autism with hold-out accuracy of 0.650 (over baseline $= 0.516$) using functional images with personal characteristic data when trained on the Autism Brain Imaging Data Exchange (ABIDE) dataset. We also show that it is possible to diagnose ADHD and autism by using just structural brain images with accuracies of 0.661 and 0.601 respectively. Our imaging-based accuracy on the ADHD-200 dataset is about 8% higher than the best imaging-based accuracy in the ADHD-200 competition. While these results are not yet at the level of clinical relevance, they outperform all previously presented methods on both datasets. To our knowledge, this is the first demonstration of a single automated process being able to produce an effective diagnostic system for two different psychiatric illnesses (ADHD and autism). These results suggest that the learning approach using HOG features as input may produce diagnostic classifiers (from functional and/or structural brain images) that perform well for other psychiatric disorders.

This work is dedicated to my Mother, Father and Sister

Acknowledgements

First and foremost I would like to express my deepest appreciation to my supervisor, professor Russell Greiner for his endless patience. He is one of the most creative people I have ever met. He has supported me throughout my Master's and without him this dissertation would not have been completed. I am really grateful to my co-supervisor Dr. Matthew R. G. Brown who has always supported me with his insightful comments and suggestions throughout my program. I have really benefited from his feedback. Without his persistent help this dissertation would not have been possible. I am sure he was and will be the most friendly advisor I have had.

I should thank my mother, father and sister for their unconditional endless support from day one. They have supported me spiritually and financially throughout my life. I do not know how I could have done this without my mother's love, my father's generosity and my sister's care.

Last but not the least, I owe a very important debt to my friends in Edmonton. I did not have many friends but the ones I had, were always there for me.

This work was supported by the Alberta Innovates Centre For Machine Learning (AICML), Canadian Institutes of Health Research (CIHR), Alberta Innovates–Health Solutions (AIHS) and Natural Sciences and Engineering Research Council of Canada (NSERC).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Structural magnetic resonance imaging (MRI) allows for non-invasive acquisition of volumetric images of brain anatomy [Edelman and Warach, 1993]. Functional magnetic resonance imaging (fMRI) provides a measure of dynamic changes in brain activity over time, most commonly based on the blood oxygenation level dependant (BOLD) signal [Huettel et al., 2004a]. In task-based fMRI, the subject is asked to perform a task while in the scanner. Our work, however, focusses on resting-state fMRI (RS-fMRI), where the subject is asked to lie in the magnet and rest quietly [Lee et al., 2012].

There are different ways to use fMRI data to analyze the brain. Many researchers compare groups of people and/or compare brain states within the same individuals. For example, Wolf et al. [2009] applied independent component analysis (ICA) to a dataset of 12 healthy and 12 ADHD adults during a working memory task and observed that ADHD patients had significantly less activation in the left ventrolateral prefrontal cortex, cerebellar and occipital regions, compared with healthy controls. Culham et al. [2003] stated that, although both reaching and grasping require transporting the hand to the object location, grasping requires processing of object shape, orientation and size to preshape the hand. They used functional magnetic resonance imaging to determine whether grasping (compared to reaching) produced activation in dorsal areas, ventral areas, or both. They found greater activity for grasping than reaching in several regions including anterior intraparietal cortex. Such *association studies* aggregate over individuals in different groups. By contrast, our work focusses on machine learning classification studies, which use labeled datasets to produce a classifier that can then be used to classify novel individual people into different categories. Use of machine learning with structural/functional MRI data is now widespread and increasingly familiar to potential readers. In the case of diagnosis, this classifies each individual as being either healthy or falling into a disease category.

In the ADHD-200 global competition [ADHD-200 consortium, 2011a], many teams tried to learn a classifier that could determine if a subject is healthy or has ADHD, using RS-fMRI, MRI and personal characteristic data. Their results show that it is possible to classify the ADHD disorder using this data, with the caveat that it is not ready for clinical purposes since the accuracy is far from perfect (the best accuracy was 0.6154) [ADHD-200 consortium, 2011b]. Other researchers have tried to learn to classify autism, using the ABIDE dataset [ABIDE dataset, 2011, Di Martino et al., 2013]. They achieved an accuracy of 0.60 over a baseline of about 0.54. We will describe their results below and show that our methods can improve on their results.

This study describes our algorithms for learning to classify two specific kinds of psy-

chiatric disorders: ADHD and autism. One algorithm tries to find the differences in brain functional connectivities, which is defined as the temporal correlation of different brain regions over time [Friston et al., 1993]. For this purpose we use brain regions that are reported to act differently in autistic and healthy people using the Regional Homogeneity (ReHo) method [Zang et al., 2004] and resting-state functional MRI. This method measures the local synchronization of spontaneous fMRI signals. We explain our approach in detail in Chapter 4.

The dissertation also explains our MHCD learning algorithm, which learns an automated system for diagnosing ADHD from the ADHD-200 dataset or for diagnosing autism from the ABIDE dataset, using histogram of oriented gradient features (HOG, see Section 2.4 and Section 5.1) extracted from structural (or alternatively resting-state functional) MR images as input to different base learners. This algorithm makes 3 specific contributions: *a*) We show that HOG feature descriptors of either MRI or RS-fMRI data can be useful for diagnosing psychiatric disorders. *b*) Our method outperforms all previously published classification results on the two large resting-state fMRI/MRI datasets mentioned above. *c*) The effectiveness of our approach using just the MRI data suggests that the anatomy of the brain in people with these psychiatric illnesses are different from brains of healthy subjects.

The rest of this dissertation is organized as follows. Chapter 2 will define the problem and give the necessary background to make the dissertation easier to read and understand. Chapter 3 describes the preprocessing steps and challenges of preprocessing multisite datasets with large number of subjects and batch effects. In Chapter 4 we explain what we are exactly measuring over time in different brain regions and describe how we use the measure for a classification purpose. Chapter 5 provides a thorough explanation of our 3D HOG method and how it is used to learn classifiers for two large datasets. We then will present the results of our $3D$ HOG algorithm in Chapter 6 and at last we will discuss the results in Chapter 7.

Chapter 2

Background

This chapter contains the background material necessary to understand the rest of the dissertation. Section 2.1 describes the basics of functional magnetic resonance imaging and different experiment types. Section 2.2 formally defines the problem we want to solve. Section 2.3 gives information on the datasets we used to evaluate our approach and after that Section 2.4 will discuss the $2D$ histogram of oriented gradient features [Dalal and Triggs, 2005] of images, which is later expanded to the $3D$ version.

2.1 Functional Magnetic Resonance Images

Functional Magnetic Resonance Imaging also known as functional MRI (fMRI) is a neuroimaging technique that uses standard MRI scanners to investigate how the activity of the brain changes over a few minutes [Huettel et al., 2004b]. fMRI is often used to measure spontaneous low-frequency fluctuations in blood oxygen level dependant (BOLD) signal to investigate the functional architecture of the brain [Lee et al., 2012]. Figure 2.1 demonstrates a modern MRI scanner. The magnetic field of MRI scanners are typically at least 1.5 Tesla but nowadays, scanners with 3 Tesla magnetic fields are common [Huettel et al., 2004b]. While the person is inside the magnet, the instrument will scan the brain second-by-second. Scanning happens continuously, producing one snap-shot every 1 to 3 seconds. This sequence of scans over time produce a signal for each voxel of the brain (voxel is the smallest 3 dimensional unit of imaging – see Figure 2.2).

There are two different types of fMRI experiments: task-based and resting-state. In task-based fMRI, the subject is asked to do a task while in the scanner – e.g., pushing a button for certain (described) situations. In resting-state magnetic resonance imaging, the sub-

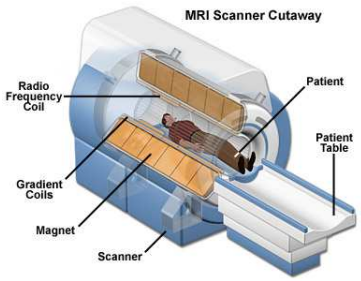


Figure 2.1: fMRI scanner; image from research.cs.queensu.ca

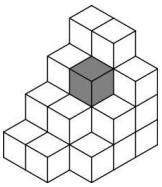


Figure 2.2: Volumetric pixel; image from en.wikipedia.org

ject is asked to lie and do nothing while in the magnet. Researchers have used resting-state functional magnetic resonance imaging (RS-fMRI) for different kinds of analysis [Konrad and Eickhoff, 2010]. Various methods could also be used for analyzing resting-state scans including independent component analysis, graph methods and neural networks [Lee et al., 2012].

2.2 Problem definition

Given a training dataset D that includes N training examples each with d features where each training example belongs to one of two possible classes (healthy or patient), the goal of learning is to find a function $C : X \rightarrow Y$ where X is the feature space and Y is the output domain – i.e., C maps each example to one of the possible classes.

$$X \subseteq \mathbb{R}^d, \quad Y = \{0, 1\}$$
$$D \subseteq (X \times Y)^N$$
$$L(D) \rightarrow C$$

In general, a learner L learns a classifier C based on the data available in dataset D.

Here we would like to learn classifiers for two specific multisite datasets, ADHD-200 competition [ADHD-200 consortium, 2011a] dataset and Autism Brain Imaging Data Exchange (ABIDE) dataset [ABIDE dataset, 2011, Di Martino et al., 2013]. In each case our learner will propose algorithms to do binary classification of ADHD versus healthy and autistic versus healthy.

2.3 Datasets

The ADHD-200 global competition *training* dataset includes resting-state fMRI and [T1-weighted] structural scans of 776 subjects from 7 different imaging sites, 491 of whom were "controls" (presumed healthy subjects) while 285 of the subjects were cases with ADHD. The ADHD-200 global competition *test* data included 197 instances including both healthy and ADHD cases from 6 different sites. We had to remove 33 of these subjects (26 from the test data and the rest from the training data) since 6 had no resting-state scan, 1 could not be preprocessed using our preprocessing pipeline (see Chapter 3) and 26 from the Brown University site as they had no diagnostic labels. We were left with 490 healthy controls and 279 ADHD cases in the training set and 94 healthy subjects and 77 ADHD cases in the test set. The ADHD-200 dataset also included other non-imaging features for each subject, including gender, age, handedness, site of the imaging, IQ measure and full4IQ score; see [ADHD-200 consortium, 2011a, Brown et al., 2012] for more details on these personal characteristics.

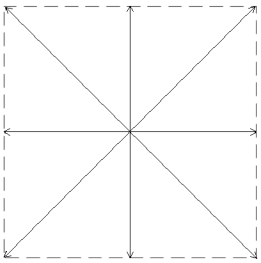


Figure 2.3: $2D$ bins

We also explored the Autism Brain Imaging Data Exchange(ABIDE) dataset, which included 1112 subjects. We had to remove one whose data could not be preprocessed with our preprocessing pipeline, leaving us with a dataset of 1111 individuals, including 573 healthy controls and 538 autistic cases. The ABIDE dataset also provided an extensive array of phenotypic information from which we used age, gender, handedness, full scale IQ, verbal IQ, performance IQ, site of the imaging and eyestat (which indicated whether the person kept his/her eyes open or not during the scan); for more information on these personal characteristics see [ABIDE dataset, 2011].

Each of the datasets included a [T1-weighted] structural scan (high resolution, for a single time point), and also 1 or more resting-state functional scans for each of the subjects. That functional scan included (for the ADHD-200) between 76 to 261 time points for each subject and between 82 and 320 time points for the ABIDE dataset. Different subjects were scanned with different temporal resolutions: ranging from $1.5(s)$ through $3(s)$ in the ADHD-200 dataset, and from $1(s)$ through $3(s)$ in the ABIDE data.

There are many challenges in processing such multisite datasets that are not present in datasets gathered in a single site, including site-specific range of intensity values, different scanning durations, different volume times and other batch effects. Our preprocessing pipeline (see Chapter 3) therefore included steps to normalize various aspects of the data across different sites. An advantage of using multi-site datasets with more diverse subjects is that as these datasets are often larger with a wider range of subjects, a system that does well here tends to be more reliable and robust versus systems honed to work on only a single site [Leek et al., 2010, Johnson et al., 2007]. Nielsen et al. [2013] observed this in their results on the ABIDE dataset. They have high accuracy values (e.g. more than 75%) when they learned on data from a single site, then classified on test data from that site. However, their best accuracy value was only 60% when they tried to classify autism versus healthy based on the whole dataset, over many sites. Although their accuracy values were lower when using the whole dataset compared to when they used single sites for classification, the results of using the whole dataset are expected to have better generalization ability. We will discuss their results in more detail later.

2.4 Histogram of Oriented Gradients

Here we provide a summary of what HOG features are and how they are extracted from $2D$ images. This will help later to understand how we extend this to $3D$ space. The idea behind the histogram of oriented gradient descriptors is that the intensity gradients' distribution can describe the object appearance [Dalal and Triggs, 2005]. The input of the HOG algorithm is an image along with the size of each cell (how many pixels/voxels should a cell contain), the size of the blocks (how many cells should a block contain) and the number of bins for each cell (see Figure 2.3). The output is a histogram over the specified number of bins for each cell of the image.

The HOG feature extraction algorithm divides the whole image into blocks and each block into several cells, where each cell involves a set of pixels (note that a cell can belong to multiple blocks). In each cell, it utilizes the histograms of gradient orientations as a new feature. It then normalizes each cell within different blocks. This can be done by considering overlapping blocks through the image (see Dalal and Triggs [2005] for a thorough explanation). Each cell then contributes to the final feature vector a few times, normalized within different blocks.

Given the function $f(x, y)$ that maps a position (x, y) in each $2D$ image to a pixel value, the derivative is:

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \end{bmatrix}$$

where:

$$f_x(x,y) = \frac{\partial f(x,y)}{\partial x} \approx \frac{f(x+1,y) - f(x-1,y)}{2}$$
$$f_y(x,y) = \frac{\partial f(x,y)}{\partial y} \approx \frac{f(x,y+1) - f(x,y-1)}{2}$$

The gradient magnitude is:

$$|\nabla f(x,y)| = \sqrt{f_x^2 + f_y^2}$$

Then we found the bin with the maximum overlap with the gradient vector as follows:

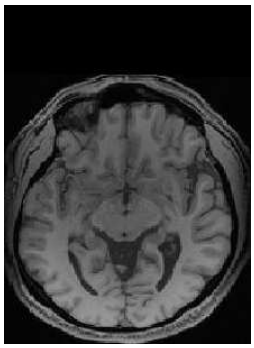
$$\alpha(\nabla f(x,y), b) = \frac{\nabla f(x,y) \cdot b}{|\nabla f(x,y)| \times |b|}$$
$$\text{dir}(x,y) = \underset{b}{\text{argmax}}\ \alpha(\nabla f(x,y), b)$$

where b ranges over each of the vectors in Figure 2.3. For each pixel at location (x, y), HOG first computes a direction $\text{dir}(x, y)$, and then increments the associated bin:

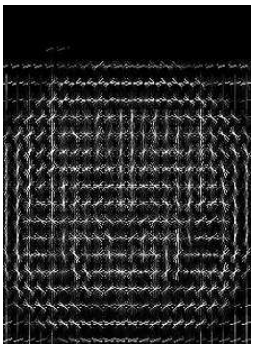
$$\text{bin}_{\text{dir}(x,y)} \mathrel{+}= |\nabla f(x,y)|$$

The idea is illustrated in Figure 2.5. See Figure 2.4 for the representation of HOG features extracted from a brain image slice.

For each cell, HOG builds a histogram, and then concatenates the cell histograms in a block for normalization – here, this produces a vector of length $i \times j$ (i cells in a block and j histogram bins for each cell) values. Here, let v represent the histogram, viewed as the



(a) An axial slice of a brain



(b) HOG features of the same slice

Figure 2.4: Input and output of $2D$ HOG on a brain image;
Here, we represent the HOG features by an 8-sided "star", where the length of each arm is the size of the histogram, in that direction. This representation is generated using the vl-feat library [Vedaldi and Fulkerson, 2008].

tuple of values in a block. One of the successful normalization schemes used in [Dalal and Triggs, 2005] is:

$$v \rightarrow \frac{v}{\sqrt{||v||_2^2 + \epsilon^2}}$$

where $||v||_2$ is the 2-norm and ϵ is a small constant, which helps in cases where all the gradient vector's magnitudes are equal to zero in a block. A thorough explanation of different block normalizations schemes can be found in the original paper; [Dalal and Triggs, 2005].

HOG has been successfully applied to $2D$ images for various tasks related to object recognition [Dalal and Triggs, 2005]. Our question was if this successful method can detect any differences in healthy control brains and non-healthy brains to diagnose different diseases.

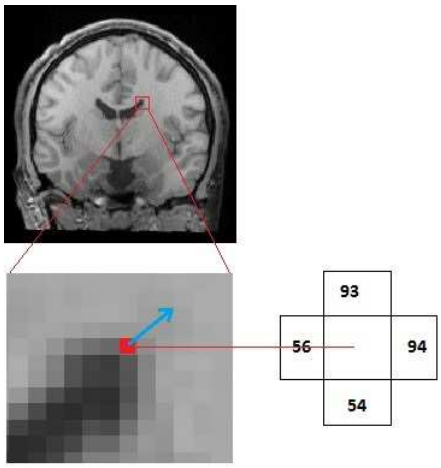


Figure 2.5: Gradient vector of a sample pixel

Here, we consider a single pixel – the one shown in red, whose neighbors have intensities 56, 93, 94 and 55. The blue arrow is the sample gradient, computed as follows:

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{94-56}{2} \\ \frac{93-54}{2} \end{bmatrix} = \begin{bmatrix} 19 \\ 19.5 \end{bmatrix}$$

$$\text{Magnitude} = \sqrt{(19)^2 + (19.5)^2} = 27.22$$

Then to find the maximum overlapping bin we compute:

$$\text{dir}(x,y) = \underset{\text{b}}{\text{argmax}}\ \alpha(\nabla f(x,y), b)$$

where b ranges over the angles shown in Figure 2.3). The biggest α is produced when $b^* = (1,1)$:

$$\alpha(\nabla f(x,y), b^*) = \frac{\nabla f(x,y) \cdot b}{|\nabla f(x,y)| \times |b|} = \frac{(19, 19.5) \cdot (1,1)}{27.22 \times \sqrt{2}} \approx 1$$

$$\text{dir}(x,y) = (1,1)$$

$$\text{bin}_{(1,1)} \mathrel{+}= 27.22$$

Chapter 3

Preprocessing

In fMRI, small but meaningful changes are buried within highly variable measurements. The measured BOLD signal change is very small when compared with the total intensity of the MR signal [Huettel et al., 2004c]. Due to issues like the movements of subject in the scanner during the scan, it is necessary to preprocess fMRI data before analyses. Computational procedures that are applied to fMRI data following image reconstruction but before statistical analysis are called preprocessing [Huettel et al., 2004c]. Preprocessing steps are meant to reduce the variability of data that is not associated with the experimental task, and to prepare the data for statistical testing [Huettel et al., 2004c]. The other big issue with fMRI data is the huge number of dimensions, which presents lots of challenges in analyzing fMRI data. There are several techniques for reducing the number of dimensions in fMRI data including focusing on certain regions of interest and averaging. We will use some of these techniques in our preprocessing pipeline.

Most of the times it is useful to preprocess structural MRI data as well. However, preprocessing MRI data needs fewer steps since it is a single time point. A summary of our preprocessing pipeline for fMRI and also MRI scans can be found in Figure 3.1.

In our preprocessing pipeline, we used standard methods from the structural MRI and fMRI literature. These include motion correction, co-registration, spatial normalization and spatial smoothing. We added some steps to the preprocessing pipeline that are expected to be useful for our analysis. Our preprocessing steps are meant to be useful for increasing the classification accuracy or dimension reduction. In Section 3.2 through Section 3.9 we describe our preprocessing pipeline and the challenges we faced in the aforementioned multisite datasets. For preprocessing, we used SPM8, a software package designed for analyzing brain imaging data and also for preprocessing fMRI data [SPM, 2009, Friston et al., 1994], and our own in-house MATLAB code. We had to use our own MATLAB code for two reasons:

- Some of the steps of our preprocessing pipeline like the temporal preprocessing step are not provided in SPM8.
- We had to write code for SPM8 to do batch preprocessing. SPM8's graphical user interface could only preprocess a single image.

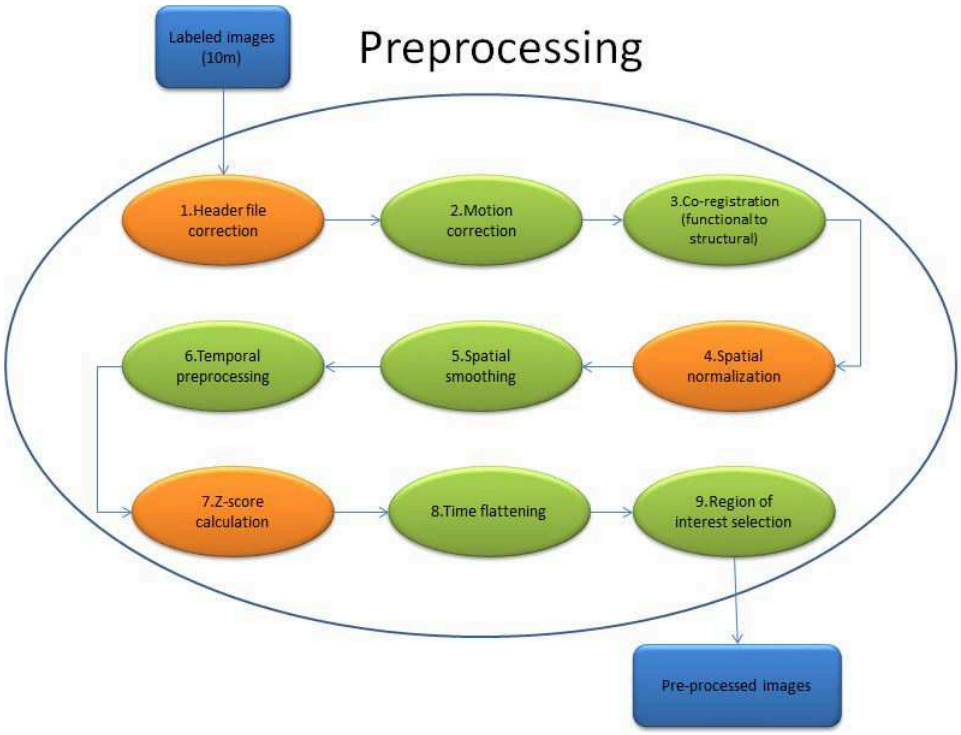


Figure 3.1: Preprocessing pipeline
The preprocessing pipeline for structural and functional magnetic resonance images is summarized in the figure. Orange shapes in the image show the steps of preprocessing necessary for both fMRI and MRI scans. Green shapes show the preprocessing steps only needed for fMRI scans.

3.1 Header file correction

Some critical information related to the images is stored in the header file of the scan including the orientation of the image and the image origin that is stored in a matrix named *sform*. In the spatial normalization step of preprocessing, we map each scan in our dataset to an atlas (see Section 3.4). For the mapping procedure, the search starts from the origin of the images. In the atlas we used for spatial normalization, the origin of the image is close to the anterior commissure (see Figure 3.2). If the origin of the image to be normalized is close to the anterior commissure and the image and the atlas have the same orientation, the spatial normalization step can easily find the global optimum and normalize the image. We reset the sform matrix so that the origin of the image is close to the anterior commissure and to make sure that the image has the same orientation (see Figure 3.3) as the atlas.

In the ABIDE dataset, the origin of the scans that came from Carnegie Mellon University (*CMU*) and University of Michigan (*UM*) sites, were set pretty badly so header file correction was necessary before continuing the preprocessing. 6% of the subjects from the *UM* site had orientation problems.

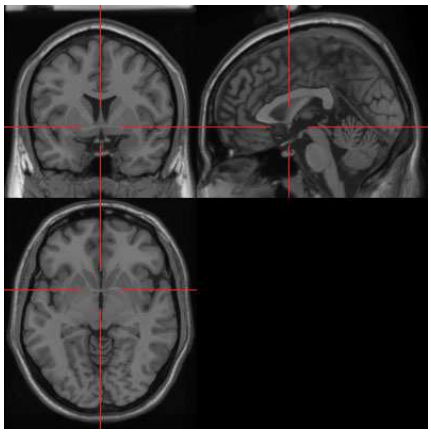
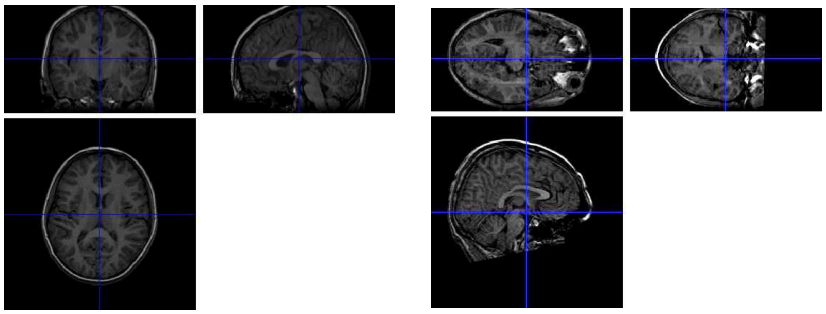
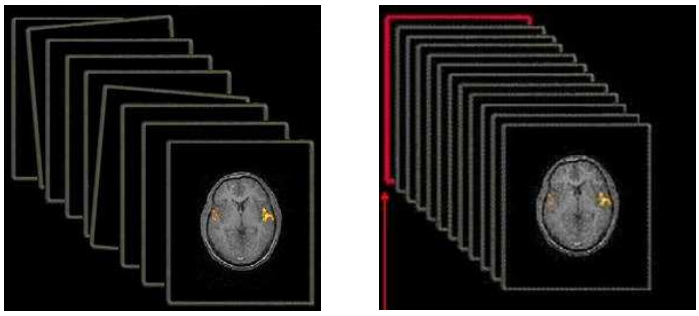


Figure 3.2: Anterior commissure



(a) Correct orientation (b) Wrong orientation

Figure 3.3: Image orientation



(a) Before motion correction (b) After motion correction

Figure 3.4: Different slices of a brain volume before and after motion correction.
All of the images are aligned with the first image (shown in red); images from fmri-easy.de

3.2 Motion correction

According to Huettel et al. [2004c] , head motion is the most damaging problem for fMRI analysis. A little head motion can make the data meaningless and unusable [Huettel et al., 2004c]. It can cause signal changes over time and it can also cause signal loss on the edges of the brain. There are some ways to avoid these problems, like prevention of head motion in the magnet, but we could only work with the data we had, which included this problem. We used the head motion correction of SPM8 to address this problem. The general process of spatially aligning two image volumes is called co-registration [Huettel et al., 2004c]. The goal of motion correction is to adjust the series of images so that the brain is always in the same position [Huettel et al., 2004c]. Motion correction is basically co-registering all of the brain volumes in a scan with the first (or another) subject-matched functional scan (see Figure 3.4). SPM8 does 6-parameter motion correction to align the images (see Figure 3.5).

3.3 Co-registration (functional to structural)

After motion correction, we ran functional-structural co-registration that is the process of mapping the functional and subject-matched structural images to each other. There are remarkable differences in functional and subject-matched structural images. Because the structural images have distinct boundaries and are higher resolution (see Figure 3.6), it is beneficial to use information available in structural scan for guiding the spatial normalization (next step) of the functional image [Huettel et al., 2004c]. We will discuss the details in Section 3.4.

Basically there are two image groups when doing coregistration.

1. **Reference image:** This image remains stationary (sometimes known as the target or template image).

2. **Source image:** This image is repositioned to best match the reference.

This brings researchers two options for doing the coregistration, either choosing the functional image as the reference image and the structural image as the source image or vice versa. We used the functional image as the source image and the structural as the reference image.

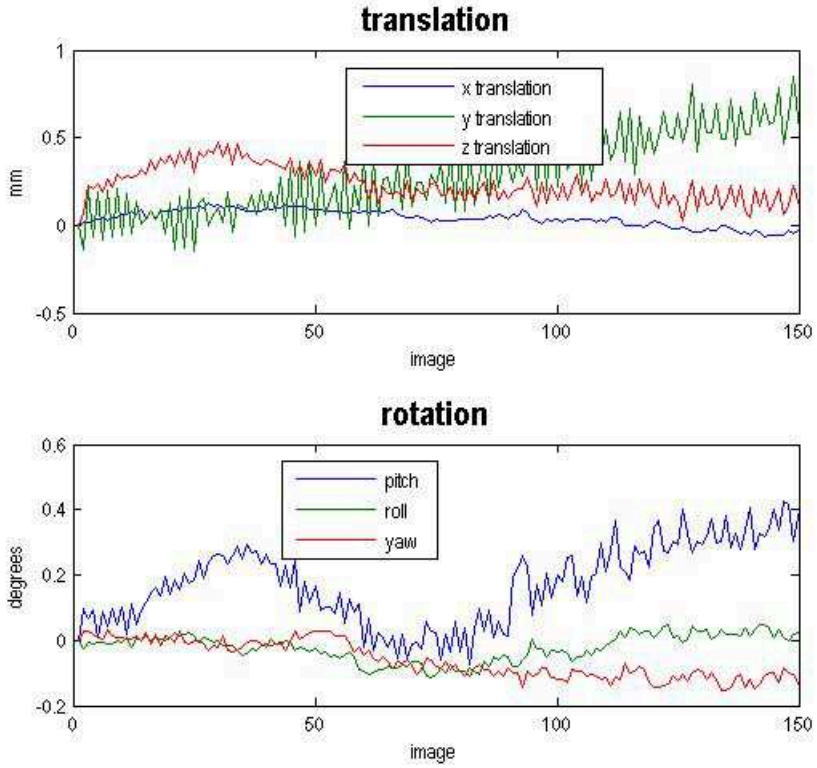
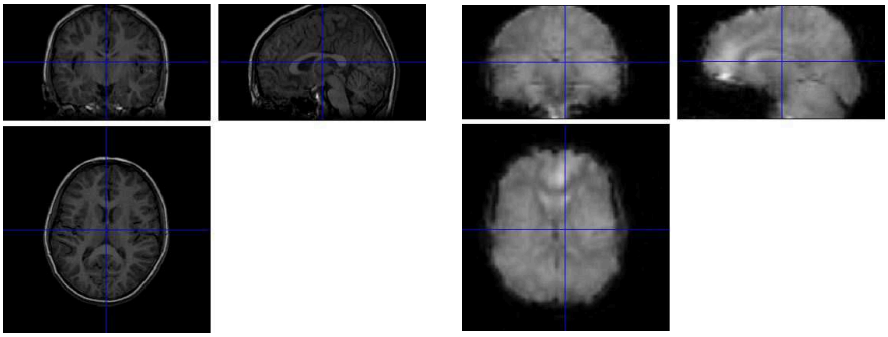
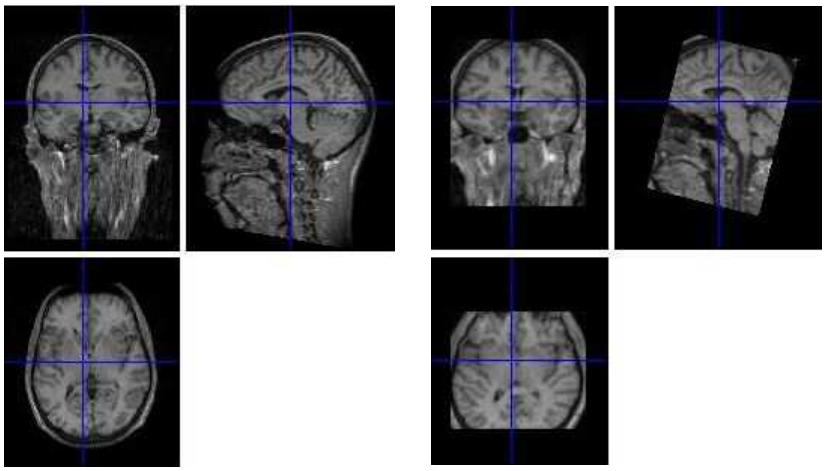


Figure 3.5: Movements of a sample subject



(a) Structural image (b) Functional image

Figure 3.6: Difference between the functional and the structural image



(a) Before (b) After, same slice

Figure 3.7: Before and after spatial normalization

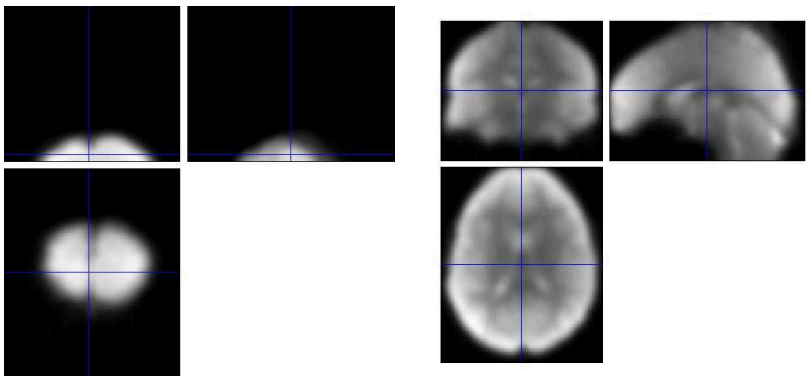
3.4 Spatial normalization

There are some differences in brain shape and also the location of brain structures like cortical gyri and sulci. The size of the adult brain is approximately in the range of $1100cc$ to $1500cc$, thus two subjects in the same brain imaging experiment may differ in their brain size by 30% or even more [Huettel et al., 2004c]. The spatial normalization step changes the images by mathematically warping, stretching and squeezing to match a template image (see Figure 3.8). After co-registering the structural and functional scans, we did non-linear spatial warping of each subject's functional volumes to the Montreal Neurological Institute (MNI) T1 template [Collins et al., 1999, Fonov et al., 2009, 2011].

There are two common atlases that might be used for normalizing the images: Talairach [Talairach and Tournoux, 1988] and MNI. Talairach atlas uses the brain of a 60-year-old French female whose brain was smaller than average. This means that if we want to match other brains to this atlas we should use lots of warps so the new image can fit the small size of the atlas. However the MNI atlas is not based on a single subject, but it is based on the average brain of many subjects. We have used the MNI atlas in our spatial normalization step. We used a bounding box of $[-78, -112, -50]$ to $[78, 76, 85]$, which are the SPM8 defaults, and a voxel size of 2 by 2 by 2 millimeters. The bounding box specifies the size of the volume (in millimeters) from the origin along the (x, y, z) directions.

According to our experiments, spatial normalization is the step most vulnerable to errors. It causes errors like losing some parts of the image, or the whole image (see Figure 3.8a for an example of a spatial normalization error). One important issue is that SPM8 can finish the normalization step without any errors or warnings while the image is not correctly matched with the atlas. One solution to avoid these problems is to make sure the origin of the image is set to the anterior commissure if the Talairach or MNI atlas is being used. The origin of the image in the Talairach atlas is the anterior commissure and the origin of the MNI atlas is very close to the anterior commissure. Although this might not be sufficient to solve the problem. It is good practice to check if the images look fine after this step of the preprocessing.

In Figure 3.8a the spatial normalization step completed without any errors but as can be



(a) Erroneous normalization (b) Valid normalization

Figure 3.8: Spatial normalization error

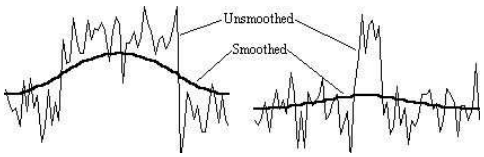


Figure 3.9: Applying a gaussian filter to two different signals; image from users.fmrib.ox.ac.uk

observed, the image is not in register with the atlas i.e., most of the image is lost.

3.5 Spatial smoothing

The strength of signal divided by other sources of variability in data is called the "signal to noise ratio" (SNR) [Huettel et al., 2004c]. In the spatial smoothing step, the data points are averaged with their neighbours to increase the signal to noise ratio. Huettel et al. [2004c] defines spatial smoothing as "the blurring of fMRI data across adjacent voxels". The most common way of spatial smoothing is using convolution with a Gaussian filter (see Figure 3.9). The width of the filter defines how many neighbour voxels it will affect: A narrow filter only affects a few neighbouring voxels while a wide one affects more [Huettel et al., 2004c]. The width of the filter is defined in millimeters at half of the maximum (see Figure 3.10) value (FWHM) [Huettel et al., 2004c]. Bigger values of FWHM results in higher SNR but will reduce the resolution of the images (see Figure 3.11), so we should find a balance to improve the SNR and maintain the resolution of the functional image.

For the ABIDE and the ADHD-200 dataset we have used $8mm$ FWHM, which is the default value in SPM8 and in our experience $8mm$ FWHM works well.

3.6 Temporal preprocessing

Although the images all had the same spatial dimension after the mentioned preprocessing steps, they varied in duration and volume time depending on the participant and the site of imaging. Temporal resolution or volume time is the duration it takes to scan the whole brain volume once – i.e., one time point. Although the images can be collected in either

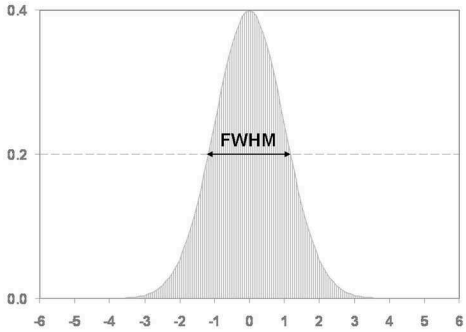


Figure 3.10: Full width half maximum (FWHM); image from support.brainvoyager.com

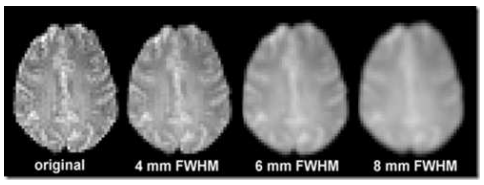


Figure 3.11: Result of using different FWHMs; image from support.brainvoyager.com

very short or long times, there are limitations of temporal resolution in functional MRI [Kim et al., 1997].

In the temporal preprocessing step, we linearly interpolated the images to have the same volume time for all of the subjects and then we chose the images with the smallest duration and we truncated all of the volumes after that duration for all other images in the dataset. After this step all of the images had the same temporal duration and volume time.

In the ADHD-200 dataset and the ABIDE dataset both, we chose a volume time of 2 seconds since many of the subjects in the datasets from different sites were scanned with the aforementioned volume time and did not need any interpolation.

3.7 Z-score calculation

There is another possible issue in a multi-site fMRI dataset that should be addressed before analyzing the data. Since the scans are gathered using different instruments and calibration settings, they will most probably have different ranges of intensity values. As an example the range might be between 20 and 50 for one site and between 1 and 1500 for another site. In the z-normalization step, we computed the mean and the standard deviation over all voxel values in an image (all voxel values for a specific patient, over all time points for fMRI data) and then subtracted the mean of the image from each voxel's value and then divided the resulting value by the standard deviation.

3.8 Time flattening

An fMRI scan is four dimensional with three spatial dimensions and one time dimension: $I(x, y, z, t)$ is the intensity value of the voxel at location (x, y, z) and at time t. We reduced the number of dimensions in the fMRI scans to 3 by setting the value for each (x, y, z)

location to the average value across time:

$$I(x, y, z) = \frac{1}{k} \sum_{t=1}^{k} I(x, y, z, t)$$

where k is the number of time points during the scan. This produced what we call a $3D$ functional MR image for each individual. (We also repeated our experiments using the median values of the voxels over time, rather than mean, and found similar results).

This step of the preprocessing is not used before the analysis in Chapter 4 since the time dimension is necessary for computing the functional connectivity.

3.9 Region of interest selection

An effective and popular method for reducing the dimensionality of the (f)MRI scans by a high amount is to select some regions of interest (ROI) and focus on those regions for analyses. It is also common to analyze ROIs to find different characteristics of specific parts of the brain and their roles in a specific disease.

We used this preprocessing step before the analysis of the ABIDE dataset in Chapter 4. However, the analysis performed in Chapter 5 and the experimental results in Chapter 6 did not focus on some sub-volumes, but instead used the whole structural/functional MRI volume.

Chapter 4

Correlation analysis and classification

In this chapter we describe our attempt to use the ABIDE dataset and the regions of interest (that are reported to act differently in healthy controls and autistic patients [Zang et al., 2004]) to learn classifiers for autism. Below, Section 4.1 explains why we thought this approach might be useful for diagnosing autism. Section 4.2 explains how we extracted the brain regions based on previous literature. Section 4.3 describes how we learned classifiers based on functional brain connectivities and our regions of interest. At the end of the chapter, Section 4.4 describes why we think this method did not lead to high classification accuracies.

4.1 Why should this approach work?

Paakki et al. [2010] claim that 15 regions in healthy and autistics act differently stating that: "The hypothesis was that the regional homogeneity (ReHo) of resting-state brain activity would be different (in resting state fMRI) between Autism Spectrum Disorder (ASD) subjects and controls in brain areas previously shown to display functional alternations in stimulus or task based fMRI studies". Since the regions have shown differences in task-based fMRI and later shown differences in resting-state fMRI using the ReHo method, we assumed that they might demonstrate differences in functional connectivities. If they do, they might then be useful for diagnosing healthy versus autistic.

4.2 Brain masking

We used the regions defined by Paakki et al. [2010] who used the ReHo method [Zang et al., 2004], to uncover some differences in autistic and healthy brains. The regional homogeneity method uses the Kendall's Coefficient Concordance (KCC) to measure the similarity of the time series of a given voxel to those of its nearest neighbors in a voxel-wise way [Zang et al., 2004]. They have presented 15 brain areas where the autistic subjects had significantly decreased or increased ReHo. Among these 15 regions we had to omit 4 because they were located down in the cerebellum that was out of the imaging bounds in many of the functional and structural images of the ABIDE dataset.

Paakki et al. [2010] used the Talairach atlas in their analysis. Since we used the MNI atlas, we converted the reported coordinates to the MNI space [Radua et al., 2012]. The Talairach and the converted coordinates are listed in Table 4.1.

X_T	Y_T	Z_T	X_{MNI}	Y_{MNI}	Z_{MNI}	Cluster size(Volume, mm^3)	Area
57	-37	6	63	-37	5	1632	MTG, STG, STS
37	43	6	41	48	-3	984	MFG, IFG
41	1	0	45	3	-6	472	Insula
45	-19	50	50	-13	52	432	Postcentral Gyrus
19	-19	6	22	-18	4	552	Thalamus
-25	31	-10	-26	33	-19	400	IFG, SG
-31	-67	8	-32	-69	12	1752	CC, OR, MOG, IOG, FG
37	-61	10	41	-62	12	1728	OR, MTG
25	-49	20	29	-48	22	1240	CC
-15	-83	16	-15	-85	22	984	OR, SOG
-13	13	26	-13	18	23	728	CC, MCC

Table 4.1: Voxel coordinates
CC=Corpus Callosum, FG=Fusiform Gyrus, IFG=Inferior Frontal Gyrus, IOG=Inferior Occipital Gyrus, ISLL=Inferior Semi-Lunar Lobule, MCC=Middle Cingulate Cortex, MFG=Middle Frontal Gyrus, MOG=Middle Occipital Gyrus, MTG=Middle Temporal Gyrus, OR=Optic Radiation, SG=Subcallosal Gyrus, SOG=Superior Occipital Gyrus, STG=Superior Temporal Gyrus, STS=Superior Temporal Sulcus. X_T, Y_T and Z_T show the Talairach coordinates and X_{MNI}, Y_{MNI} and Z_{MNI} show the corresponding MNI coordinates.

For extracting the regions of interest we used marsbar [Brett et al., 2002], a software that provides routines for region of interest analysis. After extracting the ROIs, we then computed the mean value $m^i_{t,r}$ of each region and used this value as the representative for a specific subject i, specific region r and a specific time point t. That is let $I^i_{t,v}$ represent the intensity value of voxel v at time t for subject i,

$$m^i_{t,r} = \frac{1}{p} \sum_{v \in r} I^i_{t,v}$$

where p is the number of voxels in region r.

4.3 Learning classifiers using functional brain connectivity

Brain functional connectivity is defined as the temporal correlation of a neurophysiological index measured in different brain areas [Friston et al., 1993]. If the image has n time points, let m^i_r represent the vector including all of the mean values of region r for subject i over all time points:

$$m^i_r = [m^i_{1,r}, m^i_{2,r}, \cdots, m^i_{n,r}]$$

The functional connectivity is:

$$corr(m^i_{r_j}, m^i_{r_k})$$

where r_j represents region of interest j. In our case, the blood oxygen level dependent (BOLD) signal is measured. As mentioned before (see Section 4.2) we extract 11 regions from each subject's scan and compute the mean value of that region in each time point. This forms the matrix M^i that includes the mean values for all 11 regions of interest and over all 74 time points for each subject, i:

$$M^i = \begin{pmatrix} m^i_{1,1} & m^i_{1,2} & \cdots & m^i_{1,11} \\ m^i_{2,1} & m^i_{2,2} & \cdots & m^i_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ m^i_{74,1} & m^i_{74,2} & \cdots & m^i_{74,11} \end{pmatrix}$$

Then we compute the functional connectivity of each possible two-pair of regions. The result is a matrix, C^i, that is of size 11×11 for each subject i. The matrix is symmetric since correlation(x, y) = correlation(y, x) (correlation is symmetric) and the diagonal of the matrix is always equal to 1 since correlation$(x, x) = 1$, so the matrix includes 55 ($\frac{11\times10}{2}$) useful values:

$$C^i = \begin{pmatrix} 1 & c^i_{1,2} & \cdots & c^i_{1,11} \\ c^i_{2,1} & 1 & \cdots & c^i_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ c^i_{11,1} & c^i_{11,2} & \cdots & 1 \end{pmatrix}$$

One simple approach is to try to build a classifier based on C^is. We put all 55 values for all healthy and autistic subjects in a matrix and used them to learn a classifier. If we have n healthy controls and m patients in D_{train} (we call the training set D_{train} and the test set D_{test}), X and Y matrices that we learn classifiers from, look like this:

$$X = \begin{pmatrix} C^1 \\ C^2 \\ \vdots \\ C^n \\ C^{n+1} \\ C^{n+2} \\ \vdots \\ C^{n+m} \end{pmatrix} \qquad Y = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix}$$

We used support vector machines with linear kernel and radial basis function kernel with different dimensionality reduction algorithms including principle component analysis, but the 5-fold cross validation results were never significantly better than the baseline.

We also tried another approach. We build what we call the *global mean matrix* H, for healthy controls. Here we only use D_{train}. This matrix includes all the mean values of all the regions of interest for all healthy subjects in D_{train}. H is basically all M^is of healthy controls in the training set concatenated vertically:

$$H = \begin{pmatrix} m^1_{1,1} & m^1_{1,2} & \cdots & m^1_{1,11} \\ m^1_{2,1} & m^1_{2,2} & \cdots & m^1_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ m^1_{74,1} & m^1_{74,2} & \cdots & m^1_{74,11} \\ m^2_{1,1} & m^2_{1,2} & \cdots & m^2_{1,11} \\ m^2_{2,1} & m^2_{2,2} & \cdots & m^2_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ m^2_{74,1} & m^2_{74,2} & \cdots & m^2_{74,11} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ m^n_{1,1} & m^n_{1,2} & \cdots & m^n_{1,11} \\ m^n_{2,1} & m^n_{2,2} & \cdots & m^n_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ m^n_{74,1} & m^n_{74,2} & \cdots & m^n_{74,11} \end{pmatrix} = \begin{pmatrix} M^1 \\ M^2 \\ \vdots \\ M^n \end{pmatrix}$$

We also build another matrix called S, which is the same as H but for non-healthy (sick) subjects.

Based on H and S, we build two global correlation matrices C^H (healthy control global correlation matrix) and C^S (non-healthy global correlation matrix). These matrices are built the same way we built C^is but here, we use the mean values of all healthy/non-healthy subjects' regions of interest (H and S), so they have the same number of distinct values; 55. Let $DM^{i,H}$ represent what we call the healthy difference matrix for each subject, which is basically the matrix that is the result of subtracting the global healthy correlation matrix from the subjects correlation matrix:

$$DM^{i,H} = C^i - C^H$$

and the non-healthy difference matrix is:

$$DM^{i,S} = C^i - C^S$$

and the final difference matrix (DM^i) for each subject is the concatenation of $DM^{i,H}$ and $DM^{i,S}$ after they are each reshaped into a 1×55 row vector. So the dimension of (DM^i) for each subject would be 1×110. We compute this matrix for every subject in the dataset. We assume that the DM^is that are derived from healthy controls in the training set are our *healthy* features and the DM^is that are derived from the non-healthy subjects are *non-healthy* features. If we have n healthy controls and m patients in $\mathrm{D}_{\text{train}}$, X and Y matrices that we learn classifiers from, look like this:

$$X = \begin{pmatrix} DM^1 \\ DM^2 \\ \vdots \\ DM^n \\ DM^{n+1} \\ DM^{n+2} \\ \vdots \\ DM^{n+m} \end{pmatrix} \qquad Y = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix}$$

We used a 5-fold cross validation scheme using different base learners including support vector machines with different kernel functions [Hastie et al., 2009] and different methods to reduce the dimensions and select the most useful features but the accuracy of our classifiers in cross validation was never significantly better than chance.

4.4 No good results, why?

There are 3 possible reasons why this approach did not lead to high classification accuracies:

1. Group level analysis and individual level analysis are different tools. As its name implies, in group level analysis researchers try to find differences in groups of people. However, in individual level analysis, the goal is to label a single subject as healthy or patient. Although the regions show differences in the groups of autistic and healthy, these differences might not be significant enough to be detected in individuals. Finding regions that are different in group level analysis suggests that these regions *might* be useful for classification purposes but it does not guarantee any above-chance diagnosis rates.

2. The second reason that might have caused this approach to not work is that, although these regions have different ReHos, they might not have different functional connectivities at all. ReHo and functional connectivity are different measures and differences in ReHo does not imply differences in functional connectivity. If these regions do not have different functional connectivities, the subjects cannot be classified using our functional connectivity method.

3. Another reason might be our significant reduction of the number of dimensions. Note that the number of dimensions for each subject in the ABIDE dataset was about $79 \times 95 \times 68 \times 200 \approx 100,000,000$ initially, which we reduced to $74 \times 11 = 814$ (74 time points and 11 regions of interest). We computed the mean value of each region for each person and we used this value as the representative for that specific region for a specific time point. Using this dimension reduction method, we might lose information that is necessary for the diagnosis task.

Chapter 5

3D HOG and classification

In this chapter we explain how we used preprocessed $3D$ functional or structural images and/or personal characteristic data for learning classifiers on different datasets. Below Section 5.1 describes how we expanded $2D$ HOG algorithm to the 3 dimensional space[1]. Section 5.2 explains some of the properties of HOG features. After that, Section 5.3 goes through our learning algorithm called MHCD and how it produces classifiers in detail. This learning algorithm is summarized in Figure 5.1.

5.1 3D HOG

In $3D$ HOG, we used the same ideas that were used in the $2D$ space (see Section 2.4). Given the function $f(x, y, z)$, that maps a position in a $3D$ image to its intensity value, we compute the derivative:

$$\nabla f(x,y,z) = \begin{bmatrix} \frac{\partial f(x,y,z)}{\partial x} \\ \frac{\partial f(x,y,z)}{\partial y} \\ \frac{\partial f(x,y,z)}{\partial z} \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix}$$

where as in Section 2.4:

$$f_x(x,y,z) = \frac{\partial f(x,y,z)}{\partial x} \approx \frac{f(x+1,y,z) - f(x-1,y,z)}{2}$$
$$f_y(x,y,z) = \frac{\partial f(x,y,z)}{\partial y} \approx \frac{f(x,y+1,z) - f(x,y-1,z)}{2}$$
$$f_y(x,y,z) = \frac{\partial f(x,y,z)}{\partial z} \approx \frac{f(x,y,z+1) - f(x,y,z-1)}{2}$$

The gradient magnitude is:

$$|\nabla f(x,y,z)| = \sqrt{f_x^2 + f_y^2 + f_z^2}$$

Then we found the bin with the maximum overlap with the gradient vector as follows:

$$\alpha(\nabla f(x,y,z), b) = \frac{\nabla f(x,y,z) \cdot b}{|\nabla f(x,y,z)| \times |b|}$$

[1] Many parts of this feature extraction method were developed by Ping Jin.

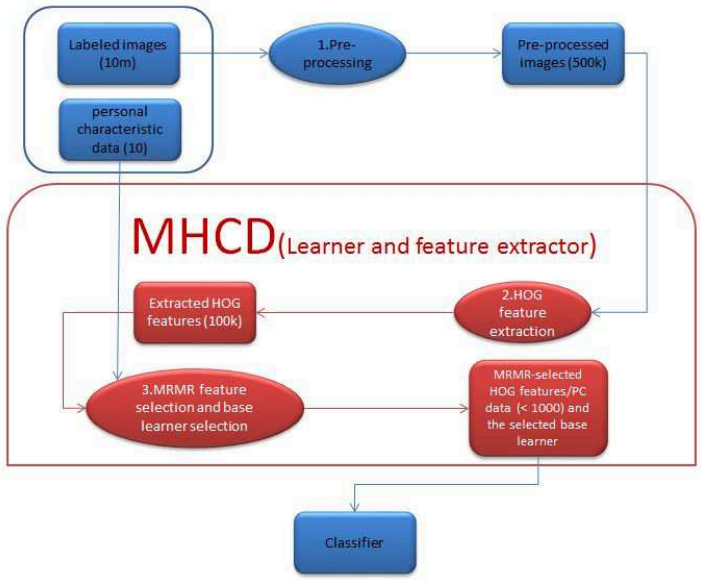


Figure 5.1: A summary of the learning pipeline

1) Each image in the datasets is preprocessed; see Chapter 3, especially Figure 3.1, reducing the dimensions from about $100,000,000$ ($79 \times 95 \times 68 \times 200$) to about $500,000$. 2) The MHCD system then extracts the $3D$ HOG features of each image, reducing the number of dimensions to about $100,000$; see Section 5.1. 3) The last step tries to select the best learner (from the initial set of base learners) and feature set, based on running an internal 5-fold cross validation over the training set, using different combinations of the number of features and base learners. This step reduces the number of dimensions to a number under 1000; see Chapter 6. The selected features and learner are then run on the training set to produce a classifier. HOG feature extraction, MRMR feature selection and base learner selection are all parts of the MHCD algorithm (shown in the red box above). See Algorithm 1 for details.

$$\text{dir}(x, y) = \underset{b}{\text{argmax}}\ \alpha(\nabla f(x, y, z), b)$$

where b ranges over each of the 26 vectors (see Figure 5.2). Then the vector contributes to its corresponding bin as in the $2D$ space.

Each cell in our $3D$ HOG consists of $8 \times 8 \times 8 = 512$ voxels. A histogram of size 26 is built for each cell and each block includes $2 \times 2 \times 2 = 8$ cells. Using these parameters the vector representing each block has a size of $8 \times 26 = 208$. $3D$ HOG, with these parameters, identifies each subject with $116,480$ features. We used the same normalization scheme as mentioned above for the $2D$ space ($v \to \frac{v}{\sqrt{||v||_2^2 + \epsilon^2}}$) for each block of the image.

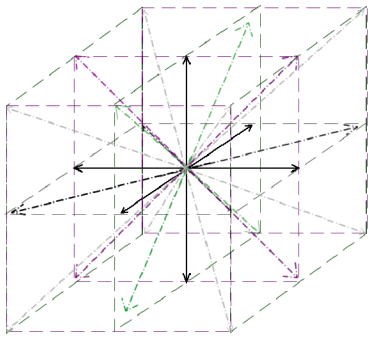


Figure 5.2: $3D$ bins

(a) Primary image (b) Shifted image

Figure 5.3: HOG features are shift invariant in each cell of the image

5.2 Properties of HOG features

Image texture gives information about the spatial arrangement of intensities in an image or selected region of an image [Shapiro and Stockman, 2001]. The basic question was if texture differences could be found in structural (respectively functional) images that distinguish different groups – here, we chose to use (f)MR images to see if they can be useful to distinguish healthy versus patient. We can consider different ways to compare the image textures in healthy controls and patients.

One of the ways to describe the image texture is to compute the gradient vector's angle and magnitude for each pixel/voxel of the image. There are different feature extraction algorithms that use gradient vectors to describe the image texture. Among the algorithms that represent the image texture by gradient vectors, Scale Invariant Feature Transform (SIFT) [Lowe, 1999] and Histogram of Oriented Gradients [Dalal and Triggs, 2005] were very successful in another range of tasks.

Recall that all of the images are normalized and registered to the MNI T1 template (see Section 3.4), which means that each specific point (x, y, z) in the image should refer approximately to the same area (e.g. cerebellum) in the brain for all of the images. Therefore, when all of the HOG features are extracted and put in a long row vector, each element in the vector is referring to the gradient vector extracted from the same area in the brain. HOG features have one important property:

- HOG features capture the orientation and the magnitude of the gradient vectors of each pixel/voxel in each cell of the image but not the location within the cell, which

means they are shift invariant (over small distances – i.e., within a cell). This means that if the voxels in a cell of the image are shifted (see Figure 5.3) over a small range of locations, the HOG features will remain the same. We know that the spatial normalization of the brain images is not perfect so it is better if our algorithm is not sensitive to errors in the spatial normalization process. This means that small registration errors will not make the HOG features of the images different.

Our empirical question was to see if these features can find any differences in healthy and non-healthy brains and produce classifiers that can successfully diagnose different mental diseases.

5.3 Classifier for automated diagnosis

Here, we consider the MHCD learner, which uses the HOG features of labeled brain images and personal characteristic data (over a large number of instances) as input, to learns binary classifiers to diagnose either ADHD versus control (using the ADHD-200 data) or autism versus control (using the ABIDE data). MHCD returns the best classifier over a subset of HOG features of the images and perhaps other patient features, based on a 5-fold cross validation on the training set. MHCD uses several base learners (each a learning algorithm like support vector machine with radial basis function kernel and a specific sigma value, naive bayes or k-nearest neighbours [Hastie et al., 2009]), the $3D$ HOG feature extraction method [Dalal and Triggs, 2005] and the MRMR feature selection algorithm [Peng et al., 2005, Ding and Peng, 2005]. As each dataset had only about 1000 individuals, there was a high chance the learning algorithms would overfit to the training data if we used all of the features. We therefore used MRMR (maximum relevance minimum redundancy) [Peng et al., 2005, Ding and Peng, 2005] as a preprocessing step to select the most relevant features (Figure 5.1 part 3).

Our MHCD system is summarized in Algorithm 1. For notation, let:

- acc(L, D, FS) be the 5 accuracy values computed from each fold of 5-fold cross validation, using base learner L on dataset D with feature set FS,
- Eacc(L, D, FS) be the mean of these 5 values,
- Racc(L, D, FS) to be the range of accuracy values over the 5 folds (see the range variable in Algorithm 1),
- D_{train} be the training data, and D_{test} the hold-out data,
- $FS_k(D)$ be the top k MRMR features over the dataset D,
- LS be the set of base learners: LS $= \{$ SVM-RBF-1, SVM-RBF-2, ..., SVM-RBF-9, NB, KNN $\}$,
- SVM-RBF-i denote support vector machine with radial basis function kernel and sigma value equal to i,
- L^* denote the best base learner and
- $FS^*(L^*)$ denote the best feature set associated with the best learner.

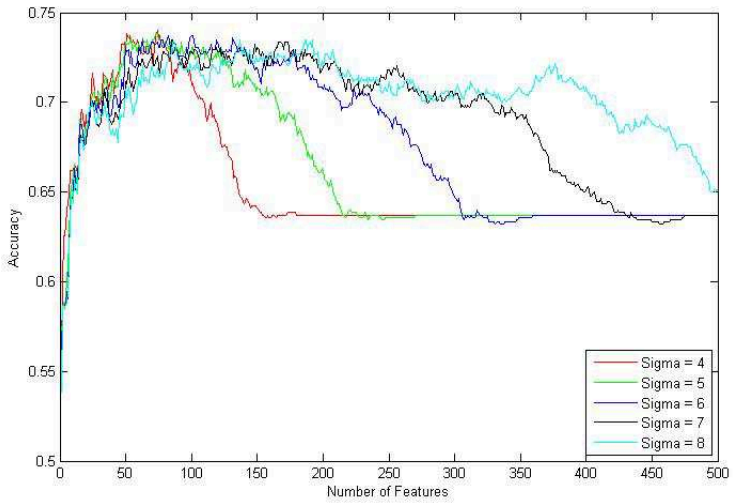


Figure 5.4: 5-fold cross validation accuracies on the training set
The accuracies are obtained using RBF SVM (with various sigma values), on training portion of the ADHD-200 dataset using functional images plus personal characteristic data.

Our MHCD system considers various base learners (with various parameters). Using dataset D (here we used D_{train} only), for each base learner $L \in LS$, MHCD sequentially considers using the top $k = \{1, 2, 3, ...\}$ MRMR features. It uses a 5-fold cross validation approach, stopping when there is no accuracy increase in all of the learners and the Eacc(L, D, FS_k) reaches a plateau. Figure 5.4 shows that each of the learners reaches a peak and then drops significantly after that point. We identified each base learner with both the mean accuracy achieved using the best feature set, Eacc(L, D_{train}, FS) and also the range Racc(L, D_{train}, FS) on D_{train}. We found the 5 base learners with the top 5 mean accuracy values. Because the top 5 accuracies were very close, our MHCD chose the learner with the smallest Racc(L, D_{train}) and returned that learner. After learning this L^*, and $FS^*(L^*)$, we then ran this learner and features on D_{test}.

Algorithm 1 : MHCD

```
1:  procedure MHCD(D: training data, 𝓛: set of base learners)
2:      for each base learner L in 𝓛 do
3:          FS ← top 1 MRMR feature ( on D )
4:          repeat
5:              vals[L, 1 : 5] ← acc( L, D, FS ) // 5-fold cross validation accuracy
                of L on D, using only the features FS
6:              mean ← ave_i {vals[L, i]}
7:              range ← max_i{vals[L, i]} − min_j{vals[L, j]}
                //range of accuracy over the 5 folds
8:              AveAcc[L, FS] ← [mean, range]
9:              FS ← FS + top 1 new MRMRSelectedFeatures from D
10:         until (No accuracy increase in all of the learners)
              // see text and Figure 5.4 for more information
11:     end for
12:     i ← 1
13:     for each base learner L in 𝓛 do
14:         bestFS ← argmax_FS{AveAcc[L, FS][1]}
          // the feature set that has the largest mean value for this learner L
15:         TopAccs[i] ← [L, bestFS, AveAcc[L, bestFS][1], AveAcc[L, bestFS][2]]
          //this tuple lists the best feature set for this L, along with the average accuracy for
          this [L, bestFS] pair, and the range of accuracies
16:         i ← i + 1
17:     end for
18:     TopAccs ← sort TopAccs corresponding to third element
          // which is the mean value of each [L, FS] pair, sorted large to small
19:     BestAcc[ 1 : 5 ] ← TopAccs[ 1 : 5 ]
20:     BestL ← tuple in BestAcc with smallest 4th entry ( range )
21:     L* ← BestL[ 1, 1 ]
22:     FS* ← BestL[ 1, 2 ]
23:     classifier ← L*(D, FS*)
          // run learner L* on whole training set D, using only the features FS*
24:     return classifier
25: end procedure
```

Chapter 6

Experiments

The ADHD-200 Global Competition divided the ADHD-200 dataset into training and test sets; we used this split. In the ABIDE dataset, we randomly selected a label-balanced $4/5$ of the data as the training data, and left the remaining $1/5$ as the testing data. We ran our MHCD system on each training dataset using different kinds of training and testing data that varied based on whether it included personal characteristic (PC) and/or RS-fMRI and/or structural MRI features. (Below we consider 5 of the $2^3 = 8$ sub-collections of these 3 types of features). We ran MHCD on various sub-collections of these feature sets – in each case, following the methodology mentioned above.

Below, Section 6.1 represents the results of the learning using only RS-fMRI. Section 6.2 explains the results of using only structural MRI scans. The results of using only personal characteristic data can be found in Section 6.3. When we combined personal characteristic data with structural or functional images, we got the results listed in Section 6.4.

6.1 Functional images

We ran our MHCD algorithm using the $116,480$ HOG features derived from only the fMRI data from the ADHD-200 (respectively ABIDE) dataset. For the ADHD-200 training dataset, MHCD determined that the best learner was the RBF SVM with Sigma $= 9$ and a specific set of 469 HOG features. When we ran this learning and feature set on the test set, its accuracy was 0.5965. For the ABIDE dataset, MHCD decided that SVM with RBF kernel with Sigma $= 6$ with 110 features was the best; when it was run on its hold-out set, its accuracy was 0.5919.

Tables 6.1 and 6.2 list the top 5 learners, their accuracy on the training set using 5-fold cross validation and the best learner's accuracy on the test set. When using only functional images we found that all top 5 learners were support vector machines with RBF kernels, with different Sigma values. In all of the tables, the learners are sorted in decreasing value of the training accuracy.

6.2 Structural images

We then explored the performance using structural MRI (structural images) – that is, just using T1-weighted images. We therefore ran the same processing, including HOG feature extraction, MRMR feature selection, and using the exact same set of base learners. The results for both datasets are shown in Tables 6.3 and 6.4.

Learner^{+}, L	Number of features, $\lvert FS^{*}(L)\rvert$	Training Accuracy, Eacc(L, D_{train}, $FS^{*}(L)$)	Range	Test Accuracy, acc(L^{*}, D_{test}, $FS^{*}(L^{*})$)
RBF-9	469	0.7035	0.0736	0.5965
RBF-8	311	0.7021	0.1126	
RBF-7	227	0.6931	0.1104	
RBF-6	204	0.6931	0.1191	
RBF-5	145	0.6892	0.0909	

Table 6.1: ADHD-200, functional images

$^{+}$RBF-i represents SVM with RBF kernel with Sigma = i.

(This is true for all subsequent tables as well).

Learner, L	Number of features, $\lvert FS^{*}(L)\rvert$	Training Accuracy, Eacc(L, D_{train}, $FS^{*}(L)$)	Range	Test Accuracy, acc(L^{*}, D_{test}, $FS^{*}(L^{*})$)
RBF-8	128	0.5890	0.0530	
RBF-9	248	0.5889	0.0730	
RBF-3	53	0.5878	0.0843	
RBF-5	110	0.5856	0.0393	
RBF-6	110	0.5856	0.0249	0.5919

Table 6.2: ABIDE, functional images

Here we found that the top 5 learners for the ADHD-200 dataset were all SVM with RBF kernels while in the ABIDE dataset MHCD chose Naive Bayes as one of the top 5 learners, although the Naive Bayes learner was not selected as the best. The test accuracy in the ADHD-200 dataset was 0.6608 using RBF SVM-8 with 285 features. The ABIDE test set diagnosis accuracy was 0.6009 using RBF SVM-6 with 194 features.

Learner, L	Number of features, $\lvert FS^{*}(L)\rvert$	Training Accuracy, Eacc(L, D_{train}, $FS^{*}(L)$)	Range	Test Accuracy, acc(L^{*}, D_{test}, $FS^{*}(L^{*})$)
RBF-8	285	0.7230	0.0519	0.6608
RBF-9	467	0.7217	0.0649	
RBF-7	274	0.7166	0.0764	
RBF-6	204	0.7153	0.0763	
RBF-5	131	0.7088	0.0699	

Table 6.3: ADHD-200, structural images

Learner, L	Number of features, $\|FS^*(L)\|$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
RBF-9	590	0.5889	0.0755	
Naive Bayes	11	0.5867	0.0449	
RBF-6	194	0.5811	0.0418	0.6009
RBF-7	357	0.5810	0.1093	
RBF-8	651	0.5788	0.0812	

Table 6.4: ABIDE, structural images

Learner, L	Number of features, $\|FS^*(L)\|$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
RBF-1	5	0.7178	0.0584	0.6901
RBF-2	4	0.7087	0.0974	
Naive Bayes	6	0.7074	0.0996	
RBF-3	4	0.7022	0.1039	
RBF-4	5	0.6996	0.0714	

Table 6.5: ADHD-200, personal characteristic data

6.3 Personal characteristic data

We also investigated if personal characteristic (PC) data can help the classification of these psychiatric disorders. For the ADHD-200 dataset we used age, gender, handedness, IQ measure, full4IQ score and site of the imaging. For the ABIDE dataset we used age, gender, handedness, fIQ standard score, pIQ standard score, vIQ standard score, site of the imaging and eyestat. Some of the values of the mentioned features were missing in the dataset; here we replaced each missing value of a feature with the mean of the values we had available for this feature.

Using the MHCD algorithm with only PC data as the input for classifying the ADHD disease gave an accuracy of 0.6901 (over baseline $= 0.5497$), which is consistent with the result of Brown et al. [2012]. The same process with the ABIDE dataset resulted in 0.5964 accuracy over the baseline of 0.5157.

In the ADHD-200 dataset, our MHCD system selected SVM with RBF kernel with Sigma $= 1$ and 5 features as the best learner. The only feature that was not chosen by our algorithm was IQ measure. In the ABIDE dataset, our method chose SVM with RBF kernel with Sigma $= 2$ and all 8 features as the best learner. The MHCD-chosen learners and the test accuracy of both datasets are shown in Tables 6.5 and 6.6.

Learner, L	Number of features, $\lvert FS^*(L)\rvert$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
Naive Bayes	8	0.6273	0.0654	
RBF-3	8	0.6261	0.0599	
RBF-4	8	0.6205	0.0654	
RBF-2	8	0.6194	0.0337	0.5964
RBF-6	7	0.6194	0.0767	

Table 6.6: ABIDE, personal characteristic data

Learner, L	Number of features, $\lvert FS^*(L)\rvert$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
RBF-8	83	0.7452	0.0714	
RBF-9	127	0.7451	0.0649	
RBF-3	42	0.7451	0.0603	
RBF-7	117	0.7412	0.0779	
RBF-5	129	0.7412	0.0390	0.6959

Table 6.7: ADHD-200, personal characteristic data with structural images

6.4 Adding personal characteristic data to functional or structural images

We simply added these PC features to structural images (respectively functional images). For example in the ADHD-200 dataset, we used $116,480$ HOG features extracted from each functional image, plus 6 more personal characteristic features (PC data) to produce a $116,486$-sized feature set, then let MHCD choose which features to use for classification.

Brown et al. [2012] showed that a learned classifier that uses only personal characteristic data without any imaging data, can diagnose ADHD with an accuracy higher than any of the other approaches proposed in the ADHD-200 competition. Note that these other approaches

Learner, L	Number of features, $\lvert FS^*(L)\rvert$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
RBF-4	74	0.7399	0.0844	
RBF-5	73	0.7386	0.0909	
RBF-6	101	0.7374	0.0634	
RBF-7	85	0.7348	0.0779	
RBF-8	137	0.7347	0.0455	0.6433

Table 6.8: ADHD-200, personal characteristic data with functional images

Learner, L	Number of features, $\|FS^*(L)\|$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
RBF-9	5	0.6295	0.0843	
RBF-4	54	0.6285	0.1059	
RBF-6	66	0.6273	0.0993	
RBF-7	55	0.6273	0.0712	0.6413
RBF-8	90	0.6262	0.0824	

Table 6.9: ABIDE, personal characteristic data with structural images

Learner, L	Number of features, $\|FS^*(L)\|$	Training Accuracy, Eacc(L, D_{train}, $FS^*(L)$)	Range	Test Accuracy, acc(L^*, D_{test}, $FS^*(L^*)$)
RBF-7	50	0.6183	0.0861	
RBF-3	47	0.6171	0.0827	0.6502
RBF-4	56	0.6160	0.0895	
RBF-6	39	0.6148	0.1009	
RBF-2	32	0.6115	0.0934	

Table 6.10: ABIDE, personal characteristic data with functional images

used both functional images and personal characteristic data for classification and had structural images available for all of the subjects. The best binary imaging-including classification accuracy (brain images plus personal characteristic data) achieved in the ADHD-200 competition was 0.6154 when trying to build a two way classifier for healthy versus ADHD [ADHD-200 consortium, 2011b]. Here we have confirmed that personal characteristic data can classify the ADHD disease with an accuracy that is over 14 % more than the baseline (almost 0.69). We have also shown that personal characteristic data can diagnose autism with about 0.60 accuracy. We also investigated if personal characteristic data can help imaging data in diagnosing psychiatric diseases or not.

Figure 6.1 and Tables 6.5, 6.1 and 6.8 (respectively) show that the accuracy (on the ADHD-200 dataset) of the learned classifier using only personal characteristic data is 0.69 and using only functional images is 0.5965, while the accuracy of using personal characteristic data with functional images is 0.6433. When we only use structural images as the input of our system, we achieve an accuracy of 0.6608. When we add personal characteristic data to structural images the accuracy increases to 0.6959 (see Tables 6.3 and 6.7). Comparing Tables 6.1 and 6.5 with Table 6.8, we conclude that adding personal characteristic data to functional images in ADHD-200 dataset does not improve the classification performance compared to using only PC data. Note that the results of using functional images and personal characteristic data were better than using only functional brain images for classification (see Table 6.1 and Table 6.8 only). This result is also consistent with the results of Brown et al. [2012], as they too achieved their best accuracy using only PC but not by using PC and functional images together. Brown et al. [2012] did not test combined PC and fMRI data, just PC or fMRI data separately.

On the other hand, in the ABIDE dataset (Tables 6.6, 6.2 and 6.10 and Figure 6.2),

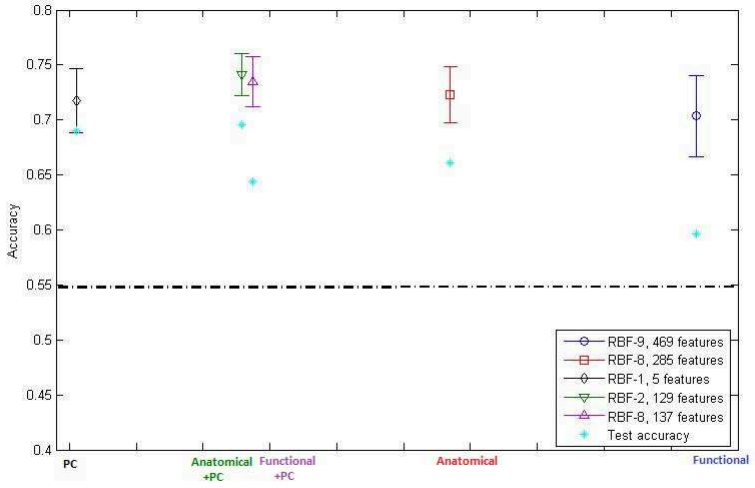


Figure 6.1: Summary of ADHD-200 dataset classification results
The horizontal dotted line shows the baseline accuracy of the test set. Each vertical bar shows the mean and range of the cross validation results for the selected base learner (L) and feature set (FS*(L)) as mentioned in Algorithm 1. The blue '*' shows the accuracy of the classifier on the hold-out set. The classifiers on the x-axis are ordered by the number of features they used. The legend also identifies the actual classifier used.

we see that the accuracy of using personal characteristic data as input is about 0.60 and the accuracy when using functional images is also about 0.60 but when we use these data together we achieve over 0.65 accuracy (see Table 6.10). The same phenomena happens for personal characteristic data and structural images (see Tables 6.4 and 6.9). The results of all of the experiments are summarized in Figure 6.1 (ADHD-200 dataset) and Figure 6.2 (ABIDE dataset).

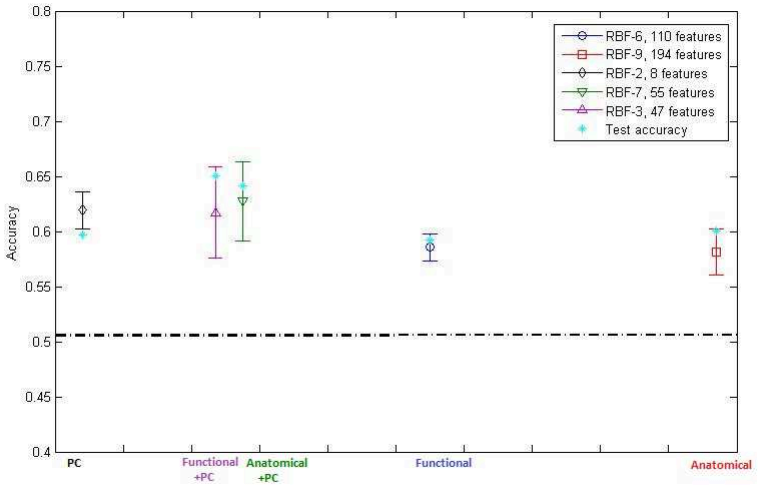


Figure 6.2: Summary of ABIDE dataset classification results

Chapter 7

Discussion

The best 2-class (ADHD versus healthy) imaging-including (imaging data and personal characteristic data) accuracy in the ADHD-200 dataset we could achieve was 0.6959 using structural images and personal characteristic data, which was 8% better than the best imaging-including (functional scans, structural scans and personal characteristic data) diagnostic performance, 0.6154, achieved in the ADHD-200 global competition [ADHD-200 consortium, 2011b]. (Note that our accuracy scores for the ADHD-200 test set did not include the 26 subjects from the Brown site, as their diagnostic labels have not been released). Sidhu et al. [2012] achieved an average accuracy of 0.7604 using cross-validation on 668 subjects of the ADHD-200 training dataset. Their best 2-class accuracy result on the hold-out set was 0.6667 using imaging and personal characteristic data. In a recent article, Dey et al. [2014] achieved an accuracy of 0.7355 on the test data but they only used 4 of the imaging sites for their analysis (they used only 487 of the subjects in the dataset). In another article Chang et al. [2012] used 436 male subjects in the ADHD-200 dataset and they achieved an accuracy of 0.6995. Their results are not comparable with our results since we used all of the available subjects in the dataset for our analysis. They used a 10-fold cross validation while we used the ADHD-200 test set for analyzing the generalization ability of the method.

For the ABIDE dataset we could achieve an accuracy of 0.6502 on a hold-out set, using functional images with personal characteristic data (note on this hold-out set, the baseline was 0.5157). This is better than the result of Nielsen et al. [2013], who achieved 0.60 accuracy against their baseline of 0.5363 . (The difference in baseline accuracies was because Nielsen et al. omitted 148 of the individuals due to preprocessing problems). To our knowledge, our results are the best to date.

Note that our results are not directly comparable to the results of Nielsen et al. [2013] because their feature selection method (called "binning") was run on the dataset a few times (with a leave-one-out scheme each time) using different number of "bins"(brain connections). This "best accuracy" was actually based on examining the test set scores of all of the bins; that is, it was based indirectly on all of the data (not just the training set). This means that their true (generalization) accuracy may be under the reported 0.60 if run on a *hold-out* set.

We also found that, when using personal characteristic data along with structural or functional images, MHCD chooses models that require fewer features than ones based on only structural or only functional images. As an example, when using only functional images, we could achieve an accuracy of 0.5965 with 469 features, but when using personal

characteristic features as well as functional image features the accuracy increased to 0.6433 while using only 137 features in the ADHD-200 dataset. We can see this by comparing Tables 6.1 and 6.8. We can also observe in Figures 6.1 and 6.2 that MHCD chooses models with fewer features when using personal characteristic data (models that do not use PC data, are on the right side of the figure for both ADHD-200 and ABIDE datasets).

This report shows that it is possible to diagnose either ADHD or autism with accuracy levels 14.5% and 12.5% above chance using structural MR brain images – indeed, more accurately than any other method presented to date.

To summarize, we have improved the results on classification of ADHD and autism on 2 large datasets, ADHD-200 and ABIDE, which suggests that both ADHD and autism can be diagnosed using brain images and personal characteristic data. We have also shown that HOG features, which are well known for object detection, can be useful for classification of psychiatric illnesses using brain images. Since we successfully applied our method to learn two diagnosis methods from two large multisite datasets, we expect that our approach might be appropriate for other datasets of brain images from psychiatric patients. Further research will be needed to address this question.

Bibliography

Robert R Edelman and Steven Warach. Magnetic resonance imaging. *New England Journal of Medicine*, 328(10):708–716, 1993.

Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*. 2004a.

MH Lee, CD Smyser, and JS Shimony. Resting-state fmri: A review of methods and clinical applications. *American Journal of Neuroradiology*, 2012.

Robert C Wolf, Michael M Plichta, Fabio Sambataro, Andreas J Fallgatter, Christian Jacob, Klaus-Peter Lesch, Martin J Herrmann, Carlos Schönfeldt-Lecuona, Bernhard J Connemann, Georg Grön, et al. Regional brain activation changes and abnormal functional connectivity of the ventrolateral prefrontal cortex during working memory processing in adults with attention-deficit/hyperactivity disorder. *Human brain mapping*, 30(7):2252–2266, 2009.

Jody C Culham, Stacey L Danckert, Joseph FX De Souza, Joseph S Gati, Ravi S Menon, and Melvyn A Goodale. Visually guided grasping produces fmri activation in dorsal but not ventral stream brain areas. *Experimental Brain Research*, 153(2):180–189, 2003.

ADHD-200 consortium. Adhd-200 global competition. `http://fcon_1000.projects.nitrc.org/indi/adhd200/`, 2011a.

ADHD-200 consortium. Adhd-200 global competition results. `http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html`, 2011b.

ABIDE dataset. Autism Brain Imaging Data Exchange. `http://fcon_1000.projects.nitrc.org/indi/abide/`, 2011.

A Di Martino, CG Yan, Q Li, E Denio, FX Castellanos, K Alaerts, JS Anderson, M Assaf, SY Bookheimer, M Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 2013.

KJ Friston, CD Frith, PF Liddle, and RSJ Frackowiak. Functional connectivity: the principal-component analysis of large (pet) data sets. *Journal of cerebral blood flow and metabolism*, 13:5–5, 1993.

Yufeng Zang, Tianzi Jiang, Yingli Lu, Yong He, and Lixia Tian. Regional homogeneity approach to fmri data analysis. *Neuroimage*, 22(1):394–400, 2004.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*. 2004b.

Kerstin Konrad and Simon B Eickhoff. Is the adhd brain wired differently? a review on structural and functional connectivity in attention deficit hyperactivity disorder. *Human brain mapping*, 31(6):904–916, 2010.

Matthew RG Brown, Gagan S Sidhu, Russell Greiner, Nasimeh Asgarian, Meysam Bastani, Peter H Silverstone, Andrew J Greenshaw, and Serdar M Dursun. Adhd-200 global competition: diagnosing adhd using personal characteristic data can outperform resting state fmri measurements. *Frontiers in systems neuroscience*, 6, 2012.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Jared A Nielsen, Brandon A Zielinski, P Thomas Fletcher, Andrew L Alexander, Nicholas Lange, Erin D Bigler, Janet E Lainhart, and Jeffrey S Anderson. Multisite functional connectivity mri classification of autism: Abide results. *Frontiers in human neuroscience*, 7, 2013.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*. 2004c.

Statistical parametric mapping, version 8. `http://www.fil.ion.ucl.ac.uk/spm/`, 2009.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

D Louis Collins, Alex P Zijdenbos, Wim FC Baaré, and Alan C Evans. Animal+ insect: improved cortical structure segmentation. In *Information Processing in Medical Imaging*, pages 210–223. Springer, 1999.

VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009.

Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, and D Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

Jean Talairach and Pierre Tournoux. Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: an approach to cerebral imaging. 1988.

Seong-Gi Kim, Wolfgang Richter, and Kamil Ugurbil. Limitations of temporal resolution in functional mri. *Magnetic Resonance in Medicine*, 37(4):631–636, 1997.

Jyri-Johan Paakki, Jukka Rahko, Xiangyu Long, Irma Moilanen, Osmo Tervonen, Juha Nikkinen, Tuomo Starck, Jukka Remes, Tuula Hurtig, Helena Haapsamo, et al. Alterations in regional homogeneity of resting-state brain activity in autism spectrum disorders. *Brain research*, 1321:169–179, 2010.

J Radua, D Mataix-Cols, ML Phillips, W El-Hage, DM Kronhaus, N Cardoner, and S Surguladze. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European psychiatry*, 27(8):605–611, 2012.

Matthew Brett, Jean-Luc Anton, Romain Valabregue, and Jean-Baptiste Poline. Region of interest analysis using the marsbar toolbox for spm 99. *Neuroimage*, 16:S497, 2002.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

Linda Shapiro and George C Stockman. Computer vision. 2001. *ed: Prentice Hall*, 2001.

David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

Gagan S Sidhu, Nasimeh Asgarian, Russell Greiner, and Matthew RG Brown. Kernel principal component analysis for dimensionality reduction in fmri-based diagnosis of adhd. *Frontiers in systems neuroscience*, 6, 2012.

Soumyabrata Dey, A Ravishankar Rao, and Mubarak Shah. Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects. *Frontiers in Neural Circuits*, 8:64, 2014.

Che-Wei Chang, Chien-Chang Ho, and Jyh-Horng Chen. Adhd classification by a texture analysis of anatomical brain mri data. *Frontiers in systems neuroscience*, 6, 2012.