DOCTOR OF MEDICINE

Development of Machine Learning-Based Techniques in Psychiatric Neuroimaging

Blair Johnston

2014

University of Dundee

Development of Machine Learning-Based Techniques in Psychiatric Neuroimaging

Blair Alexander Johnston

A thesis presented for the degree of Doctor of Philosophy at the University of Dundee

February, 2014

Table of Contents

Table of Contents ii
Figures vi
Tables xi
List of Abbreviations xii
Acknowledgements xiii
Declaration xiv
Abstract xv
Publications xvi
Awards xvii
Chapter 1: Introduction 1
Chapter 2: Methods 6
2.1 Neuroimaging data quality 6
2.2 Pre-processing 9
2.3 Structural MRI Pre-processing and DARTEL 9
2.4 fMRI Pre-processing 13
2.5 fMRI first level analysis 14
2.6 Second level analysis 15
2.7 Neuroimaging data quality - Outlier analysis 15
2.8 Multivariate Pattern Analysis 16
2.9 Support Vector Machine 19
2.10 Application of SVM 21
2.11 Relevance Vector Machine 26
2.12 Application of RVR 27
2.13 Alternative Classifiers 29
2.14 Feature Selection 30
2.15 Feature selection methods used in this thesis 33
2.15.1 Mean-thresholding method 33
2.15.2 Thresholded t-test method 34
2.15.3 Thresholded linear regression method 34
2.15.4 Recursive Feature Elimination 35
Chapter 3: Literature review 36
3.1 Introduction 36

3.2 Overview of Neuroimaging Studies applying MVPA techniques 36
3.3 Child and adolescent ADHD MVPA Neuroimaging studies 38
3.4 MDD MVPA Neuroimaging studies 41
3.5 Summary 51
Chapter 4: Predicting Methylphenidate Treatment Response in Drug-Naïve Boys with ADHD. 53
4.1 Introduction 53
4.2 Methods 54
4.2.1 Subjects 54
4.2.2 Variable preparation 55
4.2.3 Discriminant Analysis 55
4.2.4 Individual Scan Classification 55
4.2.5 Euclidean Distance from the SVM Hyperplane Investigation 59
4.3 Results 59
4.3.1 Participant Characteristics 59
4.3.2 Discriminant Analysis 59
4.3.3 Individual Subject SVM Predictions 61
4.3.4 Euclidean Distance from the SVM Hyperplane Investigation 63
4.4 Discussion 69
Chapter 5: ADHD diagnostic classification using structural MRI data 70
5.1 Introduction 70
5.2 Methods 74
5.2.1 Subjects 74
5.2.2 Image Acquisition 76
5.2.3 Image Pre-processing 76
5.2.4 Group Level Comparisons 77
5.2.5 Individual Scan Classification 78
5.2.6 Multivariate Feature Selection 87
5.2.7 Calculating the number of Voxels in each cluster 88
5.3 Results 92
5.3.1 VBM Analysis 92
5.3.2 Individual Subject SVM Predictions 96
5.3.3 Brain Regions identified using Feature Selection 96
5.3.4 Comparison between VBM analysis and Classification 97

5.3.5 Multivariate Feature Selection 103
5.3.6 Group level differences between previously medicated and unmedicated ADHD subjects 103
5.4 Discussion 107
Chapter 6: The iBOCA study 112
6.1 Introduction 112
6.2 Recruitment Criteria 113
6.3 fMRI paradigm 114
6.4 Personal input towards the fMRI paradigm 119
6.5 Review of the Practicalities of Scanning Children and Adolescents 122
6.6 Practical Aspects of Preparing Volunteers for Scanning 128
6.7 Practical Aspects of Scanning: Data quality checks 130
6.8 Theory of ADHD Syndrome Mechanism: the Dopamine Transfer Deficit (DTD) Theory of Altered Reinforcement Mechanisms in ADHD 131
6.9 Discussion 134
Chapter 7: Diagnostic Classification and Prediction of Symptom Severity in MDD 136
7.1 Introduction 136
7.2 Method 139
7.2.1 Subjects 139
7.2.2 Image Acquisition 140
7.2.3 Image Pre-processing 140
7.2.4 Neuroimaging data quality - Outlier analysis 140
7.2.5 Individual Scan Classification 145
7.2.6 Group Level Comparisons 146
7.3 Results 146
7.3.1 Participant Characteristics 146
7.3.2 Individual Subject SVM Predictions 149
7.3.3 Brain Regions identified using Feature Selection 149
7.3.4 VBM Analysis (t-test) 152
7.3.5 Whole Brain Individual Patient RVR Severity Score Predictions 157
7.3.6 Individual Patient RVR Severity Score Predictions using Feature Selection 160
7.3.7 VBM Analysis (Multiple Linear Regression) 166

7.4 Discussion 174
Chapter 8: High Accuracy Individual Diagnostic Classification in MDD using fMRI 180
8.1 Introduction 180
8.2 Method 182
8.2.1 Subjects 182
8.2.2 Image Acquisition 183
8.2.3 Image Pre-processing 183
8.2.4 fMRI Analyses 183
8.2.5 Individual Scan Classification 184
8.3 Results 185
8.3.1 Participant Characteristics 185
8.3.2 Within-Group and Between-Group Analyses 187
8.3.3 Individual Subject SVM Predictions 193
8.3.4 Brain Regions identified using Feature Selection 193
8.3.5 Correlations with Severity Scores 195
8.4 Discussion 200
Chapter 9: Conclusion 203
References 208

Figures

Figure 1: Two structural MRI images to highlight that differences in brain structure are so subtle that many disorders (such as ADHD) cannot be classified subjectively. Left: scan from healthy child. Right: scan from a child with ADHD. 2
Figure 2: Examples of poor quality images from ADHD-200 dataset. From left to right: Blurred image artefact most likely due to movement, cerebellum not fully scanned most likely due to patient positioning within the head coil, blood flow artefacts most likely due to sub-optimal image parameters, and wrap artefact most likely due to a field of view that is too small. Such gross artefacts will obscure subtle syndromal differences in brain structure. 8
Figure 3: Three stages of pre-processing a structural MRI image of a healthy control. Left: A 'raw' structural MRI image. Centre: A segmented grey matter image. Right: A DARTEL-processed, smoothed grey matter image. 11
Figure 4: A summary of the SVM process. If there are 10 patients and 10 healthy controls suppose the 'left out' subject in this example of the LOOCV step is a patient. The training set therefore consists of 9 patients and 10 healthy controls. In this example only two voxels from the brain are selected to simplify the description of SVM principles. Left: SVM training – A plot of the two selected features of each subject in the training set (patients represented by blue squares and healthy controls represented by red circles). The best fitting decision boundary is calculated from the subjects in the training set. Right: SVM testing – The 'left out' subject (represented by the star) is plotted and the prediction depends on which side of the decision boundary the subject lies. In this case the 'left out' subject scan lies above the decision boundary so would be predicted to be a scan of a patient (if below the decision boundary it would be predicted to be from a control). 25
Figure 5: A flowchart of the SVM prediction technique. 57
Figure 6: A flowchart of the variable and parameter selection stage in Figure 5. 58
Figure 7: Individual subject's Euclidean distance from the linear hyperplane in the classification which achieved 76.7%. If the y-value = 0 the subject was correctly classified, if the y-value = 1 the subject was a non-responder who was predicted to be a responder (false positive) and if the y-value = -1 the subject was a responder who was predicted to be a non-responder (false negative). 64
Figure 8: Histogram of individual subject's Euclidean distance from the linear hyperplane in the classification which achieved 76.7%. The blue bars show the correctly classified subjects and the green bars show the incorrectly classified subjects. 65
Figure 9: Histogram displaying individual subject's absolute Euclidean distance from the linear hyperplane in the classification that achieved 76.7%. The blue bars show the correctly classified subjects and the green bars show the incorrectly classified subjects. 67
Figure 10: A plot of the ratio of correctly (blue) and incorrectly (green) classified subjects to total number of subjects within the histogram bins displayed in Figure 9. The dashed magenta line shows the difference between the ratio of correctly classified subjects (blue) and the ratio of incorrectly classified subjects (green). 68
Figure 11: Gaussian SVM white matter prediction, average training accuracy as a function of the wide-ranging cut-off values (used to identify the narrow range of cut-off values for the 3-variable grid search – the soft-margin and kernel width parameters are fixed). 81

Figure 12: Gaussian SVM white matter prediction, average training accuracy as a function of mid-ranging cut-off values (the second iteration of the range reduction technique, vector elements are used instead of cut-off values as each subject differs on the range selected), used to identify the narrow range of cut-off values for the 3-variable grid search – the soft-margin and kernel width parameters are fixed. 82
Figure 13: Gaussian SVM white matter prediction, average training accuracy versus variable soft-margin parameters and the narrowed cut-off ranges (vector elements are used instead of cut-off values as each subject differs on the range selected), in a 2-variable grid search (the kernel width parameter was fixed for illustration purposes). 83
Figure 14: A flow diagram outlining the primary LOOCV procedure. The general procedure involved applying LOOCV on the pre-processed images, applying the 'mean-threshold' method of feature selection and SVM parameter tuning (shown in more detail in Figure 15 and Figure 16) to the training data, training the SVM using the training data and the optimised parameters, then making a prediction for the left-out subject. 84
Figure 15: Parameter range reduction stage of the 'mean-threshold' feature selection method. In order to speed up the processing time, the range of threshold cut-off values was narrowed down by testing an arbitrarily wide range using an inner LOOCV procedure with fixed SVM parameters (fixed to unity). The selected range was based on the threshold cut-off values which achieved the highest training stage accuracy..... 85
Figure 16: The grid search procedure for parameter selection (including the optimal mean-threshold cut-off value). A second inner LOOCV procedure is performed whereby all combinations of mean-threshold cut-off values (from the narrowed range identified in Figure 15) and SVM parameters are tested. The combination which achieves the highest training stage accuracy are selected for training the SVM on the full training set for the classification of the left-out subject (Figure 14). 86
Figure 17: RFE method of feature selection with SVM. 89
Figure 18: RFE feature selection details for Figure 17 90
Figure 19: Parameter selection for RFE method in Figure 17. 91
Figure 20: VBM Group Level results. Significantly ($p<0.005$) reduced grey matter (left) and white matter (right) volume in ADHD. Reduced grey matter volume in the basal ganglia (BG - putamen) and cerebellum (C) and significantly reduced white matter volume in the brainstem (BS) and cerebellum (C). 93
Figure 21: (a) Feature selection (Gaussian SVM) identified brain regions in *white* matter. BS – brainstem regions comprising a lower region in the pons and smaller bilateral region in the mid-brain; FP - frontal pole white matter; PT - pyramidal tracts (b) Feature selection (Gaussian SVM) brain regions identified using *grey* matter. BG – basal ganglia; FP – frontal pole; STS – superior temporal sulcus; IPL – inferior parietal lobule; ITG – inferior temporal gyrus; TL – temporal lobe; OG – occipital gyrus. 98
Figure 22: Locations of the noradrenergic locus coeruleus (LC) and dopaminergic ventral tegmental area nuclei (VTA), in relation to the brainstem (BS) white matter region used for classification. LC and VTA locations from previous studies (Guitart-Masip *et al.*, 2012; Keren *et al.*, 2009; Mai *et al.*, 1997). 99
Figure 23: Brain regions identified using feature selection (red), voxel based morphometry (green), and regions common to both analyses (orange). BS – brain stem; SC – superior cerebellum; BG – basal ganglia; TL – temporal lobe; STG – superior temporal gyrus..... 102

Figure 24: Group Level (VBM) significantly ($p<0.005$) increased grey matter volume (left), decreased white matter volume (centre) and increased white matter volume (right) in medicated versus unmedicated ADHD patients. Medicated patients with ADHD had significantly increased grey matter volume in BA20 (inferior temporal gyrus), insula, inferior frontal gyrus, midbrain, frontal pole and medial temporal lobe compared to unmedicated ADHD patients. A white matter volume decrease was identified in the white matter deep to BA7 (parietal lobe) and increased white matter volume was identified in the white matter deep to the lateral orbitofrontal cortex and the uncinate fasciculus. 105
Figure 25: The overlap between the regions used in the grey matter prediction and the regions which had significantly increased grey matter volume in the medicated group. Blue: significant regions identified in the group level analysis (Figure 24) but not used in the grey matter classification (Figure 21 (b)). Red: regions used in the Gaussian SVM (Figure 21 (b)) but not identified as significant in the group level analysis (Figure 24). Purple: the region (frontal pole -FP) which was both identified in the group level analysis (Figure 24) and used in the grey matter prediction (Figure 21 (b)). 106
Figure 26: The stimuli displayed during the modified version of the Pessiglione task. In this example the reward pair is represented by square-shaped fractal images, neutral by circular-shaped fractal images and the loss pair by triangular shaped fractal images but the shapes are randomly assigned to the 3 stimulus types..... 115
Figure 27: The two possible outcomes for each stimulus type. The subjects were informed during task training that there was no difference between "No change in vouchers" and "Nothing". 116
Figure 28: An example trial to illustrate the sequence of images and the task timing. In this example the circular fractal pair is the neutral pair (as in Figure 26). This illustration displays the example when an image was selected within the time limit of 3 seconds (precisely after 'x' seconds in this example). As this was a neutral pair there were two possible outcomes: "No change in vouchers" and "Nothing" (which are essentially the same thing – see Figure 27 for the possible outcomes for each trial type). Each trial took between 10-20.75 seconds depending on the inter-trial interval. 118
Figure 29: Comparison between the timing of the Presentation code (left) and the Psychtoolbox code (right). The Psychtoolbox code (written by the author) shows an improved overall timing over the same session with identical responses..... 121
Figure 30: The mock scanner setup used to train volunteers in the iBOCA study. . 129
Figure 31: (A) A pictorial representation of the expected magnitude and timing of anticipatory dopamine cell firing in controls, compared to (B) the abnormal dopamine cell firing in the DTD theory. The DTD theory suggests that children with ADHD fail to correctly adjust their dopamine cell firing rate from the reward reinforcement/feedback time towards the response time in anticipation of a reward. 133
Figure 32: A plot showing the percentage of voxels considered to be outliers in each diagnostic group (blue- controls, red- MDD) when using DARTEL. The dotted lines indicate the mean values for each diagnostic group and each imaging modality. ... 142
Figure 33: A plot showing the percentage of voxels considered to be outliers in each diagnostic group (blue- controls, red- MDD) when using standard VBM. The dotted lines indicate the mean values for each diagnostic group and each imaging modality. 144
Figure 34: Feature selection (Gaussian SVM) identified brain regions in *grey* matter. 150

Figure 35: Feature selection (Gaussian SVM) identified brain regions in *white* matter. CG – cingulate gyrus; PC – posterior cingulate; IN – white matter deep to the insula. 151
Figure 36: *Group* level *grey* matter volume *reductions* in patients with MDD compared with healthy matched controls. PV- periventricular grey matter, C – caudate reductions, H – habenula and IN – insula. 153
Figure 37: *Group* level (a) *reductions* and (b) *increases* in *white* matter volume in patients with MDD compared with healthy matched controls. FR- frontal region, IN – white matter deep to the insula, CG – cingulate gyrus and PC – posterior cingulate. 154
Figure 38: Overlapping grey matter regions between features selected during classification (purple/blue) and regions selected in the VBM analysis (red/purple). 155
Figure 39 Overlapping white matter regions between features selected during classification (purple/blue) and regions selected in the VBM analysis (green/blue). 156
Figure 40: The best fit lines for whole brain severity score predictions (top: grey matter predictions, bottom: white matter predictions, left: HAM-D predictions, right: MADRS predictions). 158
Figure 41: The brain regions which were identified as the most predictive during the whole brain severity score predictions (top: grey matter predictions, bottom: white matter predictions, left: HAM-D predictions, right: MADRS predictions). 159
Figure 42: The best fit line for the prediction of the BDI score using thresholded multiple linear regression feature selection and grey matter images. 162
Figure 43: The best fit lines for thresholded multiple linear regression-based white matter severity score predictions (left: HAM-D prediction, right: MADRS prediction). 163
Figure 44: The best fit lines for RFE-based white matter HAM-D predictions (left: Gaussian kernel, right: RBF kernel). 164
Figure 45: The regions identified using RFE-based feature selection on a non-linear kernel and white matter images to predict HAM-D scores (left: Gaussian kernel, right: RBF kernel). 165
Figure 46: Group-level positive correlations between grey matter volume and HAM-D (top) and MADRS (bottom). 167
Figure 47: Group-level negative correlations between grey matter volume and HAM-D. 168
Figure 48: Group-level negative correlations between grey matter volume and MADRS. 169
Figure 49: Group-level positive correlations between white matter volume and HAM-D (left) and MADRS (centre and right). 170
Figure 50: Group-level negative correlations between white matter volume and HAM-D (A and B) and MADRS (C and D). 171
Figure 51: *Group* level *grey* matter volume *decreases* in patients with MDD with increased BDI scores. 172
Figure 52: *Group* level *white* matter volume *decreases* in patients with MDD with increased BDI scores. 173
Figure 53: Within-group analyses of the controlled win contrast, displaying activations in controls (left) and patients (right). 188
Figure 54: Between-group analysis of the controlled win contrast, displaying regions of increased activation in controls compared with patients (left) and increased activation in patients compared with controls (right). 189

Figure 55: Within-group analyses of the controlled loss contrast, displaying deactivations in controls (left) and patients (right). 190
Figure 56: Within-group analyses of the controlled loss contrast, displaying activations in the patient group. 191
Figure 57: Between-group analysis of the controlled loss contrast, displaying regions of increased activation in patients compared with controls. 192
Figure 58: The overlap between the regions identified during VBM analysis (blue and pink) and the brain regions identified during the classification of MDD patients and controls (pink) using a controlled win contrast (left) and a controlled lose contrast (right). 194
Figure 59: Group-level negative correlations between the basic loss contrast and total scores on the MADRS (left) and HAM-D (right). 196
Figure 60: Group-level positive correlations between the controlled loss contrast and total score on the BDI. 197
Figure 61: Group-level positive correlations between the controlled loss contrast and HAM-A. 198
Figure 62: Group-level positive correlations between the controlled loss contrast and total score on the BHS 199

Tables

Table 1: Clinical descriptors for responders and non-responders to MPH. Variables are shown as mean (standard deviation). *chi-square test with other tests being t-tests..... 60
Table 2: Frequency of variable selection in leave-one-out method. 62
Table 3: MNI coordinates of each cluster of grey matter volume decreases. 94
Table 4: MNI coordinates of each cluster of white matter volume decreases. 95
Table 5: MNI coordinates of each cluster of white matter identified using mean-threshold feature selection with the Gaussian SVM. The number of resampled voxels contained in each cluster was calculated as discussed in Chapter 2. 100
Table 6: MNI coordinates of each cluster of grey matter identified using mean-threshold feature selection with the Gaussian SVM. The number of resampled voxels contained in each cluster was calculated as discussed in Chapter 2. 101
Table 7: The number of ADHD subjects in a selection of fMRI studies. 126
Table 8: Clinical descriptors for the MDD and healthy control groups in the structural MRI analysis. Variables are shown as mean (standard deviation). *chi-square test with other tests being t-tests. 148
Table 9: Current Medication and State Illness Severity. No patients had psychotic symptoms and quetiapine was prescribed as an augmentation agent for antidepressants (Dorée *et al.*, 2007), similar to the long established use of lithium, L-tryptophan and triiodothyronine in treatment resistant depression. No obvious relationship between current medication and state illness severity was present. ‘mg’ indicates total dose per day, ‘mcg’ total micrograms per day..... 178
Table 10: Clinical and task peformance descriptors for the MDD and healthy control groups in the fMRI analysis. Variables are shown as mean (standard deviation). *chi-square test with other tests being t-tests. 186

List of Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism Spectrum Disorder
BDI	Beck Depression Inventory-II
BHS	Beck Hopelessness Scale
BOLD	Blood-oxygen-level dependent
BPVS	British Picture Vocabulary Scale
CANTAB	Cambridge Neuropsychological Test Automated Battery
CSF	cerebrospinal fluid
DARTEL	Diffeomorphic Anatomical Registration using Exponential Lie Algebra
DMtS	Delayed Matching to Sample
DTD	Dopamine Transfer Deficit
DTI	Diffusion Tensor Imaging
EHI	Edinburgh Handedness Inventory
EPI	Echo-planar Imaging
ERD	mean number of errors for distractors
FA	fractional anisotropy
fMRI	functional MRI
FMRIB	Functional MRI of the Brain
FSL	FMRIB Software Library
FWHM	full-width at half-maximum
HAM-A	Hamilton Anxiety Rating Scale
HAM-D	Hamilton Rating Scale for Depression
iBOCA	Imaging Brains of Children and Adolescents
LOOCV	leave-one-out cross-validation
MADRS	Montgomery-Åsberg Depression Rating Scale
MAE	mean absolute error
MD	mean diffusivity
MDD	Major Depressive Disorder
MGH-S	Massachusetts General Hospital (MGH-S) staging method
MNI	Montreal Neurological Institute
MP-RAGE	magnetisation-prepared rapid acquisition gradient echo
MPH	methylphenidate
MRI	magnetic resonance imaging
NART	National Adult Reading Test
RBF	radial basis function
RFE	Recursive Feature Elimination
RMSE	root mean squared error
RTT	reaction times to target stimuli
RVM/RVR	Relevance Vector Machine/Regression
RWTH	Rheinisch-Westfälische Technische Hochschule
SPM	Statistical Parametric Mapping
SVM	Support Vector Machine
TE	echo time
TR	repetition time
VBM	Voxel-Based Morphometry

Acknowledgements

First and foremost, I wish to thank my three supervisors, Douglas Steele, Keith Matthews and David Coghill, for all their support over the three years. I would like to particularly single out my main supervisor, Douglas. Throughout my time at Dundee he has always provided support and encouragement and has always made himself available to discuss any issues.

I am extremely grateful to all my colleagues that were involved in the collection of data that was analysed in this thesis. I would like to thank Kerstin Konrad and her colleagues in Aachen for allowing me to analyse their data at the start of my PhD and I am grateful to all the research staff in Dundee, particularly, Jennifer, Ian, Pat, Elena, Kirsty, Frances, Tessa, Craig, Christine and Mairi. I want to thank my office mates Victoria, Benson, Serenella, Aistis and James for their input and for making the long commute seem a little less painful.

I wish to acknowledge SINAPSE (Scottish Imaging Network: A Platform for Scientific Excellence) and the University of Dundee for jointly funding my PhD. I would also like to thank SMHRN (the Scottish Mental Health Research Network), an Anonymous Trust and Tenovus for funding the imaging studies detailed in this thesis.

Finally, I want to thank my family and friends for their incredible backing throughout this stressful period. My final acknowledgement goes to my wife, Hannah, for her patience, love and understanding throughout.

Declaration

This thesis has been written by myself and has not been accepted for any previous application for a degree. The work presented is my own except where explicitly stated otherwise in this text.

Abstract

The primary goal of this thesis is to investigate whether machine learning-based methods can be successfully applied to make clinically relevant predictions. These techniques are applied to a range of data such as demographic, socioeconomic and neuropsychiatric variables but primarily to structural and functional magnetic resonance imaging (MRI) data. The main focus of this thesis is to investigate whether the application of these techniques can increase the understanding of psychiatric disorders.

As MR images contain a large amount of information within each image, feature selection techniques, which can identify which brain regions are most relevant to the study, are of high importance to maximise the amount of relevant information that is entered into the machine learning approaches. Successfully combining feature selection and machine learning to psychiatric imaging studies has several advantages as the machine learning methods produce output that can separate two or more groups accurately on a subject-by-subject basis or make predictions of a continuous variable and the feature selection provides information on the neurobiology by, for example, highlighting brain regions that are consistently different between groups.

Two psychiatric disorders are investigated in this thesis: Attention Deficit Hyperactivity Disorder (ADHD) and Major Depressive Disorder (MDD). ADHD core symptoms include difficulty in sustaining attention, hyperactivity and impulsive behaviour. MDD is a mood disorder that is associated with persistent and disabling symptoms of low mood, anhedonia, hopelessness, guilt, low self-worth, poor concentration, lack of energy, suicidal thoughts and altered appetite and sleep (American Psychiatric Association, 2000). Both disorders do not have any established and reliable diagnostic or prognostic biomarkers so the work undertaken in this thesis aims to identify possible biomarkers using machine learning methods.

Publications

The following has been published or is under review, presenting results and discussion in this thesis:

1) Johnston, B.A., Mwangi, B., Matthews, K., Coghill, D., Steele, J.D. Predictive classification of individual magnetic resonance imaging scans from children and adolescents. European Child & Adolescent Psychiatry, 2012.

2) Johnston, B.A., Mwangi, B., Matthews, K., Coghill, D., Konrad, K., Steele, J.D. Brainstem Abnormalities in ADHD Support High Accuracy Individual Diagnostic Classification. (under review).

Awards

Johnston, B.A. Mwangi, B., Matthews, K., Coghill, D., Konrad, K., Steele, J.D. "Diagnostic classification of Attention Deficit Hyperactivity Disorder (ADHD) using individual structural MRI scans" - the 2nd International Eunethydis ADHD conference, **Early Career Prize Lecture**, Barcelona, 2012.

Johnston, B.A., Gradin, V.B., Mwangi, B., Stirling, M., Walker, K., MacFarlane, J., Matthews, K., Steele, J.D. "Using Structural and Functional MRI to Predict Diagnostic Status and Symptom Severity in Major Depression" - Scottish Mental Health Research Network (SMHRN) ASM, **Best Student Poster Prize,** Glasgow, 2012.

Johnston, B.A., Gradin, V.B., Mwangi, B., Stirling, M., Walker, K., MacFarlane, J., Matthews, K., Steele, J.D. "Using Structural and Functional MRI to Predict Diagnostic Status and Symptom Severity in Major Depression" - SINAPSE ASM, **Best Poster for Public Communication**, Aberdeen, 2013.

Chapter 1: Introduction

Differences in brain structure between patients suffering from various 'functional' psychiatric disorders (e.g. attention deficit hyperactivity disorder (ADHD) and major depressive disorder (MDD)) and healthy subjects are sufficiently subtle that they cannot be recognised by conventional radiological methods – namely qualitative visual inspection (Agarwal *et al.,* 2010). Indeed, the term 'functional' was adopted for these disorders because it was once believed there were no structural brain abnormalities (Mwangi *et al.,* 2012a). As an illustration, Figure 1 shows T_1 weighted ('structural') MRI scans from a subject with ADHD and a healthy subject. There are no obvious visual differences between these scans.

Using imaging data such as structural and functional MRI to aid the identification of reliable biomarkers of functional psychiatric disorders could have significant implications for informing both clinical decision making and research into the causes and consequences of each disorder (Klöppel *et al.,* 2012).

Machine learning is a rapidly expanding research area that has resulted in the development of several computational models that attempt to identify patterns in data that can be constructed into a predictive model (Bishop, 2006). Depending on the machine learning approach used, this predictive model can then be applied to novel data to determine group membership or the prediction of a continuous variable related to novel data from subject(s) not used in the initial model building stage. Although these techniques originated in engineering, mathematics and computer science, they are ideally suited to psychiatric neuroimaging studies as the algorithms were originally designed to be applied to medical predictions (Kononenko, 2001). Machine learning studies have the potential to make predictions such as clinical diagnosis or prognosis, symptom severity, identification of patients at risk of developing disorder, and an estimation of the likelihood of response to treatment (Klöppel *et al.,* 2012).

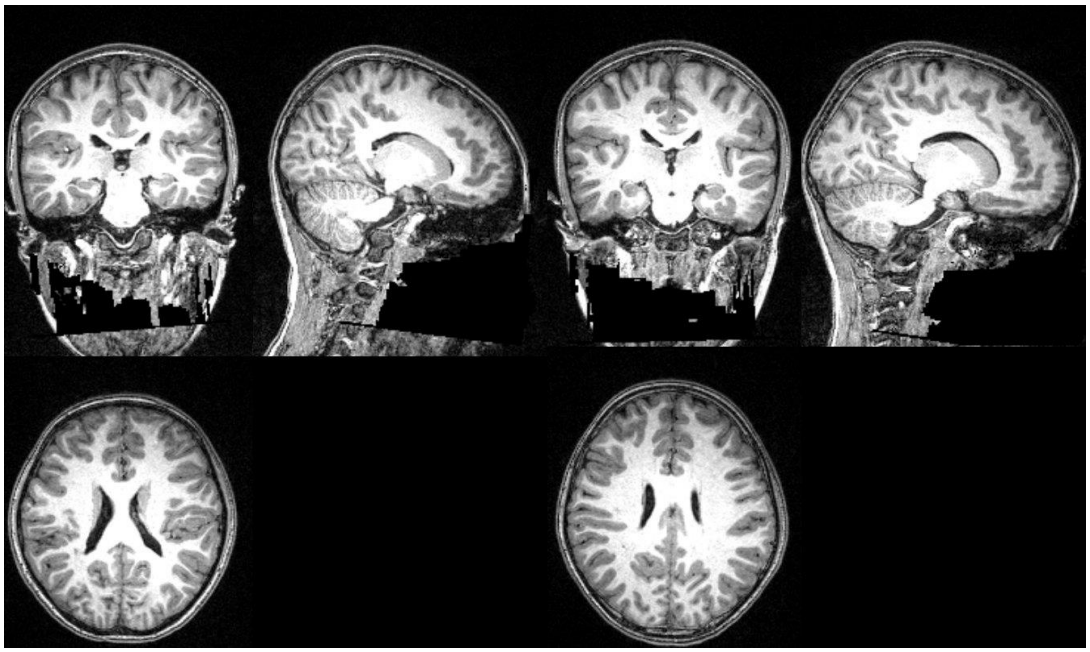

Figure 1: Two structural MRI images to highlight that differences in brain structure are so subtle that many disorders (such as ADHD) cannot be classified subjectively. Left: scan from healthy child. Right: scan from a child with ADHD.

As structural and functional MR images contain a large amount of information within each image, feature selection techniques that can identify the most relevant data to enter into the pattern recognition method, whilst discarding the less relevant or noisy data, provide two clear advantages when combined with machine learning methods. The first advantage is that reducing the number of features entered into the machine learning methods tend to reduce computational time (Theodoridis and Koutroumbas, 2006). The second advantage is feature selection provides information on neurobiology by highlighting brain regions that are most relevant to the predictive model (Mwangi *et al.*, 2013). For example, when a high classification accuracy is achieved e.g. when predicting diagnosis between a psychiatric disorder group and healthy control group, the brain regions identified through feature selection are regions that support high accuracy discrimination – thus potentially identifying a biomarker of the psychiatric disorder. In addition, the use of feature selection may to produce a more accurate predictive model when data that are not relevant to the predictive model or noisy data are removed (De Martino *et al.*, 2008; Guyon and Elisseeff, 2003; Mwangi *et al.*, 2013).

There are many different types of feature selection techniques e.g. supervised/unsupervised methods, univariate/multivariate methods. These are described in more detail in Chapter 2. There are specific drawbacks to each method of feature selection but the more general drawbacks include additional parameters to optimise, increased chance of overfitting (particularly in multivariate feature selection methods) and failure to detect the relevant features through univariate or unsupervised feature selection (Guyon and Elisseeff, 2003).

Two psychiatric disorders are investigated in this thesis: Attention Deficit Hyperactivity Disorder (ADHD) and Major Depressive Disorder (MDD). ADHD core symptoms include difficulty in sustaining attention, hyperactivity and impulsive behaviour. It affects around 5% of the general population under the age of 16 (Polanczyk *et al.*, 2007) and is the most commonly diagnosed psychiatric disorder in children.

MDD is a mood disorder that is associated with persistent and disabling symptoms of low mood, anhedonia, hopelessness, guilt, low self-worth, poor concentration, lack of energy, suicidal thoughts and altered appetite and sleep (American Psychiatric Association, 2000) with no established pathophysiological mechanisms or biomarkers. Unipolar major depression was ranked in the top ten

diseases for global disease burden in 2001 (Lopez *et al.,* 2006) and is estimated to be in the top three leading causes of burden by 2030 (Mathers and Loncar, 2006).

Unfortunately, despite a wealth of convergent evidence that both ADHD and MDD are disorders with strong biological underpinnings, it has not yet been possible to identify reliable diagnostic or prognostic biomarkers (Coghill and Banaschewski, 2009; Coghill *et al.,* 2008; Schneider and Prvulovic, 2013). As a consequence, the diagnostic process remains dependent upon clinical history and rating scales.

In this thesis, the main machine learning predictions were made using a patient group and matched control group, to provide a proof of concept that these techniques can be applied to clinical studies and to investigate which brain regions differentiate each disorder from controls. These machine learning approaches can also be applied to more clinically relevant questions, such as prediction of response to clinical treatment. This has the obvious advantages of potentially providing a reliable predictor of treatment response prior to exposure of the medication. Furthermore, it has the potential to increase the understanding of the mechanisms which underlie treatment response.

Chapter 2 provides an overview on the methods used throughout this thesis with an emphasis on the main machine learning method used, the Support Vector Machine (SVM). Following this, Chapter 3 discusses studies that have applied machine learning methods to neuroimaging, with a focus on ADHD and MDD studies. Chapter 4 describes a preliminary investigation into whether sociodemographic, clinical and neuropsychological measures can be used to predict treatment response in ADHD using machine learning methods. Chapter 5 explores whether ADHD and control subjects can be accurately classified using structural MR images and machine learning. An ongoing study that involves scanning medication-naïve children and adolescents with ADHD and healthy controls, with the ADHD group beginning a trial of medication after the scan, is outlined in Chapter 6. Whilst this study did not have enough subjects for an analysis due to unforeseen delays in the study outwith the control of the author, the work involved in this study and the planned future work are described, such as the prediction of both diagnosis and medication response. The results obtained when attempting to predict MDD vs. healthy controls using structural MRI is outlined in Chapter 7. Chapter 7 also outlines the results obtained when MDD patients' symptom severity scores were predicted on the basis of their structural images. Chapter 8 describes the application

of machine learning methods to functional MR images to see if it is possible to classify MDD patients and healthy controls, on the basis of their brain activity, when receiving rewarding (‘win’) and aversive (‘loss’) stimuli. Finally, the overall results are briefly summarised and potential directions for future research discussed in Chapter 9.

Chapter 2: Methods

This chapter provides a review of all the methods used in the experimental studies described in this thesis. Methods developed by the author will be highlighted.

2.1 Neuroimaging data quality

Structural and functional neuroimaging studies of psychiatric disorders, such as ADHD and MDD, aim to either identify subtle average differences at a group level in comparison to controls, or use multivariate techniques aimed at predictively classifying individual subjects. As the signal/noise (signal = clinical features being tested for, noise = random and systematic confounds) ratio in MRI scans is low, it is very important to make sure one avoids the introduction of gross imaging artefacts, such as blurring due to movement and blood flow artefacts likely due to a sub-optimal choice of image parameters.

Gross artefacts have two potential effects. First, they can substantially increase inter-subject variance and so obscure clinical features of interest. Second, and more problematic, both the number and nature of gross artefacts are less likely to be balanced when comparing healthy controls with populations that may be unable to tolerate scanning such as those with Parkinson's disease or ADHD. As a consequence, significant between-group differences (or individual-subject classification methods) may be driven by artefact rather than clinically relevant differences. In essence, differences in the number and nature of artefacts between groups (e.g. ADHD vs. control) could result in misleading results whereby between-group artefact differences are misinterpreted as syndromal differences.

The situation for subjective assessment of scans is somewhat different. Clearly it is still important to have good quality scans, but as radiological reporting can only detect relatively high signal/noise abnormalities, radiological reporting is less susceptible to such artefacts. This means that the quality of an image may be acceptable for radiology purposes, but not for inclusion into a quantitative neuroimaging study.

Examples of various types of gross artefact affecting a high proportion of child and adolescent scans from across a number of different neuroimaging labs can be found in the on-line ADHD-200 downloadable dataset:

http://www.incf.org/community/competitions/adhd-200-global-competition.

Examples of commonly occurring artefacts from the ADHD-200 dataset are shown in Figure 2 and include: a blurred image - likely due to movement, part of brain missing - probably due to poor positioning of the subject relative to the head coil, blood flow artefacts - likely due to sub-optimal choice of image parameters, and a 'wrap' artefact - most likely due to a field of view that was too small. A readable account of MRI physics imaging artefacts and their avoidance is available (McRobbie *et al.*, 2010). It should be noted that there are many other sources of artefact including static magnetic field inhomogeneities and radiofrequency coils not producing the intended pulse shapes (McRobbie *et al.*, 2010). Establishment of a good quality assurance (QA) program for neuroimaging research sites is therefore critical. Additionally, computational methods to detect some types of scanning artefacts http://www.fil.ion.ucl.ac.uk/spm/ext/ have been described.

Every scan should be visually inspected for gross artefacts, before applying quantitative imaging methods, irrespective of the size of the dataset. It is reassuring if artefacts are rarely found in a dataset that has been acquired with a stringent QA program in place. In contrast, data that has been acquired without QA and that contains a high proportion of gross artefacts is, arguably, uninterpretable. Note, however, scans that do not contain gross artefacts may have more subtle artefacts that could still result in misleading findings. Examples of subtle artefacts include the susceptibility artefact (an artefact due to the proximity of air filled spaces in the head to various brain regions), localised blurring and flow artefacts (more subtle in spin-echo sequences) (McRobbie *et al.*, 2010).

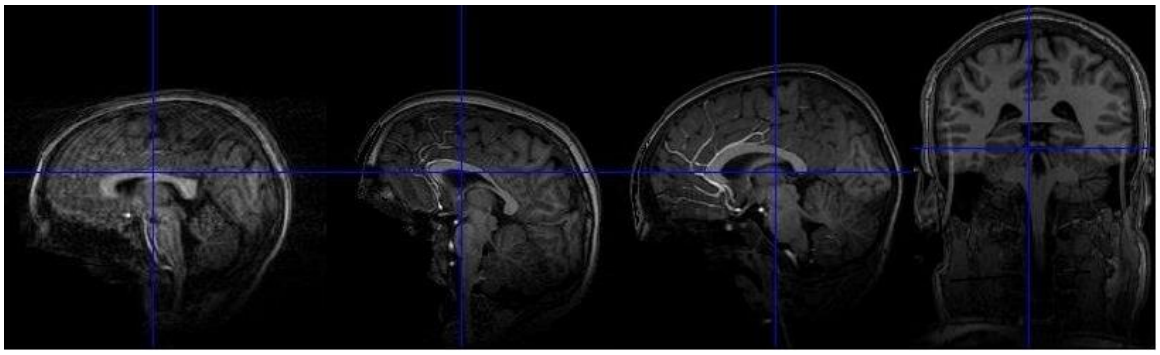

Figure 2: Examples of poor quality images from ADHD-200 dataset. From left to right: Blurred image artefact most likely due to movement, cerebellum not fully scanned most likely due to patient positioning within the head coil, blood flow artefacts most likely due to sub-optimal image parameters, and wrap artefact most likely due to a field of view that is too small. Such gross artefacts will obscure subtle syndromal differences in brain structure.

2.2 Pre-processing

The first step used to analyse all structural or functional MRI (fMRI) datasets in this thesis (after checking data quality) was image pre-processing.

MR images consist of many voxels (a unit of image information defining a region in three dimensional space – similar to a pixel, but in 3D). The primary goal of normalisation is to make images comparable on a voxel-by-voxel basis. There are various pre-processing techniques that can be used to normalise MR images such as Statistical Parametric Mapping (SPM - http://www.fil.ion.ucl.ac.uk/spm/) and the FMRIB (Functional MRI of the Brain) Software Library (FSL - http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). All imaging studies described in this thesis used SPM8 for pre-processing. The process of normalising each brain on a voxel-by-voxel basis and then comparing each image volume between subjects at every voxel is called Voxel-Based Morphometry (VBM).

All raw MR images acquired from MR scanners in this study were in DICOM format. Before image pre-processing can take place, the images were converted from DICOM to Analyze format using MRIConvert (http://lcni.uoregon.edu/~jolinda/MRIConvert/).

Once the data was in Analyze format, the images were spatially normalised. This is an important stage to allow inter-subject comparisons (or in the case of fMRI data, which consists of a time series of images for each subject, intra-subject analysis followed by inter-subject testing) (Friston *et al.*, 2007).

2.3 Structural MRI Pre-processing and DARTEL

The standard approach to pre-processing structural (e.g. T_1 or T2 weighted) MR images include: segmentation, spatial normalisation and smoothing. A newer technique for pre-processing structural MRI images known as 'DARTEL' (Diffeomorphic Anatomical Registration using Exponential Lie Algebra) (Ashburner, 2007) has become available and is implemented in SPM8 (Friston *et al.*, 2007).

The first step when using either the default SPM spatial normalisation routine (Ashburner *et al.*, 2012) or DARTEL involves "segmenting" the images. Segmentation involves aligning each image and estimating the probability of grey

matter, white matter or cerebrospinal fluid (CSF) within each voxel in a brain scan. The procedure produces three separate images containing the respective voxel probabilities. As an example, a structural MR image of a healthy control is shown in Figure 3 (left) and the corresponding segmented grey matter image is shown in Figure 3 (centre).

When spatially normalising using the default SPM method, it is usual to rescale images such that they match a standard anatomical template. SPM includes a number of templates including the Montreal Neurological Institute (MNI) template (Mazziotta *et al.*, 1995). This was created by averaging a large number of normal adult brain scans scaled to the Talairach Atlas (Talairach and Tournoux, 1988). However, use of a standard adult spatial normalisation template is problematic in studies of children and adolescents and could also be an issue when investigating subjects with gross structural brain abnormalities, as the brain structure of these populations is not the same as a healthy adult's brain structure.

To address this issue, the DARTEL method introduced a stage in the analysis, following segmentation, involving the creation of a 'study-specific brain template'. This template is then used in place of a standard adult template (such as the MNI template). Prior to the introduction of DARTEL it was still possible to create study-specific templates (Good *et al.*, 2001) but this was not always done. The DARTEL method employs a sophisticated normalisation technique that involves 'spatial normalisation', 'warping' (non-linear geometric transformations of the image) and a choice whether to include a 'modulation' step (Ashburner, 2007).

The decision as to whether to include modulation is important. Modulation is used to control the effect of volumetric differences that would otherwise occur during spatial normalisation. SPM uses the terms "Preserve Amount" to describe the inclusion of modulation during normalisation and the term "Preserve Concentration" to describe normalisation without modulation.

When "Preserve Amount" is selected (modulation applied) the regional and global intensity is preserved. In this case, when a region's volume is increased during warping, the region's tissue (e.g. grey or white matter) probability is proportionally reduced – thus the probability within the region remains constant. This means the total amount of matter present in the original and the normalised images should be identical.

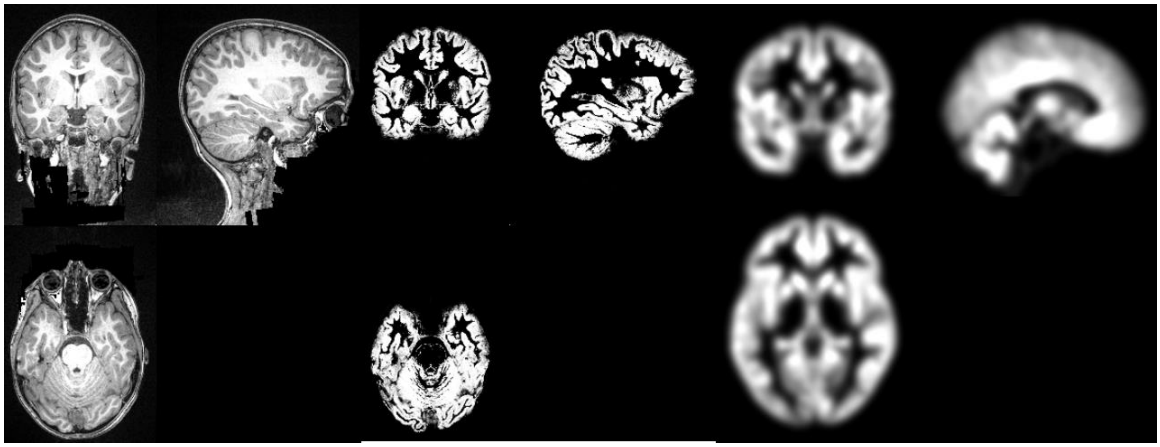

Figure 3: Three stages of pre-processing a structural MRI image of a healthy control. Left: A ‘raw’ structural MRI image. Centre: A segmented grey matter image. Right: A DARTEL-processed, smoothed grey matter image.

Conversely, when “Preserve Concentrations” is used (no modulation applied) the concentration (ratio of intensity and volume) of the tissue probabilities in the original image is preserved. Therefore, when a region’s volume is increased the probability is increased accordingly such that the ratio remains constant.

Ashburner (2012) advises that “Preserve Amount” (modulation) should be used when normalising structural MR images but “Preserve Concentration” (omitting the modulation step) should be chosen when normalising functional MR images. However, this is simply advice and there are no definitive conclusions as to which normalisation method is best. There are currently no studies that compare these two approaches.

In summary, the SPM DARTEL technique aligns the images by creating a mean image for both grey matter and white matter segmented images (with an option to perform the equivalent calculation on the CSF segmented images). It then involves transformations to deform each individual image towards a study-specific template. The newly deformed images can again be averaged to create an anatomically 'sharper' template and the process repeated.

Another advantage of DARTEL is that the warping method preserves more of the image than previous warping methods (Ashburner *et al.*, 2012). In essence, the key difference is that instead of fitting to the template by ‘pulling’ voxels from the original image, DARTEL ‘pushes’ each individual voxel in the original image to fit to the template. However, this method has a potential drawback as it tends to introduce aliasing artefacts (which appear as grid lines across the entire volume) in the images unless they are smoothed (Ashburner *et al.*, 2012).

The final step of both the default SPM8 normalisation and DARTEL is to smooth the images with a Gaussian kernel, to remove artefacts (e.g. aliasing effects) and optimise the signal/noise ratio (Friston *et al.*, 2007). Figure 3 (right) shows an example of a DARTEL-processed image. The mathematics implemented in DARTEL is beyond the scope of this thesis but a detailed description is provided by Ashburner (2007).

The result of DARTEL pre-processing is a series of brain images that are all accurately aligned to a study-specific template, with a given anatomical coordinate location in one image corresponding to the same anatomical coordinate location in the other subjects’ images.

Following pre-processing of structural MRIs, it is usual to perform conventional univariate group level image analysis to test null hypotheses of no significant effect using SPM (e.g. t-tests, linear regression). The pre-processed images can also be used for individual subject predictive modelling using multivariate techniques. The analyses that may be performed after pre-processing are discussed from section 2.6 to the end of Chapter 2. As DARTEL is a relatively new technique, released in 2007 (Ashburner, 2007), group-level analyses in the literature have mostly used standard normalisation pre-processing methods.

2.4 fMRI Pre-processing

fMRI pre-processing consists of: slice timing correction (optional in some circumstances), realignment, coregistration (optional), normalisation and smoothing.

Slice timing correction attempts to compensate for differences in slice acquisition times. This correction is required if it is important to correct for each complete image consisting of slices taken at slightly different times, although this stage is often omitted as there is a debate regarding the effectiveness and importance of the slice timing correction (Sladky *et al.*, 2011).

The realignment ensures all images within the fMRI time series are oriented to a selected reference image. This is usually done by either aligning all scans to the first image or, more commonly, aligning to the mean image across the time series. The purpose of realignment is to remove movement related artefacts which can occur during an fMRI scan. The amount of realignment required for each image is recorded and is also used as a covariate 'of no interest' during first level fMRI analyses.

The coregistration step is not always used during standard fMRI normalisation. When used in standard fMRI normalisation, coregistration is used to register each subject’s realigned fMRI volume to a structural scan from the same subject. Coregistration then involves spatially normalising the structural scan (which in theory might be done more accurately than spatially normalising the mean functional scan) and applying the spatial normalisation parameters to the functional scans.

The normalisation and smoothing processes are identical in principle to the default SPM approach for structural scans: i.e. it involves spatially normalising all images towards a template image with a final step of smoothing the images.

The DARTEL method can also be applied to fMRI data when each subject has a corresponding structural MRI of high quality. The approach to normalising fMRI images using DARTEL involves performing the structural MRI DARTEL calculation (as described in section 2.3), coregistering each subject's fMRI volume to the corresponding original structural MR image, then using the flow field created during each subject's structural normalisation to warp the fMRI images towards the study-specific template created during the structural MRI DARTEL process. Unfortunately, the author found, through visual inspection of the normalised images, that this approach did not provide as accurate normalisation as the standard method for fMRI data as a number of brain landmarks were positioned and oriented incorrectly in comparison to the template image in many of the subjects. It is unclear what caused this poor normalisation but as the lower resolution fMRI images do not have as clearly defined landmarks compared to structural images, the error most likely occurs during coregistration towards the structural images or when using the flow fields generated during the structural normalisation to normalise the fMRI images. Accurate normalisation is essential to all neuroimaging analyses in this thesis in order to ensure each voxel corresponds to the same brain region in all subjects so the standard fMRI normalisation was used throughout.

2.5 fMRI first level analysis

When analysing fMRI data from multiple subjects, using a random effects design, the analysis initially takes place at the "first level" which is a within-subject analysis. This is followed by a "second level", between subjects, analysis using summary images generated from each subject from the first level analysis. Structural MRI analysis does not involve an analogous first level analysis. It is essential in fMRI studies as it extracts and summarises the relevant information from the total fMRI volume for each subject to create a contrast image.

SPM performs statistical analysis of fMRI data using a mass-univariate approach based on the General Linear Model (GLM). The standard approach for task-based fMRI involves: extraction of timing parameters and task information from the logfiles to create a design matrix, the estimation of the parameters for the GLM and the analysis of the results obtained from the relevant contrast vectors.

All fMRI first level analyses undertaken by the author have used an event-related, random effects design. The first level model was created by identifying relevant time points for the contrast(s) of interest, creating a truncated delta function at each time point, which is then convolved with the haemodynamic response function (Ashburner *et al.,* 2012).

2.6 Second level analysis

A comparison between subjects (or groups of subjects) is defined as a second level analysis, using the GLM to implement t-tests, correlations, etc. As mentioned previously, between subject analyses can be done on structural MRI data directly after pre-processing and also on the output from a first level analysis using fMRI.

2.7 Neuroimaging data quality - Outlier analysis

A key idea to analyse both the quality of the normalisation and the quality of each subject's images in general was an outlier analysis. This involved comparing each voxel in a given subject's spatially normalised structural scan with the corresponding voxels from other subjects spatially normalised structural scans (within the same group: e.g. patient or control).

Using a boxplot function, outliers were defined as data which were located outwith the "whiskers" of the boxplot. The default setting for the maximum whisker length, *W*, was 1.5 (Tukey, 1977), this corresponds to approximately 2.7 standard deviations (99.3% coverage if the data are normally distributed). This meant points were identified as outliers if they fell outwith the range Q1-W*(Q3-Q1) and Q3+W*(Q3-Q1), where Q1 and Q3 are the 25th and 75th percentiles, respectively.

The threshold that was empirically selected for an acceptable percentage of voxels that were identified as outliers across the whole brain was 10%. This relatively lenient threshold allowed subjects to have a small proportion of the brain defined as an outlier (in comparison to the same diagnostic group) yet was low enough such that structural anomalies in anatomy or preprocessing issues could be identified in individual subjects.

2.8 Multivariate Pattern Analysis

Multivariate pattern analysis (MVPA) is a term used to describe a range of approaches to pattern recognition. These techniques are being applied in various fields as they are able to detect subtle but consistent patterns of differences between groups at an individual subject level.

MVPA can be separated into two groups, supervised learning and unsupervised learning (Bray *et al.*, 2009; Shawe-Taylor and Cristianini, 2004). The difference between these two approaches is that supervised learning 'learns' using data with known class labels (training data) before predicting the class of previous unseen data (testing data), whereas unsupervised learning attempts to learn without being provided with group labels (Theodoridis and Koutroumbas, 2006). In this thesis only supervised learning is discussed in detail. For more information on unsupervised learning see Theodoridis and Koutroumbas (2006).

Overfitting, a failure to generalise from training data when the trained classifier is presented with novel data, is a common issue in machine learning studies which results in poor predictive accuracy. An excessively high accuracy occurs during the training stage as the model fits to 'noise' (random variation) within the training data, resulting in misleadingly high training stage predictive accuracies, but an inability to generalise to novel data not used for training, resulting in inaccurate predictions. Cross-validation is commonly used to address this issue, which involves separation of data into training and testing groups. This means that accuracies obtained using cross-validation are reported only from the novel data, which therefore avoids reporting inaccurately high training stage prediction accuracies if overfitting had occurred.

Cross-validation also avoids another potential issue in MVPA studies, 'double-dipping' - when prior knowledge about the testing data set is 'leaked' into the training process. It is essential to keep the training and testing stages separate during cross-validation (Kriegeskorte *et al.*, 2009). Consequently, any process occurring before MVPA takes place which could provide prior information about the testing set if performed on the full dataset (e.g. feature selection – described in section 2.14) must be performed on the training set only.

Due to the relatively small numbers of scans in typical neuroimaging studies, 'leave-one-out cross-validation' (LOOCV) is popular (Cristianini and Shawe-Taylor,

2000) as it maximises the number of the training data available. This approach involves selecting all but one of the subjects (from either group) for training and classifying the 'left out' subject in the testing stage. The selected 'left out' subject is then reintroduced to the training group and another subject removed. The process is repeated until all subjects have been left out once.

The leave-one-out method is popular as it maximises the number of subjects in the training set. It does however take considerably more time to compute as the MVPA model must be optimised for each 'left-out' subject. Furthermore, it takes even longer if feature selection is also used, as the feature selection method must be repeated on each training set (i.e. without each 'left-out' subject), in order to avoid double-dipping.

Other common cross-validation methods include two-fold and ten-fold cross-validation which involve partitioning the data into two or ten equally sized subsamples (or 'folds') and training on all folds except one, predictively classifying the left-out partition. These approaches are very useful for large datasets such as those obtained in genetics studies as it reduces the time for calculations. LOOCV is equivalent to *N*-fold cross-validation (where *N* is the number of subjects in the dataset) and is mainly used in smaller datasets to maximise the number of subjects in the training set.

As mentioned previously, overfitting is a result of fitting a pattern recognition method too closely to the training data such that it cannot generalise to novel testing data. There are a whole host of MVPA techniques but, to the author's knowledge, all use one or more parameter which alters how closely the model fits the training data. These parameters, with respect to pattern recognition methods, are discussed in section 2.10.

When overfitting occurs, the final classification accuracy for novel data is poor. A common technique to reduce overfitting is to use a *second* cross-validation process within the training set. For example, LOOCV is commonly applied within each training stage to identify which MVPA parameters (and potentially feature selection parameters, discussed in section 2.14) achieve the best results within the training set. Once the optimal parameters are found within the training set, it is then possible to apply these parameters in the classification of the previously unseen test set. The application of cross-validation within the training set allows the parameters to be tested on data not used during training to give an indication of the

generalisability of each combination of parameters prior to classification of the novel testing data. Provided there are no major differences between the training set and the testing set, this method reduces the likelihood of fitting the model too closely to the training data.

Another important factor, which must be taken into account, is the need to ensure there are no significant confounding group level differences in the data (e.g. age, gender, IQ). For example, if the patient group were significantly older than those in the healthy control group, the pattern recognition method might base its predictions on subtle features of older brains for patients and younger brains for controls. This could result in substantial errors in prediction when scans from an older healthy control, or a younger patient are analysed. Similarly, if gross artefacts are more frequent in one group (e.g. movement artefacts or scanner artefacts due to poor data quality assurance) compared to another group, the classifier may base its prediction methods on artefacts and not syndrome-related differences in brain structure.

Stratification/data selection is a practical approach to non-artefact confounds as it is not currently easy to 'covary out' such confounds, as would be done in traditional univariate group level null hypothesis testing (ANCOVA etc) (Linden, 2012; Watkins *et al.,* 2009).

The number of subjects within each group is also an important factor in MVPA analyses as many of the MVPA techniques are sensitive to what is known as the 'class imbalance problem' (Mourão-Miranda *et al.,* 2011; Theodoridis and Koutroumbas, 2006). Typically, class imbalance is only an issue when there are large imbalances between the groups. The class imbalance problem tends not to be an issue when using large datasets and classifying well separated groups. However, neuroimaging studies tend to have a small number of subjects and the groups are generally not easily separated. In the most extreme cases (such as where there are a small number of subjects and many 'noisy' or uninformative voxels – thus making the groups less separable), a few subjects' difference between groups could make the numerically larger group considerably more likely to be predicted (Theodoridis and Koutroumbas, 2006). When there are many voxels included in a classification, the accuracy may be less than the expected 50% chance performance as too many noisy or irrelevant voxels are included when attempting to train the SVM, provoking the class imbalance problem. Similarly, when very few voxels are included, the accuracy

may drop due to too few voxels which support the classification. In these cases feature selection is crucial to ensure the classification is based on an optimal fit to the underlying pattern – making class imbalance less important.

2.9 Support Vector Machine

This section will discuss the mathematical principles of SVM. The SVM toolbox used in this study was written by Anton Schwaighofer (Schwaighofer, 2001) and part of the Piotr Dollar's Image and Video Matlab Toolbox (Dollar, 2011) (http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html - V2.4 and V2.6).

In general, SVM can be described as a two-class pattern recognition technique. During training, the SVM method uses a kernel function to construct a hyperplane ('decision boundary') that best separates the two groups. During testing, predictive classifications using novel (not used for training) data are done by identifying which side of the hyperplane a given novel datum lies.

Given N subjects in a training set (the set of subjects used for identifying the optimal hyperplane), then $\{\boldsymbol{x}_i, y_i\}$ where $i = 1 \ldots N$, $\boldsymbol{x}_i$ represents a vector for each subject (i.e. the selected voxels from a structural MR image), and y_i represents a subject's group label (e.g. -1 or 1 – where class labels are arbitrarily assigned to each value). The hyperplane can be described using the equation $\boldsymbol{w}.\boldsymbol{x} + b = 0$, where $\boldsymbol{w}$ is normal to the hyperplane and $b/\|\boldsymbol{w}\|$ is the perpendicular distance from the hyperplane to the origin.

When, for example, a linear kernel is being used to classify groups which are linearly separable, there exists a vector $\boldsymbol{w}$ and a scalar b such that the inequalities:

$$\begin{aligned} w.x_i + b &\geq 1 \quad if\ y_i = 1 \\ w.x_i + b &\leq -1 \quad if\ y_i = -1 \end{aligned} \qquad \text{Eq. 1}$$

are consistent throughout the training set. In order to classify data which is not linearly separable Cortes and Vapnik (Cortes and Vapnik, 1995) introduced a 'slack variable', ξ_i (where $i = 1 \ldots N$), which variably penalises misclassified data:

$$\begin{aligned} w.x_i + b &\geq 1 - \xi_i \quad if\ y_i = 1 \\ w.x_i + b &\leq -1 + \xi_i \quad if\ y_i = -1 \\ \xi_i &\geq 0\ \forall_i \end{aligned} \qquad \text{Eq. 2}$$

The above equations can be combined:

$$y(w.x_i + b) - 1 + \xi_i \geq 0 \quad where\ \xi_i \geq 0\ \forall_i \qquad \text{Eq. 3}$$

The optimal hyperplane is identified as the boundary that minimises classification errors and maximises the 'margin' (defined as the shortest distance between the hyperplane and the closest subject) separating two groups. To allow the SVM approach to cope with datasets that are not perfectly separable (datasets whereby no choice of hyperplane could perfectly separate the data without classification errors) Cortes and Vapnik (1995) introduced the "slack variable". In a simple classification problem the classes are linearly separable and the, slack variable ξ_i (where i = 1…N) can be omitted. However, the slack variable 'soft-margin' means that classification can be performed even when there are data (subjects) located on the incorrect side of the hyperplane, as it acts to variably penalise each misclassified data as a function of distance from the hyperplane (Cortes and Vapnik, 1995; Fletcher, 2009). Provided the conditions of Eq. 3 are satisfied, the hyperplane is identified by minimising:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \qquad \text{Eq. 4}$$

The parameter *C* corresponds to the soft-margin parameter which requires optimisation when applying linear and non-linear SVMs, as outlined in section 2.10. A full derivation of the mathematics involved in SVMs can be found elsewhere: Bishop (2006), Cristianini & Shawe-Taylor (2000), Fletcher (2009).

Once the optimal hyperplane has been identified from the training data (e.g. a suitable soft-margin parameter has been selected - and also any additional parameters - such as the kernel width parameter - if using a non-linear SVM), the novel testing data can then be classified. SVM kernels are either linear or non-linear, with the latter often achieving higher classification accuracy (Song *et al.*, 2011). Using a linear kernel, a new subject $\{\boldsymbol{x}^*, y^*\}$ is classified by:

$$f(x^*) = \sum_{i=1}^{N} w_i x_i^* + b \qquad \text{Eq. 5}$$

where the sign of $f(\boldsymbol{x}^*)$ determines which class label the subject is predicted to have (i.e. $f(\boldsymbol{x}^*)$ -the predicted class label- is compared with y^* - the true class label - in order to determine whether that subject's data has been correctly classified.

The main equation for SVM (and Relevance Vector Machines; RVMs) classification is:

$$f(x) = \sum_{i=1}^{N} w_i K(x, x_i) + b \qquad \text{Eq. 6}$$

where $K(x,x_i)$ describes the kernel function (e.g. linear, polynomial, radial basis function, etc) (Bishop, 2006). In this study results have been reported for both a linear kernel (as shown in Eq. 5) and a Gaussian kernel (Eq. 7). There are other kernels such as a polynomial kernel and sigmoid kernel but the linear and Gaussian kernels were selected as they are commonly used in the literature (Fletcher, 2009; Shawe-Taylor and Cristianini, 2004).

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \qquad \text{Eq. 7}$$

2.10 Application of SVM

Training a *linear* SVM involves varying one parameter, the soft-margin parameter, during the training stage of cross-validation. The soft-margin parameter, C, determines how closely the SVM 'tries' to fit to the training data; a low C value allows a small number of misclassifications in the training set to obtain the best predictive model whereas a high C value allows no misclassifications in the training set, potentially at the expense of the generalisability of the model. Therefore, if a large soft-margin parameter is selected then it is likely to lead to overfitting during training. Training a *non-linear* SVM involves optimising two or more parameters depending on the choice of non-linear kernel. For the most popular non-linear kernel functions (e.g. Gaussian, radial-basis function and polynomial kernels), two

parameters require optimisation (the soft-margin parameter and an additional ‘kernel width’ parameter, the latter having an effect of scaling the non-linear kernel function). There are non-linear kernels that require the optimisation of additional parameters, such as the sigmoid function but these are not discussed in this thesis. Typically, the more parameters that require optimisation, the longer the optimisation takes during training.

During cross-validation, for each prediction of a left-out subject, all parameters must be 'tuned' using the training data set prior to predictive classification on the test data. In the studies outlined in this thesis, parameter selection was performed using an inner LOOCV procedure *within the training set* with all possible parameters tested using a ‘grid search’ procedure. The optimal combination of variables was selected based on the highest ‘training accuracy’ during the inner LOOCV. It is important to distinguish the *accuracy during training* (obtained from the inner LOOCV), which is used to guide parameter selection, from the *true predictive accuracy achieved during the testing stage* with *novel* data. Only the true predictive accuracy achieved during the *testing* stage is reported here.

The combination of parameters that achieved the highest training stage accuracy is then applied during *testing* (i.e. when classifying the novel ‘held-out’ subject), to assess classifier performance using conventional methods: accuracy, sensitivity, specificity and chi-square significance of classification.

SVM is one of the most popular supervised learning MVPA techniques, used most often in this thesis as it typically gives good classification performance (Craddock *et al.,* 2009) and can be described relatively intuitively, without extensive mathematical presentation. Other methods such as one-class SVM, RVM and Gaussian Processes are briefly discussed in section 2.13.

As an illustration of the general SVM approach, suppose MRI data from 10 patients and 10 healthy controls have been pre-processed (in general, more subjects would be required to achieve high accuracy using this technique) and, as the technique is easiest explained with a two dimensional example, predictions are made based on data from only two voxels without feature selection and a linear SVM (to facilitate description).

As mentioned previously, the classification process involves two stages: training, when the SVM ‘learns’ from the training set, and testing, where the testing set is classified (e.g. predicting whether each scan in the testing set belongs to the

patient or control group). In experimental studies, these predictions are then compared to the actual group membership to determine the level of success of the classifier.

In our example LOOCV is used. Therefore, 19 of the 20 MRI images (9+10) were selected for SVM training; the 'left out' subject's diagnostic status was then predicted in the SVM testing stage. This process was repeated until all 20 images were predictively classified.

Considering now the SVM ***training*** stage in more detail, a way to visualise SVM training is a plot of the voxel value (e.g. grey or white matter probability value) from each subject in the training set. In our two voxel example, this corresponds to a two-dimensional plot. As the diagnostic status of each subject in the training set (all subjects other than the 'left out' subject) is known, the SVM training stage identifies a line which best separates the two groups - the 'decision boundary' (or 'hyperplane') (Bishop, 2006), by optimising the SVM parameters. Once the optimal decision boundary has been estimated, the SVM training is complete. In our example, consider the 'left out' subject being a scan from a patient; this leaves 10 controls and 9 patients in the training set. As this example is only considering two voxels (and not hundreds of thousands), the respective grey matter voxel probabilities can be plotted on a two-dimensional plot as described above. An illustration is shown in Figure 4 (left) where the decision boundary is the central line. When applying SVM to real data, the number of voxels required to achieve a high classification accuracy is obviously going to be larger than two and separation is usually never as straightforward as this example, as the SVM typically must allow for errors in classification, to find the best fitting decision boundary.

The SVM ***testing*** stage involves plotting the 'left out' subject's voxel probability values on the plot created in the SVM training stage. The prediction of which group the 'left out' subject belongs depends on which side of the decision boundary the subject is found. In our example, the subject is predicted to belong to the patient group as the 'left out' subject is plotted above the decision boundary (as shown in Figure 4 (right), with the 'left out' subject represented by a star). If the data point was plotted below the decision boundary, the scan would be predicted to be from a subject belonging to the control group. When more voxels are included in the SVM analysis the situation becomes complex to visualise, because the number of voxels included in the prediction equates to the number of dimensions the decision

boundary must be calculated over, but the basic mathematical concepts remain the same.

By separating the total training set into a number of subsets it allows the decision boundary to be tested on various subjects, before the SVM testing stage, in order to make the decision boundary as generic as possible (thus avoiding overfitting).

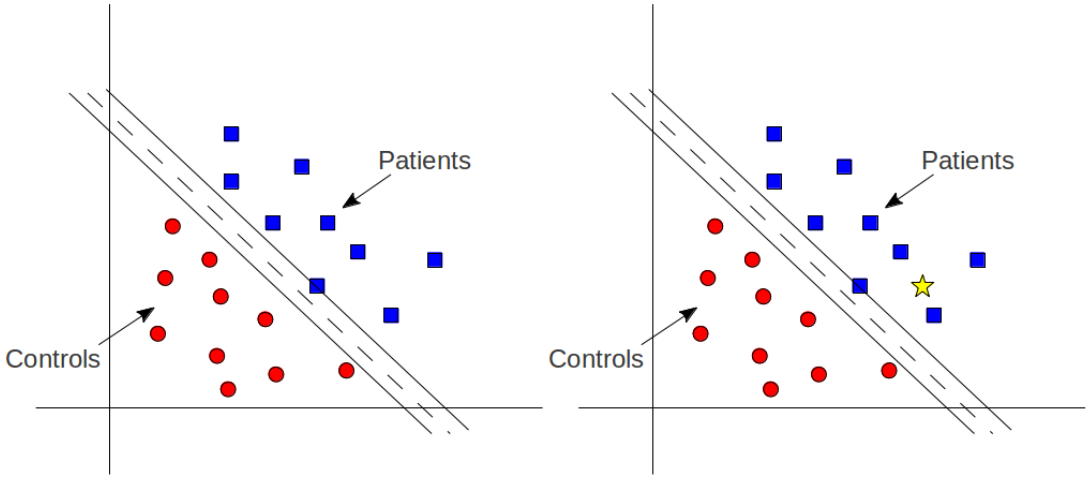

Figure 4: A summary of the SVM process. If there are 10 patients and 10 healthy controls suppose the ‘left out’ subject in this example of the LOOCV step is a patient. The training set therefore consists of 9 patients and 10 healthy controls. In this example only two voxels from the brain are selected to simplify the description of SVM principles. Left: SVM training – A plot of the two selected features of each subject in the training set (patients represented by blue squares and healthy controls represented by red circles). The best fitting decision boundary is calculated from the subjects in the training set. Right: SVM testing – The ‘left out’ subject (represented by the star) is plotted and the prediction depends on which side of the decision boundary the subject lies. In this case the ‘left out’ subject scan lies above the decision boundary so would be predicted to be a scan of a patient (if below the decision boundary it would be predicted to be from a control).

2.11 Relevance Vector Machine

This section will briefly outline the RVM method. The RVM algorithm was written by Mike Tipping (Tipping, 2001) and in this study the algorithm programming interface was created from a section of PRoNTo (Pattern Recognition for Neuroimaging Toolbox) Matlab code (Schrouff *et al.,* 2013).

Although the Bayesian learning framework on which RVM was built on is far more general, the RVM method was designed to follow the same functional form as SVM. Therefore, RVM also relies on the selection of a kernel and shares the same basis equations as SVM such as Eq. 6.

$$f(x) = \sum_{i=1}^{N} w_i K(x, x_i) + b \qquad \text{Eq. 6}$$

The main difference between RVM and SVM is that RVM employs a Bayesian framework to make predictions (Tipping, 2001). This means that there are several advantages of the RVM algorithm (compared to SVM). First, it can output probabilistic predictions. The standard SVM algorithm can only provide a binary output, whereas RVM can provide a level of confidence or uncertainty in each prediction. Second, and most importantly for the work undertaken in this thesis using the RVM algorithm, RVM is able to predict continuous variables, this technique is called Relevance Vector Regression (RVR). The RVM algorithm also has other advantages over SVM such as it automatically estimates the soft-margin parameter, kernel functions do not have to satisfy Mercer's condition, and the method tends to be sparser.

The RVM classification method was investigated to see if it could improve on the results obtained using SVM. However, despite all the advantages of RVM, the SVM algorithm has been found to be *more robust* by the author as it achieved consistently better results compared to RVM. For this reason RVM classification has not been discussed in more detail in this thesis - detailed description can be found elsewhere (Bishop, 2006; Tipping, 2001). Nevertheless, the RVM algorithm was found to accurately predict continuous variables, which can be used to predict clinically relevant information, such as symptom severity.

The RVM code is freely available (http://www.vectoranomaly.com/downloads/downloads.htm) and the PRoNTo

toolbox includes this algorithm amongst their pattern recognition tools and created an interface to implement the code. Custom Matlab code was written by the author to interface with the PRoNTo RVM toolbox.

Another addition to the RVM toolbox was the use of 'sensitivity map' calculations with non-linear kernels. When applying a linear kernel to SVM or RVM, the relative contribution of each voxel to the prediction, the 'voxel weight', could be output from the procedure. However, a limitation of these algorithms was that the same approach could not be applied when applying non-linear kernels. However, Rasmussen *et al.* (2011) identified a method to calculate equivalent voxel weights for non-linear kernels. This method was used in this thesis.

2.12 Application of RVR

The RVM toolbox is implemented in an almost identical process as the SVM toolbox. The main difference is that the soft-margin parameter is automatically estimated rather than manually optimised.

When performing RVR, however, the process was altered, as it was no longer possible to optimise the parameters using the training stage accuracies. Instead of using the training accuracy to test how well the model fit the data, the true continuous variables had to be compared with the predicted variables to determine the goodness of fit. There are a number of measures of goodness of fit, eleven of which are included in the gfit2 toolbox (http://www.mathworks.co.uk/matlabcentral/fileexchange/22020-goodness-of-fit-modified/content/gfit2.m). The measures calculated using this toolbox include: the mean squared error, the normalised mean squared error, the root mean squared error (RMSE), the normalised root mean squared error, the mean absolute error (MAE), the mean absolute relative error, Pearson's coefficient of correlation (R), the coefficient of determination, the coefficient of efficiency, the maximum absolute error and the maximum absolute relative error.

Only three of these measures were selected for evaluating the goodness of fit during the training stage of regression studies undertaken by the author: RMSE, MAE and Pearson's coefficient of correlation (R). The RMSE is calculated using the following equation:

$$\sqrt{\frac{\sum_{i=1}^{N}(t_i - y_i)^2}{N}} \qquad \text{Eq. 8}$$

where *N* describes the number of subjects in the training set, *t* describes the target values for the regression model and y describes the predicted values output from the regression model.

Using the same notation the MAE is calculated using the following equation:

$$\frac{\sum_{i=1}^{N}|t_i - y_i|}{N} \qquad \text{Eq. 9}$$

Finally, the sample Pearson correlation coefficient, R, is defined by the equation:

$$\frac{\sum_{i=1}^{N}(t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(t_i - \bar{t})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad \text{Eq. 10}$$

where $\bar{t}$ and $\bar{y}$ represent the mean values for the target values for the regression model and the predicted values output from the regression model respectively. As the Pearson correlation coefficient is a parametric test, the Shapiro-Wilk test was used to test whether each continuous variable followed a normal distribution before attempting to make predictions using RVR.

There were two main methods used to optimise the training stage: a one variable approach and a multi-variable approach.

The one variable approach is similar to the approach taken when comparing training stage accuracies. One of the three main goodness of fit variables highlighted was selected, the goodness of fit measure selected was computed for each combination of parameters (in the same way as the training stage accuracies) and the parameters which produced the best fit (the lowest score for MAE and RMSE and the highest for R) was selected. If more than one combination of parameters shared the optimal score then a 'tie-breaker' decision was required. If the correlation coefficient, R, was not the main comparison variable then this score was used as a tie-breaker, otherwise the MAE score was used. In the case of the same optimal scores being achieved in the tie-breaker also, Ockham’s Razor was implemented. In other words, when presented with two competing models that make exactly the same

predictions, the simpler model was considered better. Therefore the combination of parameters that produced the simplest model was selected (i.e. the lower kernel width parameters produce smoother, more general kernel functions, which tend to avoid overfitting).

When training RVR using the multi-variable approach, the measures selected were normalised (brought within the range 0-1) and combined to make a single score. This approach resulted in very few tie-breaker situations when the three aforementioned variables were combined. In the cases where the same optimal combined score was achieved, Ockham's Razor was implemented as described above.

2.13 Alternative Classifiers

Thus far, only SVM for classification and RVM for classification and regression have been described. The linear SVM is especially intuitive, easy to visualise and often gives good classification performance but has disadvantages: e.g. SVMs do not give a probabilistic output and the standard SVM algorithm cannot perform regression (although extensions to the SVM algorithm include regression (Smola and Schölkopf, 2004; Vapnik *et al.*, 1997) – not discussed further here). Non-linear SVMs are slightly more complicated to conceptualise, but they are very similar to the linear SVM except they attempt to identify a decision boundary described by non-linear functions such as a polynomial function, a radial basis function (RBF) or a Gaussian function, rather than a linear function. As mentioned, the RVM algorithm is very similar to the SVM algorithm as it uses the same basic equations for prediction. As RVM uses a Bayesian treatment to eliminate some of the limitations of SVM (Tipping, 2001), it provides probabilistic predictions and automatically selects the soft-margin parameter, generally making it faster to train. Furthermore, RVM is able to perform both classification and regression. As above though, despite the obvious advantages of RVM over SVM, the SVM algorithm tends to be more robust and consistently successful when performing predictive classification. However, there are many other pattern recognition methods, each with advantages and disadvantages.

An adaptation to the SVM method is the use of one-class SVMs which attempt to identify outliers from a specific class or population (Schölkopf *et al.*, 2001; Shawe-Taylor and Cristianini, 2004). This method will not be discussed in

detail here: for an example of the application of this method to psychiatric research, Mourão-Miranda *et al.* applied one-class SVMs to classify fMRI images of depressed patients as outliers relative to healthy controls (Mourão-Miranda *et al.*, 2011).

Another popular MVPA method is the use of Gaussian processes. As this method has not been used in this thesis, a detailed description and the mathematical underpinnings has not been provided here, see Rasmussen & Williams (2006) and Bishop (2006). Unlike SVM, where one function must be selected (e.g. linear, polynomial, RBF etc), a Gaussian process uses a distribution of every possible function, with higher probabilities for distributions which are more likely (Meyfroidt *et al.*, 2009; Rasmussen and Williams, 2006). Gaussian processes apply Bayesian statistics to perform both classification and regression (Bishop, 2006; Meyfroidt *et al.*, 2009; Rasmussen and Williams, 2006). An advantage of Gaussian processes is that they can classify more than two classes at a time. RVM is a special case of Gaussian processes theory (Rasmussen and Williams, 2006).

As there are many MVPA methods it can be difficult deciding which classifier to choose. The 'No Free Lunch Theorem' suggests there is no single learning algorithm that always produces the highest accuracy (Alpaydin, 2004). A common approach is to try a range of classifiers for a given problem and dataset and select the one that performs the best on a separate validation set (Alpaydin, 2004).

2.14 Feature Selection

The resolution of structural and functional MR images is such that there are typically more than one-hundred thousand voxels in a whole brain scan from a single subject. The signal/noise ratio in MRI scans is low and therefore studies aiming to predictively classify individual scans typically report poor results when voxels from the whole brain are used, due to the large number of noisy and redundant voxels (Bray *et al.*, 2009; Dash and Liu, 1997). As the number of features (or voxels) is very large in neuroimaging data compared to the number of subjects, it is common to reduce the number of features selected for entering into a classifier.

Feature selection has many benefits as it can remove noisy voxels, reduce computation time, and reduce the number of redundant voxels by removing highly correlated voxels and voxels not otherwise useful for classification (De Martino *et*

al., 2008; Guyon and Elisseeff, 2003; Theodoridis and Koutroumbas, 2006). Feature selection can be a critical stage before classification or regression because if features were selected that did not aid prediction, the SVM may be unable to accurately distinguish the underlying pattern. Conversely, by selecting features that identify the pattern well, the likelihood of achieving a highly accurate prediction is increased (Mwangi *et al.*, 2013). In addition, correctly identifying a subset of the most relevant features decreases the risk of overfitting.

Aside from the practical advantage of implementing feature selection (increasing the accuracy of prediction) there is also another major advantage. As the methods identify which features (or in a neuroimaging study, brain regions) are relevant to the prediction, feature selection can identify which brain regions are consistently different between groups in a classification study, or which regions are correlated with certain continuous variables. This means that if a successful prediction can be achieved, the feature selection can potentially provide an insight into structural or functional brain differences that are consistent enough to produce a strong prediction - potentially identifying biomarkers or elucidating brain mechanisms.

As with machine learning methods, there are supervised and unsupervised feature selection methods. However, as with the MVPA methods, only supervised methods are used in this thesis. The methods used in this thesis are outlined in more detail in section 2.15. The 'No Free Lunch Theorem' described at the end of section 2.13 also applies to feature selection techniques. Each group of feature selection techniques have different advantages and disadvantages and there is no single method that will always select features that produce the highest accuracy (Alpaydin, 2004). As with machine learning algorithms, a common approach is to try a range of feature selection methods for a given problem and dataset and select the one that performs the best on a separate validation set (Alpaydin, 2004).

There are a number of proposed categorisations for many feature selection methods (Guyon and Elisseeff, 2003; Saeys *et al.*, 2007), however only a brief introduction to this topic is possible here and these approaches are discussed in more detail in Saeys *et al.* (2007) and Mwangi *et al.* (2013). The first categorisation of feature selection methods is the distinction between supervised and unsupervised techniques. Like machine learning methods, supervised techniques selects features using data with known class labels (within the training data), whereas unsupervised

feature selection attempts to learn without being provided with group labels (Mwangi *et al.,* 2013).

Supervised methods can be separated into univariate and multivariate approaches. There are a wide range of alternative supervised feature selection methods, some of which can be used to filter the data (e.g. T-tests, Analysis of Variance and Pearson Correlation Coefficient (Chaves *et al.,* 2009; Costafreda *et al.,* 2009a; Duchesnay *et al.,* 2011; Guyon and Elisseeff, 2003)), others that use the output from machine learning predictions to test the importance of different combinations of variables (e.g. RFE (De Martino *et al.,* 2008; Saeys *et al.,* 2007)) and there exist some MVPA methods which perform the feature selection stage as part of the classification process (e.g. L_1-regularisation and the elastic net (Park and Hastie, 2007; Shen *et al.,* 2011)). T-tests and Analysis of Variance are examples of univariate feature selection methods (i.e. voxels are considered individually) which have been applied in the classification of Alzheimer's disease (Chaves *et al.,* 2009), Autism Spectrum Disorder (ASD) (Duchesnay *et al.,* 2011) and MDD (Costafreda *et al.,* 2009a). Using univariate statistical tests to filter the variables is common in the literature and has the advantages of being computationally fast, independent from the machine learning algorithm and, as they are univariate, they can easily be applied to both small and large datasets (Saeys *et al.,* 2007). The fact that these methods are univariate is also a disadvantage as they ignore possible interactions between variables which, when combined, may improve classification accuracy (Saeys *et al.,* 2007). There are also multivariate feature selection methods that attempt to find an optimal combination of voxels such as Recursive Feature Elimination (RFE) (De Martino *et al.,* 2008; Saeys *et al.,* 2007). Multivariate techniques are able to identify interactions between variables but tend to be more computationally intensive and have a higher risk of overfitting (Saeys *et al.,* 2007).

The most common unsupervised feature selection methods include principal component analysis and independent component analysis. Principal component analysis uses a linear transformation to reduce the number of correlated variables such that it is able to capture most of the variance within the original data with fewer variables (Mwangi *et al.,* 2013). As imaging data typically contain many correlated voxels, this analysis can significantly reduce the number of features. Independent component analysis assumes that the data comprises of a combination of non-Gaussian, linearly combined, and statistically independent signals and Gaussian-

distributed noise. Using these assumptions, this method attempts to extract the statistically independent multivariate signals to describe the data (Mwangi *et al.*, 2013).

An important factor when applying feature selection algorithms is that double-dipping must be carefully avoided (Kriegeskorte *et al.*, 2009). As described in section 2.8, double-dipping is a situation whereby information from the testing set is "leaked" into the training set. If feature selection was performed on the entire data set prior to cross-validation, then the features used in the classification would have been selected using the testing data. In this case the cross-validation process has been compromised as the test data is no longer novel to the classifier. Therefore in all analyses outlined in this thesis, the feature selection stage is performed *N* times (on each individual training set during each LOOCV loop) which ensures the testing set remains novel.

2.15 Feature selection methods used in this thesis

2.15.1 Mean-thresholding method

The mean-thresholding method is essentially a simplified version of a t-test, which was developed by the author (similar to a t-test but without taking into account of the variance within the data). The method involves calculating the average image for each group, during the training stage (i.e. not including the single 'left-out' scan to be predictively classified). The absolute difference between each average image in the training set is then calculated. Each voxel is then sorted from the lowest to the highest absolute difference between the average images. The thresholding process of the mean-thresholding technique involved identifying the optimal absolute difference cut-off value for the difference between the average images –all voxels above this cut-off value were included in the SVM calculation. The thresholding component of this approach inspired the thresholding component of the thresholded t-test method (described in section 2.15.2).

In order to identify the optimal cut-off value for each leave-one-out loop, a broad range of potential values was investigated. Within each (outer) leave-one-out loop the starting range was chosen to be very large (enough such that it would always include the optimal cut-off value). As it would take an unfeasibly long time to

optimise the parameters (e.g. two for the linear SVM - the soft-margin parameter and the optimal cut-off value - and three for the Gaussian SVM - the soft-margin parameter, the ‘kernel width’ parameter and the optimal cut-off value) for such a wide range of thresholds the soft-margin parameter (and the ‘kernel width’ parameter in the non-linear SVM) was initially set to unity in order to narrow down the wide range of potential thresholds. This range was reduced by performing an inner LOOCV with the SVM parameters fixed, identifying the thresholds which achieved a high training accuracy, centring the narrowed range on these thresholds, and then reducing the step size to investigate more thresholds within the range.

The author also developed a variation on the mean-thresholding method that used the number of voxels as the threshold (e.g. top one-hundred voxels) rather than an arbitrary difference between groups. The method achieved similar results as the mean-thresholding method but tended to be less robust due to the variability in the number of predictive voxels contained in each combination of training data, so was not used further in work presented in this thesis.

2.15.2 Thresholded t-test method

T-tests are one of the most popular feature selection methods. The method used in this thesis involved calculating one t-test between the two groups in each training set (as implemented in SPM). The significance threshold of the t-test was set to the highest 'acceptable' level of significance ($p < 0.05$), aiming to include as many significant voxels as possible in the optimisation stage of the feature selection process. The z-scores at each of the significant voxels were then ranked in order of significance. This allowed a threshold to be defined which could optimise the number of relevant features included in the prediction. As described in section 2.10, the threshold that optimised the feature selection was identified during LOOCV at the same stage as the MVPA parameter selection, and kept constant during the testing stage accuracy determination.

2.15.3 Thresholded linear regression method

The previous two methods are appropriate for classification studies when the feature selection depends on the comparison between groups, however, they cannot be

applied to a one-group regression study. Using the group level regression framework in SPM, the method is identical to the thresholded t-test except continuous variables are entered rather than class labels.

2.15.4 Recursive Feature Elimination

The most popular multivariate feature selection technique is RFE (De Martino *et al.,* 2008; Duchesnay *et al.,* 2011; Guyon and Elisseeff, 2003; Somol *et al.,* 1999). This method can be used in both classification and regression studies. The main argument against univariate feature selection is that it is not able to take into account any interactions between voxels.

As with the other feature selection approaches, RFE involved using the training data to identify the optimal set of voxels. As RFE is typically more prone to overfitting than univariate methods (Saeys *et al.,* 2007), two or three-fold cross-validation was typically applied during the training stage when using RFE. Backwards elimination (iteratively removing the least relevant voxels rather than iteratively adding the most important voxels) was also used during the inner cross-validation loop for the RFE optimisation procedure.

The procedure begins by performing training on the whole brain images of all subjects in the training set, identifying the weight of each voxel, removing the voxels which were in the lower half of the weights and repeating the training on the remaining voxels. The initial removal of 50% of the voxels was used to speed up the calculation as the process was repeated but with only the lowest 5% of the weights removed each time after the initial cull. The termination criteria for the RFE loop varied, but it was typically set to when there were 200 voxels or less remaining in the calculation. Once the termination criteria were met, each iteration was evaluated either using the training stage accuracy or one of the methods described previously to evaluate goodness of fit during regression. The optimal iteration selected was then used during the testing stage.

The multivariate nature of the RFE approach meant that it became difficult to not overfit to the training data set, however, when the nested K-fold cross-validation used for optimisation was sufficiently low (e.g. 2 or 3 folds) the technique was able to generalise to novel data well.

Chapter 3: Literature review

3.1 Introduction

MVPA techniques allow predictions about a range of *individual* subjects' clinical attributes (e.g. diagnostic status, symptom severity scores, identification of patients likely to develop given disorders) based on data such as, but not limited to, MRI scans (Klöppel *et al.*, 2008; Koutsouleris *et al.*, 2009; Mwangi *et al.*, 2012b). MVPA techniques (e.g. using T_1 weighted MRI images) have been successfully used to predict diagnostic status (in comparison to controls) in a range of psychiatric conditions such as: Alzheimer's disease (Klöppel *et al.*, 2008), ASD (Ecker *et al.*, 2010b), and MDD (Gong *et al.*, 2011). MVPA techniques can be applied to quantitative data such as structural and fMRI and Diffusion Tensor Imaging (DTI) as well as a wide range of non-scanning data.

The general direction this exciting field is headed is outlined, with an emphasis on the potential clinical implications of applying MVPA methods.

3.2 Overview of Neuroimaging Studies applying MVPA techniques

A number of published studies have now used MRI data in conjunction with MVPA techniques to make predictions about individual subject data. The most common MVPA neuroimaging studies attempt to use structural MRI to train a classifier to predict diagnostic status (patients versus controls). High classification accuracies have been reported in these studies in a number of disorders: e.g., predicting individual scans from patients with Alzheimer's disease vs. healthy controls with 96% accuracy (sensitivity – 0.97, specificity – 0.94) (Klöppel *et al.*, 2008), MDD vs. controls with 90% accuracy (sensitivity – 0.93, specificity – 0.88) (Mwangi *et al.*, 2012a), and a study of ASD with 81% accuracy (sensitivity – 0.77, specificity – 0.86) (Ecker *et al.*, 2010b).

In addition to the prediction of diagnosis there are also studies which have attempted to predict continuous scores such as symptom severity scores. Mwangi *et al.* (2012b) used structural MRI images of patients with MDD and RVR to predict depressive illness severity rating scores. They found that the T_1 weighted scans allowed prediction of the self-rated Beck Depression Inventory-II (BDI) score, which

correlated significantly ($p < 0.0001$) with the actual patient self-rated BDI scores (Mwangi *et al.,* 2012b). Clinical scores have also been successfully predicted from obsessive compulsive disorder using Support Vector Regression (Hoexter *et al.,* 2013) and patients with Alzheimer's disease and mild cognitive impairment using RVR (Stonnington *et al.,* 2010; Wang *et al.,* 2010). Stonnington *et al.* found that predictions of the Mini-Mental State Examination (MMSE), Dementia Rating Scale (DRS) and Auditory Verbal Learning Test (AVLT) scores all correlated significantly ($p < 0.0001$) with the actual scores (Stonnington *et al.,* 2010).

The methods used to make individual predictions of diagnostic group or predicting rating scale values can be applied to create predictive models that can address clinical prediction for which there are currently no useful methods. For example, MDD has a large number of potential treatments that can take months or often years to evaluate for a patient. Identifying predictors of response to one therapy, versus another, early in the illness could therefore be beneficial. As progress towards this goal, Gong *et al.* (2011) reported that approximately 70% of depressed patients following standard antidepressant treatment showed some improvement. Using structural MRI and SVM these authors managed to predict responders vs. non-responders to antidepressant treatment with an accuracy of 70% (sensitivity – 0.70, specificity – 0.70) (Gong *et al.,* 2011).

Other aims of applying pattern recognition techniques include the prediction of clinical outcome and/or trajectory of illness severity over time in patients and the identification of individuals who may be 'at-risk' for developing a given disorder. Koutsouleris *et al.* used structural MRI images to make individual predictions of healthy controls and 'at-risk' subjects, some of whom developed psychosis ('converters') and some of whom who did not ('non-converters') (Koutsouleris *et al.,* 2011). The authors reported accuracies of 92.3% (sensitivity – 0.94, specificity – 0.91), 66.6% (sensitivity – 0.43, specificity – 0.91), and 84.2% (sensitivity – 0.81, specificity – 0.88) when making predictions between "controls vs. converters", "controls vs. non-converters", and "converters vs. non-converters" respectively (Koutsouleris *et al.,* 2011). A similar study by Plant *et al,* also using MRI images, achieved 75% accuracy (sensitivity – 0.56, specificity – 0.87) when predicting which subjects with mild cognitive impairment (MCI) would go on to develop Alzheimer's disease (Plant *et al.,* 2010).

MVPA methods can also be applied to imaging modalities other than 'structural' MRI. Zhu *et al.* (2008; 2005) reported significant classification (85%, sensitivity – 0.78, specificity – 0.91) of ADHD children/adolescents vs. controls using resting state fMRI. Craddock *et al.* used resting state functional connectivity data to achieve a 95% accuracy when predicting clinically depressed patients vs. healthy controls (Craddock *et al.,* 2009). Similarly, Ingalhalikar *et al.* used DTI to make individual predictions of schizophrenia vs. controls (91%) and ASD patients vs. controls (90%) (Ingalhalikar *et al.,* 2010). Using event-related potential (ERP) EEG data, Mueller *et al.* reported predicting adult ADHD vs. controls category with an accuracy of 94% (Mueller *et al.,* 2011).

At this stage, it is important to mention typical limitations related to the current machine learning field in neuroimaging. It is generally ***not possible*** to train using data from one scanner and successfully predict, to a high degree of accuracy, using novel data from another scanner (Mwangi *et al.*, 2012a). The reason is that even if scanners are nominally of the same field strength (e.g. 3T), there are differences in field strength nevertheless and these cause subtle distortions in brain regions close to air filled cavities (e.g. orbitofrontal cortex and subgenual anterior cingulate, inferior temporal lobe and brainstem) which are similar to the locations of the subtle psychiatric syndrome-linked signals, used to classify images - so a major confound. Even when the same scanner is used, there are slow drifts in scanner performance over time and 'upgrades' can radically change a scanner's performance - causing the same problems. There are other potential reasons why a predictive classifier from one dataset cannot be naïvely used on a different data-set, as other subject-related confounds may also be present: e.g. differences in average age of subjects providing the 'independent' data, compared to the data the classifier was trained on. Recognising what the real limitations are with machine learning based psychiatric studies allows new studies to be designed which can address the problems.

3.3 Child and adolescent ADHD MVPA Neuroimaging studies

There have been very few studies which have applied machine learning methods to structural or functional MR images of children and adolescents with ADHD. The majority of these studies participated in the ADHD-200 competition, or used the data

released as part of it. The ADHD-200 competition, mentioned previously, was released by the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) and consisted of a multi-centre dataset of T_1 weighted scans and resting state fMRI scans of 285 children with ADHD and 491 controls (Milham *et al.*, 2012). Participants were given a first dataset with diagnostic labels to 'train' a predictor. They were then required to predict the diagnostic status and ADHD subtype (between inattentive-type (ADHD-I) and combined-type (ADHD-C) ADHD) of participants in a second, unlabelled, dataset.

The winner of the ADHD-200 competition achieved a predictive accuracy of 58% (sensitivity 21%, specificity 94%) for ADHD vs. controls (but achieved a higher accuracy of 89% when predicting between inattentive-type ADHD and combined-type ADHD) using combined resting state fMRI functional connectivity and T_1 data (Eloyan *et al.*, 2012).

Eight other studies have published results using the ADHD-200 dataset. These reported the following for ADHD vs. control prediction: 67% using a texture analysis approach on T_1 weighted brain images (Chang *et al.*, 2012), 76% accuracy using resting state fMRI and T_1 weighted brain images (Cheng *et al.*, 2012), 56% and 66% accuracy using resting state fMRI functional connectivity data respectively (Colby *et al.*, 2012; Dai *et al.*, 2012), 65% accuracy using resting state fMRI data (Solmaz *et al.*, 2012). Another study which used resting state fMRI data was not able to predict diagnostic state much above chance, but could predict between ADHD subtypes (inattentive (ADHD-I) vs. combined (ADHD-C) type ADHD) with 70% accuracy (Sato *et al.*, 2012). Similarly, Fair *et al.* attempted a 3 group classification (Controls vs. ADHD-C vs. ADHD-I) using resting state functional connectivity after controlling for micro-movements which achieved 69% accuracy (Fair *et al.*, 2012b). Bohland *et al.* (2012) combined resting state fMRI functional connectivity, T_1 weighted brain images and non-imaging features into one classifier which achieved 80% accuracy.

Interestingly, the group which actually achieved the highest accuracy in the ADHD-200 competition used phenotypic data in the prediction and *excluded* all imaging data (Brown *et al.*, 2012). However, they were disqualified from the competition as "*the intent of the competition was imaging-based classification*". The accuracy achieved by Brown *et al.* was 63% for ADHD vs. controls. The fact that the highest accuracy achieved in this competition was found using non-imaging data is

unsurprising, given the significant quality issues in the ADHD-200 data set outlined in Chapter 2 and in the literature (Johnston *et al.*, 2012; Lim *et al.*, 2013). While investigating the quality of the dataset through visual inspection, 113 of the 176 (64.2%) structural MRI scans randomly selected from seven scanning centres were found to contain a gross artefact. While investigating whether there were enough high quality scans from the two scanning centres with the most subjects, Peking University and New York University Child Study Center, 37/60 (61.7%) and 72/72 (100%) randomly selected structural MRI scans were found to contain gross artefacts respectively. The NeuroImage dataset was also investigated in full but contained gross artefacts in 26/48 (45.8%) of the structural MRI scans. Finally, as the Oregon Health and Science University dataset was found to contain the lowest ratio of poor quality structural MRI scans from the random multi-centre sample (2/17, 11.8%, excluding the Washington University sample as it only contained controls), the full dataset was investigated for gross artefacts. Of the individual scanning centres investigated in the consortium, the Oregon sample was found to have the lowest proportion of gross artefacts (33/79, 41.8%). Given better quality imaging data, it would be expected that the classifiers would be more likely to identify genuine differences in psychopathology leading to more accurate predictions. Instead, the predictions may have been based on artefactual differences or differing signal-to-noise ratios between images.

Of the childhood ADHD, MRI and MVPA studies which did not use the questionable ADHD-200 data set, there are two which classified using resting state fMRI, one using resting state fMRI functional connectivity and one which used structural MRI.

Zhu *et al.* reported significant classification (accuracy - 85%, sensitivity – 0.78, specificity – 0.91) of ADHD children/adolescents vs. controls using resting state fMRI (Zhu *et al.*, 2008; 2005). Liang *et al.* (2012) achieved an 80% accuracy (sensitivity – 0.81, specificity – 0.79) when classification was based on functional connectivity from resting state fMRI data.

The only study which predicted between ADHD vs. controls using structural MRI data (which was not part of the ADHD-200 data set) achieved an accuracy of 79% (sensitivity 76%, specificity 83%) when using the grey matter component (Lim *et al.*, 2013). The brain regions which were used to separate the groups included the caudate, ventral striatum/putamen, insula, brainstem, thalamus, hypothalamus,

precuneus/cuneus, hippocampus, amygdala, cerebellar vermis and inferior and superior parietal regions (Lim *et al.,* 2013) – regions which have been previously reported in group level analyses.

This study also attempted to provide evidence that the classifier was disorder-specific by including a third group, which consisted of boys with ASD group. The study aimed to predict ADHD vs. non-ADHD (Controls and ASD groups combined), ADHD vs. ASD and a 3 class classification process which achieved balanced accuracies of 77.1%, 85.2% and 68.2% respectively (Lim *et al.,* 2013).

To date, there have been no published manuscripts that have attempted to classify using the white matter compartment from structural MRI images or event-related fMRI from children and adolescents with ADHD. Furthermore, there have been no machine learning studies which have attempted to predict medication response, symptom severity, and clinical outcome and/or trajectory of illness severity over time in children with ADHD. These are clear gaps in the literature, which could potentially aid diagnosis, treatment and understanding of ADHD.

In summary, the application of machine learning techniques to ADHD neuroimaging is still a developing field. These methods have a large range of applications, and have the potential to be far more informative than the more popular mass-univariate methods. The goal of machine learning research in ADHD is to develop these methods for the prediction of diagnosis, while potentially identifying reliable biomarkers of ADHD – increasing the understanding of the disorder, before attempting to predict more challenging issues such as the trajectory of the disorder or response to treatment (Bray *et al.,* 2009).

3.4 MDD MVPA Neuroimaging studies

Machine learning studies that attempt to make predictions using neuroimaging data of adults with MDD are far more common in the literature than the corresponding literature on ADHD. Alongside popular diagnostic prediction, studies have attempted to predict symptom severity (Mwangi *et al.,* 2012b) and treatment response (Fu *et al.,* 2008; Gao *et al.,* 2012).

One study by Fang *et al.* (2012) attempted to predict MDD diagnosis using DTI and SVM. The study achieved a classification accuracy of 92% (sensitivity – 86%, specificity – 96%, $p < 0.0001$) with increased connectivity in the depressed

group compared to controls in the cortical-limbic network and, to a lesser extent, the temporal-occipital network (including connections coming from regions such as the cingulate gyrus, insula, hippocampus, caudate, putamen, pallidum and thalamus) (Fang *et al.*, 2012). Unfortunately, this high accuracy may have been obtained through overfitting as the method does not explain how it was decided that the kernel-width parameter should be set equal to three throughout the classification process. Despite this concern, the demonstration that DTI data can be used to predict MDD to such a high accuracy is very encouraging. This is the only paper to have predicted MDD diagnosis using anatomical connectivity, but the senior author in this study, Dewen Hu, has also appeared as the senior author of five out of the seven manuscripts which apply machine learning methods to MDD functional connectivity data – all but one of which have been published in the last two years.

The first study to successfully develop a predictor enabling diagnosis of major depression using functional connectivity was performed by Craddock *et al.* (2009). In this study, a number of different feature selection methods were tested with classification accuracies ranging between 63% and 95%. Lord *et al.* (2012) also applied a feature selection method and SVM to functional connectivity data to distinguish unipolar depression from healthy controls. In this study, Lord *et al.* found that their approach could achieve almost perfect separation between the groups (above 99% classification accuracy). The most influential features which guided this classification included the putamen, thalamus and insula, among other regions (Lord *et al.*, 2012).

Using a different approach, Zeng *et al.* (2012) achieved a 94% classification accuracy when predicting between major depression and healthy controls using whole brain functional connectivity. Regions which were most relevant to the prediction included the amygdala, anterior cingulate cortex, parahippocampal gyrus and hippocampus. Ma *et al.* (2012) also performed a diagnostic classification of MDD using functional connectivity but with a focus on connections from the cerebellum, achieving an impressive accuracy of 91%.

The final study which attempted to classify solely between MDD and healthy controls using functional connectivity is, perhaps, the most interesting, despite having a lower accuracy than some of the other studies. This is because Zeng *et al.* (2013) successfully applied an unsupervised machine learning technique to predict diagnosis, achieving a classification accuracy of 93%. As mentioned previously, the

majority of machine learning studies which are implemented in the literature are supervised techniques (when the class membership of subjects in the training set is known by the classifier) but unsupervised classification involves the identification of an underlying pattern in the data without any knowledge of class membership (Bishop, 2006; Orrù *et al.,* 2012). This result is significant because successful unsupervised classification means that the prediction provides a more objective diagnosis from neuroimaging evidence rather than relying on potential biases from patients or clinicians. The unsupervised machine learning approach used maximum margin clustering (MMC) and was developed as a result of SVM. The key principle in SVM is to maximise the separation (margin) between the two classes, but MMC develops this principle by defining class membership by identifying the maximal margin between subjects in the training set (Zeng *et al.,* 2013).

Yu *et al.* (2013) performed a multiclass diagnostic prediction between major depression, schizophrenia and healthy controls. In this study they achieved 81% classification accuracy (84.2% for MDD patients, 81.3% for schizophrenic patients and 78.9% for healthy controls). Regions which showed abnormal connections in both schizophrenia and major depression compared to controls included the medial prefrontal cortex, anterior cingulate cortex, thalamus, hippocampus and cerebellum and the regions which differentiated the two disorders included the prefrontal cortex, amygdala and temporal poles (Yu *et al.,* 2013).

The final functional connectivity MDD MVPA study did not attempt to predict diagnosis, but rather prediction of recovery. In this study by Gao *et al.* (2012), three groups were involved in the classification process, medication-naïve patients with major depression, healthy controls and previously depressed patients who achieved clinical remission through treatment. The aim was to train a classifier to predict between the depressed group and the control group to investigate which class the remission group would be predominantly attributed to. When using no feature selection, the remitted group were predicted to be healthy controls in 14 out of the 16 subjects (88%), but all sixteen of the remitted group were predicted to be controls when feature selection was implemented as part of the process.

Functional connectivity is an emerging field with a large number of approaches to preparing and analysing the data. The results when applying functional connectivity data to machine learning are very impressive, however these results require independent replication using an identical approach as the optimal

preprocessing techniques for functional connectivity remain actively debated (Lord *et al.,* 2012). Biomarkers identified using functional connectivity measures may require additional evidence such as the identification of a structural abnormality biomarker to provide a more reliable clinical diagnosis of MDD (Zeng *et al.,* 2012).

The application of machine learning methods to task-based fMRI images is another active research area. The first study which applied machine learning using task-based fMRI to MDD was by Fu *et al.* (2008). The task focused on the identification of the intensity of sadness in male and female faces. Considering no feature selection was used, the diagnostic classification accuracy of 86% (sensitivity – 84%, specificity – 89%, $p < 0.0001$) is very impressive when the prediction was based on the lowest intensity of sadness stimuli, however this accuracy was the highest of the reported techniques which ranged from 53% to 86% accuracy (Fu *et al.,* 2008). In addition to attempting to predict diagnosis of depression, Fu and colleagues also attempted to predict which depressed patients (all of whom were psychotropic medication free for at least 4 weeks at the time of recruiting) would respond after 8 weeks to the antidepressant medication fluoxetine. Given the small sample size of responders vs. non-responders, the authors acknowledged that this goal was less likely to succeed which, unfortunately, was shown to be the case as their best classifier was only able to correctly identify 75% of non/partial-responders and 62% of full responders ($p = 0.11$). The same (or similar) emotional processing fMRI paradigm (displaying a varying degree of happiness or sadness expressed on a face) has been used in a large proportion of fMRI machine learning studies in depression.

Mourão-Miranda *et al.* attempted to distinguish three groups, bipolar disorder, unipolar depression and healthy controls using responses to happy and neutral faces (Mourão-Miranda *et al.,* 2012a). However, the only significant classification was when using the mildly happy faces vs. neutral faces contrast to classify bipolar disorder vs. unipolar depression (accuracy = 67%, specificity = 72%, sensitivity = 61%, $p = 0.02$). They also investigated whether it would be possible to predict between the fMRI contrasts intensely happy faces vs. neutral faces contrasts and mildly happy faces vs. neutral faces contrasts for each diagnostic group. In the prediction of the intensely happy faces and neutral faces contrast, all diagnostic groups achieved a significant classification accuracy (BD = 61%, UD = 70% and controls = 81%) whereas in the classification of the mildly happy and neutral faces

contrast only the control group achieved an accuracy above chance (75%) (Mourão-Miranda *et al.*, 2012a). This shows that the activation pattern of the latter contrast does not have as strong a pattern in the patient groups, potentially leading to an indirect diagnosis.

Another study which attempted to classify between unipolar depression and bipolar disorder using a similar emotional processing paradigm was able to achieve an accuracy of 90% in the unipolar vs. bipolar prediction when using a happy vs. neutral faces contrast (specificity = 90%, sensitivity = 90%, $p = 0.003$), however, as this study only included ten subjects in each group, it requires replication (Grotegerd *et al.*, 2012). The classification accuracy was reduced when predicting using the sad/angry vs. neutral faces contrast and when both contrasts were entered into the SVM classifier together (respective accuracies of 75% and 80%). Interestingly, in addition to the reported SVM accuracies, Grotegerd and colleagues (2012) also implemented a Gaussian Processes Classifier, however, all the accuracies were lower than their corresponding SVM values.

The fMRI response to neutral faces in depression has also been investigated (Oliveira *et al.*, 2013). In this study the goal was to train two classifiers to predict between happy or sad faces vs. neutral faces using only healthy controls. Following training, a new sample of controls and a MDD group were tested on the classifier to test the hypotheses that the confidence of the classification in the depression group would be significantly lower than in the control group due to the patterns of brain activations to emotional and neutral faces differing in depressed patients. The healthy controls were predicted above chance for both emotional and neutral faces stimuli, but the depression group classification only achieved a significant accuracy when classification was based on the emotional faces activations and not the neutral faces (Oliveira *et al.*, 2013). However, it is possible that this result may be due to a lack of a strong pattern in either healthy controls, or MDD patients, when presented with the neutral faces stimuli as the accuracy when predicting neutral faces was greatly reduced in both groups. In other words, it is possible that the results are obtained from a classifier that identified a pattern that was based predominantly on emotional faces rather than any distinct pattern of activation from neutral faces.

An interesting MVPA method which may be able to handle data with a large imbalance in class membership (e.g. between a large group such as healthy controls and a smaller, more difficult to recruit/successfully scan group such as a patient

group), is the application of the one-class support vector machine (OC-SVM). As discussed in Chapter 2, a large class imbalance could lead to poorer results, therefore using a subset of the larger group to train a classifier of "normality" and testing whether or not each of the remaining subjects are outliers is a elegant solution. Mourão-Miranda *et al.* (2011) used this approach on emotional processing fMRI data in depressed patients vs. controls and correctly identified 79% of controls as non-outliers but only managed to correctly identify 52% of patients as outliers. Despite this disappointing result, further analysis identified that 70% of patients classified as outliers did not respond to treatment and 89% of patients classified as non-outliers responded to treatment, revealing a potential approach to identify treatment response (Mourão-Miranda *et al.,* 2011).

Hahn *et al.* (2011) proposed that combining a multitude of fMRI contrasts from a few different paradigms into one classifier would improve diagnostic classification accuracy. The fMRI paradigms implemented involved passively viewing facial expressions, the same or similar paradigm that is used in almost all fMRI MVPA MDD studies in the literature, and a monetary incentive delay task. The highest accuracy obtained from a single fMRI condition was 72% (when a contrast describing the anticipation of no loss was used, the median accuracy for all conditions was 60%), however, when 3 of the 15 conditions were combined the accuracy increased to 83% (sensitivity – 80%, specificity – 87%). The three conditions which led to the highest accuracy were neutral facial expressions, actual large reward and anticipation of no loss (Hahn *et al.,* 2011).

The only fMRI study which did not employ the emotional processing paradigm in an MPVA MDD analysis was performed by Marquand *et al.* (2008). This study achieved 68% when diagnosing MDD vs. healthy controls using a 2-back condition during an n-back working memory task.

Emotional valence has also been investigated with respect to the emotional processing paradigm (Habes *et al.,* 2013). The study was able to accurately predict different valence discriminations (between positive, neutral and negative valence stimuli) in depression.

Prediction to discriminate healthy adolescents who are genetically high or low risk of developing mood disorders has also been performed using the same passive viewing of facial expressions paradigm (Mourão-Miranda *et al.,* 2012b). This study achieved variable results when classifying based on different stimuli ranging

from 38-75% classification accuracy. The most interesting part of this study, however, was the follow up analysis, to investigate which of the at-risk group developed mood disorders. Using their best classifier (neutral faces presented during the happy face experiment which achieved 75% accuracy), Mourão-Miranda *et al.* identified that of the four at-risk individuals that were "misclassified" as healthy controls, three remained healthy and one did not take part in the follow up, potentially revealing why these subjects were not identified as "at-risk". Furthermore, the six participants in the at-risk group who developed either major depression or anxiety disorders (out of the thirteen that took part in the follow up as three withdrew after the initial scan) had significantly higher probabilities of group membership than the other at-risk adolescents who remained healthy at the follow-up (Mourão-Miranda *et al.,* 2012b).

Another application of the emotional processing task is to investigate whether it can facilitate a prediction of response to cognitive behavioural therapy (CBT) in depression (Costafreda *et al.*, 2009b). When predictions were based on neutral or the highest intensity of sad faces, the machine learning approach could distinguish responders and non-responders to CBT with 71% sensitivity and 86% specificity ($p = 0.029$), however, when the medium intensity of sadness images were used, the accuracy dropped such that the sensitivity was 57% and the specificity 43% (Costafreda *et al.,* 2009b).

With one exception, all of the studies in fMRI machine learning applied in mood disorders have used an emotional processing fMRI paradigm. In the study which attempted to combine results from fMRI paradigms, it was found that receiving a large reward and the anticipation of avoiding a loss were very relevant to distinguishing depression from healthy controls. Therefore, the application of a reward and aversive events paradigm that was used and reported later in this thesis, is very relevant, yet relatively novel, to the literature. In addition, the studies outlined above tend to have very small sample sizes (typically less than twenty per group), which means that the group sizes analysed in this thesis are actually larger or equivalent to the majority of those in this field.

Structural MRI is typically the most common imaging modality to be applied to machine learning methods due to the higher resolution, higher signal-to-noise ratio and easier interpretation of the results, in comparison to event-related fMRI. Machine learning studies are particularly interesting in neuroimaging when a study is able to

link strong classification accuracies back to the underlying biology. However, many machine learning studies have a primary goal to test a number of different classifiers or feature selection techniques rather than to discuss the results in the context of the groups being classified. One such study by Kipli *et al.* (2013) tests four different feature selection techniques on four different machine learning methods by attempting to classify structural MRI images (in particular, information extracted from structural MRI, e.g. volumes of various structures) of depressed individuals and healthy controls. The author suggests that the Information Gain algorithm outperforms OneR, SVM (using RFE) and ReliefF feature selection methods as it achieved the highest average accuracy (72%) when applied to four different classfiers (Kipli *et al.,* 2013). A concern in this study, however, is that 77% (88/115) of subjects in this study belonged to the control group. Since the sensitivity and specificity of these results are not disclosed, it is unclear if the large class imbalance is an issue. Another study which focused more on various results from feature selection rather than the neurobiology, attempted to predict diagnosis between bipolar disorder and healthy controls, achieving accuracies ranging between 60-90% (Termenon *et al.*, 2013).

Single centre studies are encouraging, but one of the next developments is to implement machine learning based diagnosis in data from multiple scanners. This has been achieved by Mwangi *et al.* (2012a) when they successfully classified structural MRI scans of people with depression and healthy controls obtained from two scanning centres. Mwangi and colleagues (2012a) implemented both an SVM and RVM approach with the latter achieving a slightly higher classification accuracy (90%). In this study, grey matter reductions were identified in MDD compared to controls in the dorsolateral prefrontal cortex, medial frontal cortex, orbitofrontal cortex, temporal lobe, insula, cerebellum and posterior lobe – consistent with the literature from group-level VBM analyses. In addition, Mwangi *et al.* identified that the weights extracted for each subject during both the SVM and, to a lesser extent, the RVM classification correlated strongly with MDD severity scores (Mwangi *et al.,* 2012a).

As well as identifying an approach to indirectly predict symptom severity scores, Mwangi *et al.* also used RVR to predict illness severity directly (Mwangi *et al.*, 2012b). In this study, they found that it was possible to significantly predict the BDI scores from the whole-brain structural MRI scans but not the Hamilton Rating

Scale for Depression (HAM-D) (Mwangi *et al.*, 2012b). As the BDI is a self-administered rating scale and the HAM-D is performed through a semi-structured interview with the patient by a trained observer, this study raised an interesting question as to which score best reflects the underlying neurobiology of the symptoms of MDD.

As well as predicting diagnosis of depression, Costafreda *et al.* (2009a) also attempted to predict response to either antidepressant medication (fluoxetine), or to cognitive behavioural therapy (CBT). The classification of response to antidepressant medication managed an impressive accuracy of 89% (sensitivity = 89%, specificity = 89%, p =0.01) while the diagnostic classification achieved a lower accuracy of 68% (sensitivity = 65%, specificity = 70%, p =0.027). The prediction of response to CBT was not significant.

The brain regions driving the prognostic classification included increased grey matter in the right rostral anterior cingulated cortex (BA 32), left posterior cingulate cortex (BA 31), left middle frontal gyrus (BA 6), and right occipital cortex (BA 19) in recovered patients and decreased regions in the orbitofrontal cortices bilaterally (BA 11), right superior frontal cortex (BA 10) and left hippocampus (Costafreda *et al.*, 2009a).

Another study that attempted to predict treatment response between treatment-resistant depression (TRD), treatment-sensitive depression (TSD) and healthy controls, using both grey and white matter images separated the diagnostic classification into two separate predictions: responders vs. controls and non-responders vs. controls (Liu *et al.*, 2012). Separating the responders' and non-responders' diagnostic predictions improved the classification accuracies compared to Costafreda *et al.* (2009a). The predictions for TSD vs. controls achieved 82% accuracy using grey matter and 91% accuracy using white matter and for TRD vs. controls achieved 86% for both grey and white matter (Liu *et al.*, 2012). When predicting treatment response, Liu *et al.* also managed to achieve 83% when using either grey or white matter images (2012).

In a similar, but larger study, Gong *et al.* (2011) attempted to predict between non-refractory depressive disorder (NDD – responders to antidepressant treatment) vs. controls, refractory depressive disorder (RDD – non-responders to antidepressant treatment) vs. controls and NDD vs. RDD, using both grey and white matter structural MRI. The patient groups had no previous psychiatric treatment, including

receiving no antidepressant treatment, prior to scanning. The NDD and RDD vs. controls predictions yielded accuracies of 76% and 67% respectively using grey matter images and 85% and 59% respectively using white matter images (Gong *et al.*, 2011). Counter intuitively, in these results, responders are easier to differentiate from controls than non-responders to antidepressant medication. The NDD vs. RDD classification yielded an accuracy of 70% using grey matter (specificity = 70%, sensitivity = 70%, p = 0.006) and 65% using white matter (specificity = 74%, sensitivity = 57%, p = 0.02). Combining the grey and white matter images did not improve accuracy more than classifying on grey matter alone. Despite this study having more subjects, the classification accuracy was reduced compared to the other two MDD prognostic prediction studies using structural MRI.

There are two studies that have attempted multimodal classifications, both combining structural MRI and another imaging modality to classify MDD from healthy controls. The first of these studies combined structural MRI with proton magnetic resonance spectroscopy, managing to separate the females with MDD and the female controls with perfect accuracy (Floares *et al.*, 2006). The second, involved testing a new classification approach, transductive conformal predictor (TCP), on data from earlier studies (Costafreda *et al.*, 2009a; Fu *et al.*, 2008; Nouretdinov *et al.*, 2011). The goal of this study was to repeat the diagnostic prediction using fMRI data (emotional processing task) and the prognostic prediction using structural MRI data (response after 8 weeks on anti-depressant medication) using TCP. Nouretdinov *et al.* found that their results were comparable with the methods used in previous studies (Costafreda *et al.*, 2009a; Fu *et al.*, 2008; Nouretdinov *et al.*, 2011) yet the method provides a few advantages such as providing a measure of confidence for each prediction and the ability to handle multi-class predictions.

Machine learning studies in MDD are far more common than in the child and adolescent psychiatry literature. The confidence to apply machine learning methods to increasingly difficult problems has shown that it is possible to predict relevant information such as treatment response and symptom severity. However, there are still a lot of gaps in the literature that remain and a lot of replication is required to increase confidence in these techniques further. A clear gap in the literature is the application of machine learning to fMRI paradigms other than those related to emotional processing, as only Hahn *et al.* and Marquand *et al.* have applied machine learning methods to fMRI data outwith this paradigm for MDD (Hahn *et al.*, 2011;

Marquand *et al.,* 2008). Furthermore, there has been no investigation into how one imaging modality relates to another (e.g. how the structural deficits reported in MDD influences the fMRI patterns of activation and deactivation). The work outlined in this thesis either describes work that is – to date - missing in the literature, or provides independent replication of prior studies in order to increase confidence in these methods.

3.5 Summary

Significantly high accuracies of individual diagnostic classification have been reported for both adult and younger populations, specifically MDD and ADHD. These studies have the potential to identify biomarkers and elucidate the mechanisms of psychiatric disorders. However, the results require independent replication in larger samples, and potentially in a multi-centre study, before it can be considered for any clinical applications.

Furthermore, the application of pattern recognition techniques to clinically relevant questions, such as the prediction of outcome of disorders and prediction of treatment response, are very promising in the MDD neuroimaging literature. However, to the author's knowledge, there are currently no ADHD studies that have reported similarly high accuracies in the prediction of treatment outcome, or syndrome outcome. Successful and reliable application of these techniques to ADHD populations could allow pattern recognition techniques to have a major role in future clinical practice.

The key areas for future development include performing multi-class predictions (e.g. develop a classifier to be able to predict between ADHD, Autism Spectrum Disorder and healthy controls), an increase in the robustness of the machine learning methods, working with larger datasets to investigate particularly heterogeneous disorders and a process to highlight less certain predictions by including a probabilistic measure of confidence in the prediction (Klöppel *et al.,* 2012).

Finally, it is important to emphasise that machine learning is an active research area in itself so methods are currently in development and being tested. These methods do not guarantee good results and it takes some time to develop an understanding of the various pitfalls involved in this research field. Failure to

appreciate the complexity of machine learning could lead to either falsely enhanced prediction or poor classifier performance. However, when applied correctly, these methods compliment neuroimaging studies perfectly as they can combine to create a multivariate insight into brain structure, function and connectivity.

Chapter 4: Predicting Methylphenidate Treatment Response in Drug-Naïve Boys with ADHD.

4.1 Introduction

Methylphenidate (MPH) is the most commonly prescribed stimulant medication for ADHD. Whilst ~70% of children will respond to and tolerate MPH, 30% do not (Humphrey, 1992).

Denney and Rapport (1999) evaluated models designed to predict MPH response. None of the MPH response prediction studies they investigated could be replicated. Denney and Rapport proposed that a comprehensive model of MPH response must include both a biological and a behavioural component (Denney and Rapport, 1999).

Coghill *et al.* (2007) investigated whether there was a correlation between clinical response to MPH and neuropsychological measures, and found that poor performance in the Delayed Matching to Sample (DMtS) task was the only predictor of response at baseline.

The present study aimed to predict MPH response using a multivariate approach on a subset of the data identified by Coghill *et al.* (2007) as being the most relevant to medication response. The identification of this subset was through a principal components analysis which was not performed by the author and is discussed in more detail elsewhere (Coghill *et al.*, 2007). Treatment response was determined using the method of Jacobson and Truax (Jacobson and Truax, 1991). Using this method, both "clinically significant change" and "reliable change" are required as criteria to determine full response. Following treatment, if the patients' post-treatment severity scores move toward the controls' scores, beyond a specified threshold (defined using pre-treatment severity scores for patients and controls) then the subject has achieved clinically significant change. Reliable change is a measure of how much each subjects' scores changed during treatment and is calculated by dividing the difference between the pre-treatment and post-treatment scores by the standard error of difference between the two scores. Subjects who experienced significant adverse side-effects – irrespective of symptom changes – that caused the treatment to be stopped were also classed as non-responders.

4.2 Methods

4.2.1 Subjects

Of the 75 boys with ADHD included in Coghill *et al.* (2007), thirty-two were excluded, either due to incomplete data, or to ensure there were no significant differences in age or verbal IQ (estimated using the British Picture Vocabulary Scale (BPVS)) between the group of responders and group of non-responders. Of the 43 boys with ADHD included in the present study, 30 were classed as responders (i.e. showing both "clinically significant change" and "reliable change") and 13 were classed as non-responders to MPH. The neuropsychological tasks included several tasks from CANTAB (Cambridge Neuropsychological Test Automated Battery) and a separate computerised Go/NoGo task.

Thirteen variables were included in the analysis: three demographic variables, three clinical variables and seven neuropsychological test scores. The demographic variables included were the BPVS percentile rank, decimal age, and deprivation (SIMD) score (an estimation of the socioeconomic background). The three clinical variables of interest included were the presence of comorbid oppositional defiant disorder or conduct disorder and the t-score baseline Parents ADHD Conners' questionnaire. Finally, four neuropsychological task scores were taken from a Go/NoGo task and three from the CANTAB Visual Memory Battery (Pattern recognition, Spatial recognition, and DMtS total percent correct z scores[*]). These thirteen variables were a subset of the total number of variables and were identified by Coghill *et al.* (2007) using principal component analysis to be variables which would most likely distinguish responders to MPH from non-responders.

The Go/NoGo task involved subjects being presented with a sequence of letters and numbers on the screen. The 'type 1' block corresponds to a 'switch' trials where subjects were required to withhold response when the stimulus has changed from letters to numbers, or vice-versa. The 'type 2' block corresponds to 'non-switch' trials where the subjects were required to withhold response when a letter is presented if the previous stimulus was a letter and likewise for numbers. For both 'type 1' and 'type 2', the output variables are the 'mean number of errors for distractors' (ERD), a measure of the average number of times the subject responded

[*] adjusted for age and BPVS

when they were required to inhibit their response, and the 'reaction times to target stimuli' (RTT), a measure of the reaction time when a correct response (key press) was required. Four variables were extracted from the Go/NoGo task: Go/NoGo - type 1 RTT and ERD and Go/NoGo - type 2 RTT and ERD. The Visual Memory Battery tests the ability to recognise a previously presented abstract pattern in a forced choice procedure for the pattern recognition task, the ability to recognise the spatial locations of target stimuli for the spatial recognition task, and the ability to remember the visual features of a complex, abstract, target stimulus and to select from a choice of four patterns after a variable delay in the DMtS task (Coghill *et al.,* 2007). All the variables above are described in more detail by Coghill *et al.* (2007).

4.2.2 Variable preparation

As a first step, the variables were normalised to reduce errors due to scaling. All scores were brought within the range 0-1 by simply subtracting the minimum value and dividing by the variance (maximum – minimum value). This ensured that the analysis selected variables based on their predictive value, rather than variability or magnitude.

4.2.3 Discriminant Analysis

Discriminant analysis was first conducted using IBM SPSS Statistics for Windows (v19). An automated variable selection method was used for the discriminant analysis. Variables were ranked in the order of the amount they lowered Wilks' lambda (a statistical test which reflects the importance of a variable, smaller Wilks' lambda values reflect greater importance to classification). The variable which lowered lambda the most was iteratively included with the variables used in the classification with the termination criterion that variable selection stops when the significance calculated using an F-test is less than $p<0.05$.

4.2.4 Individual Scan Classification

To attempt to classify responders and non-responders to MPH on individual subjects, a linear SVM (Vapnik, 1995; Vapnik, 1998) was explored. To avoid double dipping

(defined in Chapter 2), standard LOOCV was used, with a second (inner) leave-one-out loop used for parameter selection.

To improve prediction accuracy it is common to use feature selection methods to highlight the variables that contribute the most to the prediction (Mwangi *et al.,* 2013). The feature selection technique used here was a "mean-thresholding" technique – a simple method created by the author.

As described in Chapter 2, the mean-thresholding technique involves ranking the variables in order of the magnitude of the difference between the mean group values within each outer 'leave-one-out' loop (excluding the left out subject to avoid double dipping). As the number of variables was relatively small, the mean-thresholding method was altered such that the lowest ranked variable was removed from the analysis and the classification was repeated with the reduced number of variables (in the inner 'leave-one-out' loop). The subsequent lowest ranked variable from the following classification was again removed and the process was repeated until there was one variable remaining. Therefore, rather than optimising an arbitrary threshold, as is necessary in the standard mean-thresholding method due to the typically large number of variables (voxels) in an imaging analysis, this approach optimised the number of variables required to classify the data. The classification accuracy, sensitivity and specificity were calculated during each of the iterations.

To avoid the class imbalance problem (introduced in Chapter 2), the combination of variables/iteration which achieved the highest sensitivity (more accurately classifying those in the group with the fewest subjects – the non-responders) were then selected for training and testing the SVM on the left out subject in the outer 'leave-one-out' loop. If two or more iterations obtained identical maximum sensitivity values then the accuracy was used as a secondary selection parameter. Therefore, as variable selection took place in each 'leave-one-out' loop, it is possible a different combination of variables could be selected for each individual prediction. A flowchart representation of the method is shown in Figure 5 and Figure 6.

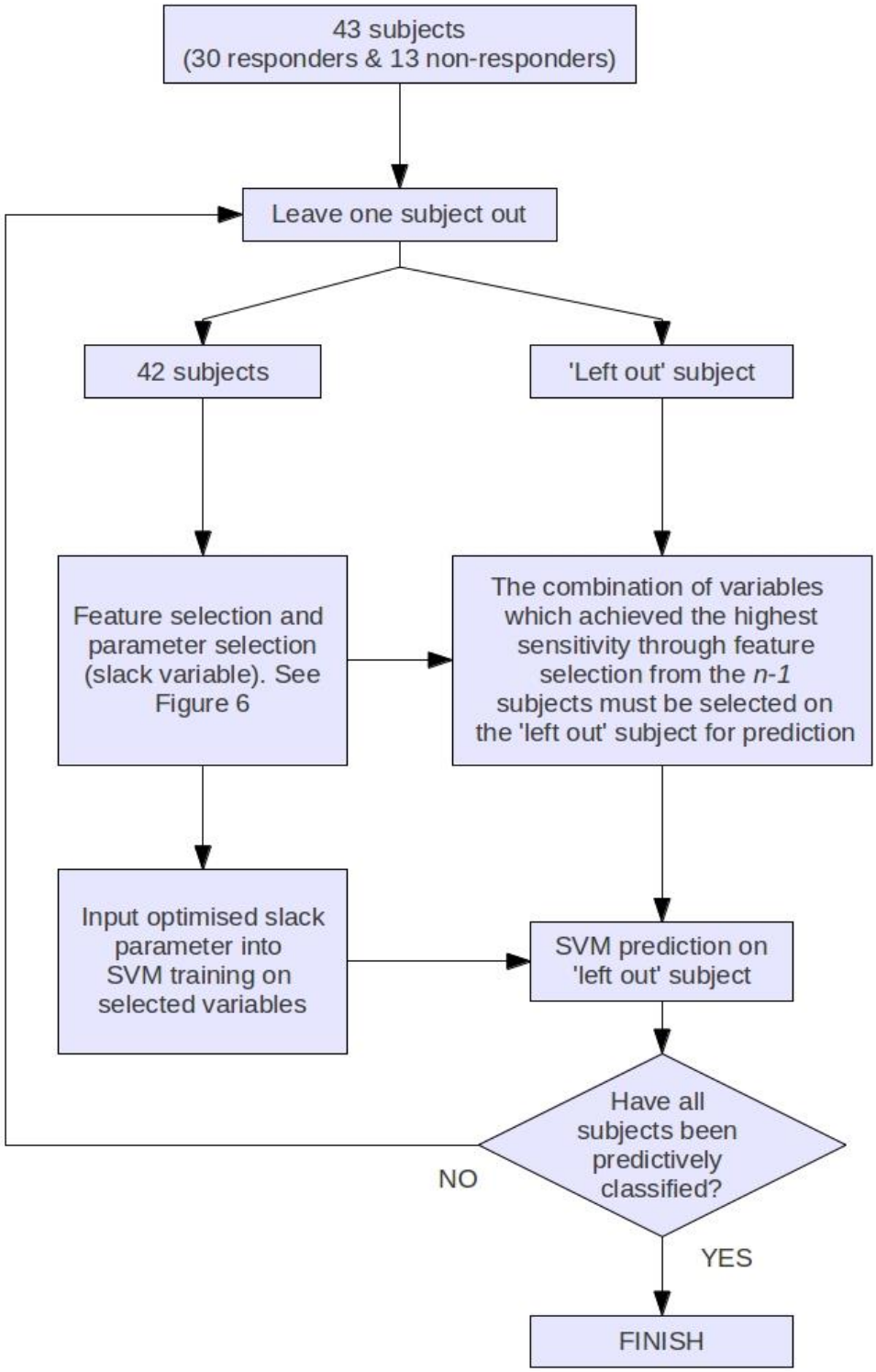

Figure 5: A flowchart of the SVM prediction technique.

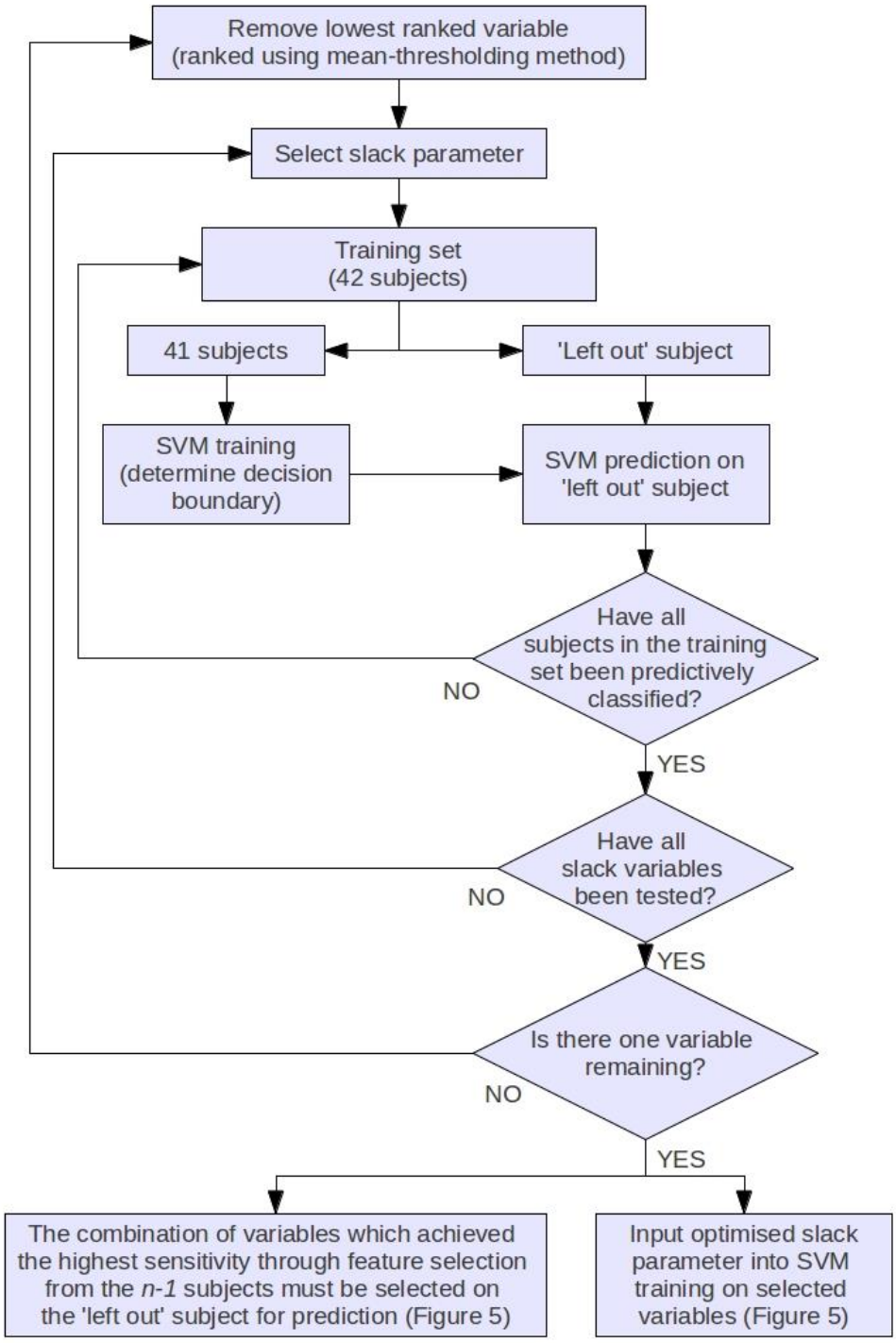

Figure 6: A flowchart of the variable and parameter selection stage in Figure 5.

4.2.5 Euclidean Distance from the SVM Hyperplane Investigation

To investigate whether the incorrectly classified subjects were closer to the linear hyperplane (in other words, close to being correctly classified) than the correctly classified subjects, the author created a function which calculated the shortest distance (the Euclidean distance) between the subject being classified and the hyperplane (which was tuned using the training set during cross-validation). The Euclidean distance was calculated by first identifying the vector normal to the hyperplane. As the SVM algorithm outputs the distance between the subject and the interception of the plane in the y-direction, the Euclidean distance was calculated using the SVM output, the vector normal to the hyperplane and standard trigonometric equations.

4.3 Results

4.3.1 Participant Characteristics

Age and verbal IQ did not differ significantly (t-test, $p>0.1$). The MPH responder group mean age was 11.2 years (standard deviation 2.4) mean IQ was 40.3 (standard deviation 30.6). The MPH non-responder group mean age was 11.3 years (standard deviation 3.0) and the mean IQ 32.2 (standard deviation 28.8). There were no significant differences in task performance between groups. These results are outlined in Table 1.

4.3.2 Discriminant Analysis

The only variables which were selected during the discriminant analysis were the presence of conduct disorder and Go/NoGo - type 1 ERD. Using these two variables, the classification accuracy achieved was 67.4% (sensitivity = 0.69, specificity = 0.67, $\chi^2 = 4.7$, $p = 0.03$). The contrast matrix is displayed below:

$$\begin{array}{cc} & \textit{Predicted} \\ & \begin{array}{cc} -1 & 1 \end{array} \\ \textit{Actual} \quad \begin{array}{c} -1 \\ 1 \end{array} & \begin{pmatrix} 9 & 4 \\ 10 & 20 \end{pmatrix} \end{array}$$

Table 1: Clinical descriptors for responders and non-responders to MPH. Variables are shown as mean (standard deviation). *chi-square test with other tests being t-tests.

	Responders (N=30)	Non-responders (N=13)	
BPVS Percentile rank	40.27 (30.58)	32.15 (28.84)	n.s.
decimal age	11.19 (2.39)	11.26 (2.99)	n.s.
diagnosis of oppositional defiant disorder*	21/30	9/13	n.s.
diagnosis of conduct disorder*	14/30	2/13	n.s.
deprivation score	4.27 (1.72)	4.08 (1.32)	n.s.
t-score baseline Parents ADHD Conners	78.07 (4.25)	80.08 (4.03)	n.s.
Go/NoGo - type 1 RTT	441.61 (91.85)	504.37 (108.07)	n.s.
Go/NoGo - type 2 RTT	457.48 (70.20)	497.13 (116.15)	n.s.
Go/NoGo - type 1 ERD	2.85 (1.54)	1.92 (1.50)	n.s.
Go/NoGo - type 2 ERD	2.63 (1.81)	1.85 (1.49)	n.s.
Pattern recognition z score*	-1.12 (1.66)	-0.56 (1.48)	n.s.
Spatial recognition z score*	-0.97 (0.89)	-0.55 (1.17)	n.s.
DMtS total percent correct z score*	-1.07 (1.23)	-0.66 (0.96)	n.s.

* adjusted for age and BPVS
* adjusted for age and BPVS
* adjusted for age and BPVS

4.3.3 Individual Subject SVM Predictions

When no feature selection was used, this method achieved a marginally improved (compared with the discriminant analysis) predictive accuracy of 69.8% (sensitivity = 0.46, specificity = 0.8, $\chi^2 = 3.1$, p = 0.08). However, due to poor sensitivity the classification was not significant. The reason that sensitivity was lower than expected may be due to the class imbalance problem (Theodoridis and Koutroumbas, 2006). The contrast matrix (below) shows that classification accuracy in the non-responder group (given the label '-1') was poor as only six out of thirteen subjects were correctly identified as non-responders:

		Predicted	
		−1	1
Actual	−1	6	7
	1	6	24

Combining feature selection and a linear SVM approach improved the accuracy to 76.7% (sensitivity = 0.54, specificity = 0.87, $\chi^2 = 7.8$, p = 0.005).

		Predicted	
		−1	1
Actual	−1	7	6
	1	4	26

Presence of conduct disorder and Go/NoGo - type 1 ERD were the only variables which were selected in all of the cross-validation predictions (the same two variables selected during the discriminant analysis), although Go/NoGo - type 1 RTT was also selected in a high proportion of predictions (38/43 subjects). On average, 4 variables were used in each prediction. The variables which were never selected (using the feature selection method) for predictions were decimal age, presence of oppositional defiant disorder, deprivation score and, interestingly, DMtS total percent correct z score* which was previously highlighted by Coghill at al (2007) as the only neuropsychological predictor of clinical response at baseline. The number of times each variable was used in each 'leave-one-out' prediction (which variables were most relevant when distinguishing responders from non-responders) is highlighted in Table 2.

* adjusted for age and BPVS

Table 2: Frequency of variable selection in leave-one-out method.

	Frequency of variable selection	Percentage of leave-one-out loops variable selected
BPVS Percentile rank	2	0.05
decimal age	0	0
presence of oppositional defiant disorder	0	0
presence of conduct disorder	43	1
deprivation score	0	0
t-score baseline Parents ADHD Conners	24	0.56
Go/NoGo - type 1 RTT	38	0.88
Go/NoGo - type 2 RTT	2	0.05
Go/NoGo - type 1 ERD	43	1
Go/NoGo - type 2 ERD	9	0.21
Pattern recognition z score*	12	0.28
Spatial recognition z score*	1	0.02
DMtS total percent correct z score*	0	0

* adjusted for age and BPVS
* adjusted for age and BPVS
* adjusted for age and BPVS

4.3.4 Euclidean Distance from the SVM Hyperplane Investigation

The results are shown pictorially in Figure 7, with the correctly classified and two types of incorrectly classified results separated on the y-axis. Subjects that were correctly classified were given the value zero, false positives (whereby non-responders were predicted to be responders) were given the value 1 and false negatives (whereby responders were predicted to be non-responders) were given the value -1. These results are also displayed on a histogram in Figure 8, with the incorrectly classified subjects shown in green and the correctly classified subjects shown in blue.

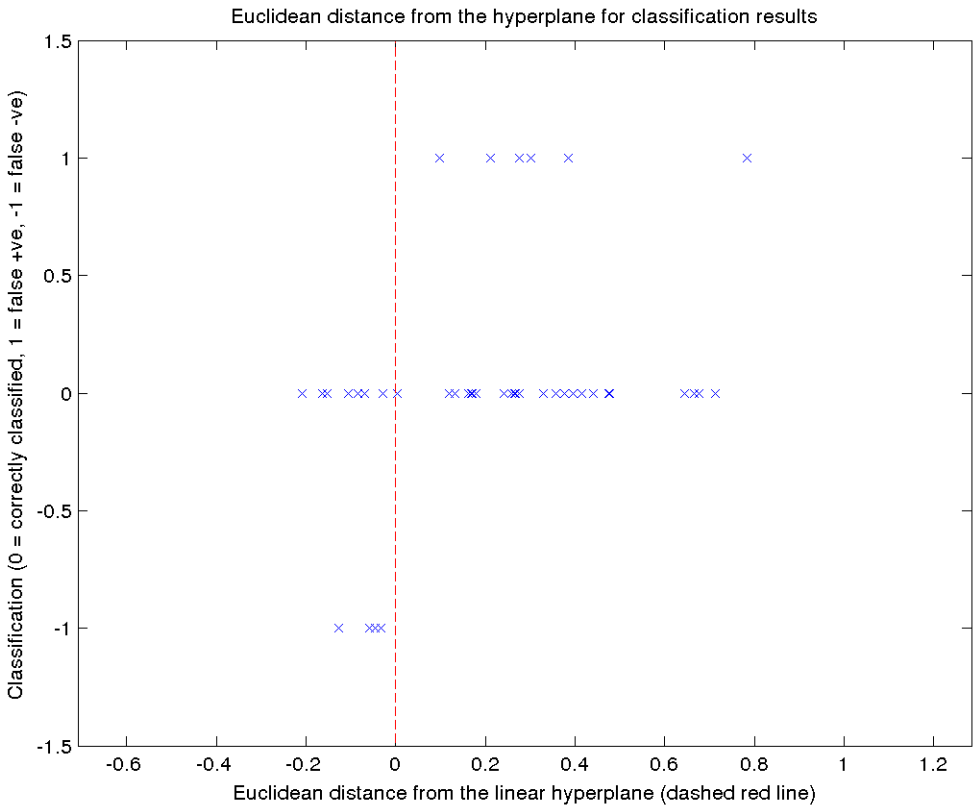

Figure 7: Individual subject's Euclidean distance from the linear hyperplane in the classification which achieved 76.7%. If the y-value = 0 the subject was correctly classified, if the y-value = 1 the subject was a non-responder who was predicted to be a responder (false positive) and if the y-value = -1 the subject was a responder who was predicted to be a non-responder (false negative).

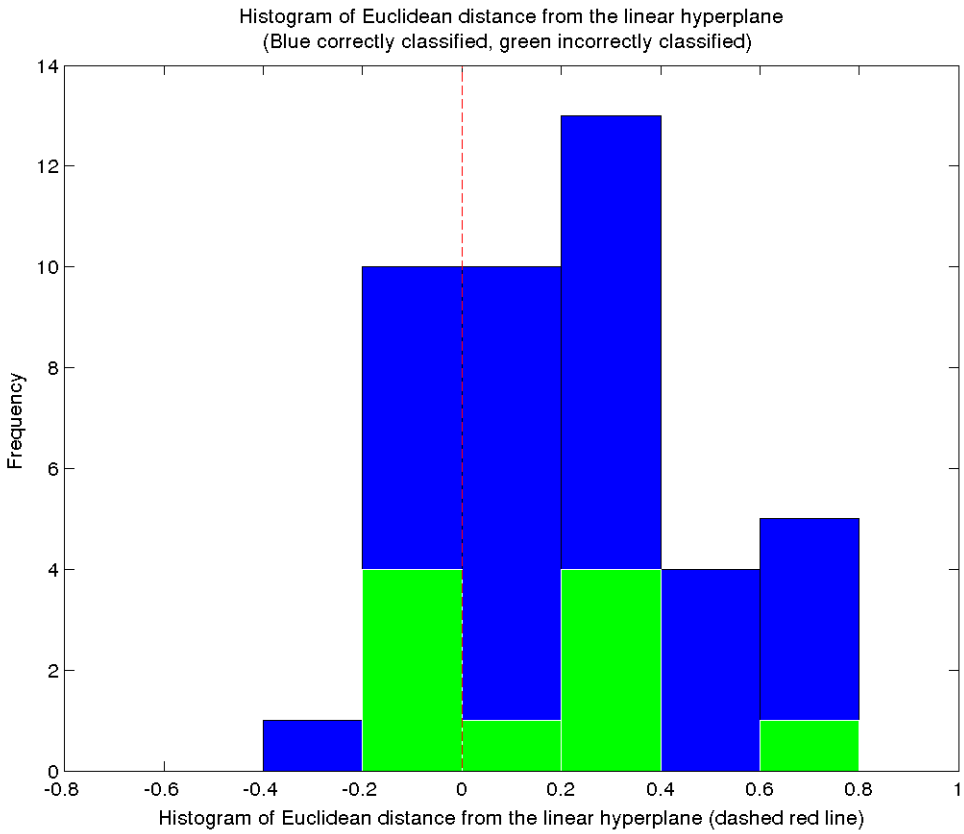

Figure 8: Histogram of individual subject's Euclidean distance from the linear hyperplane in the classification which achieved 76.7%. The blue bars show the correctly classified subjects and the green bars show the incorrectly classified subjects.

As this investigation is attempting to determine whether incorrectly classified subjects are located close to the hyperplane, the class labels of responder and non-responder are not directly relevant – the important labels are correctly or incorrectly classified and the absolute Euclidean distance. Therefore the responders and non-responders groups can be combined by taking the absolute value of the Euclidean distances. This means the number of datasets is larger and also the difference in group sizes between responders and non-responders is no longer an issue. Figure 9 shows the histogram of the absolute Euclidean distance for the correctly and incorrectly classified subjects. Other than what appears to be an outlier in the incorrectly classified group, all incorrectly classified subjects are relatively close to the hyperplane whereas the correctly classified subjects tend to have a higher proportion of subjects further from the hyperplane. This is emphasised in Figure 10 which shows the ratios of the correctly (blue line) and incorrectly (green line) classified subjects to total subjects within each of the histogram bins shown in Figure 9. As the classification accuracy was 76.7% it is unsurprising that there is a higher ratio of correctly classified subjects than incorrectly classified subjects throughout, however, the difference between the ratios (as shown by the dashed magenta line) shows that there is a clear peak in the difference further away from the hyperplane, which only decreases due to the one incorrectly classified subject that is a suspected outlier due to its large Euclidean distance from the hyperplane. Given that all the other incorrectly classified subjects are clustered around the hyperplane and the correctly classified subjects have a higher ratio of subjects with high Euclidean distances from the hyperplane, it gives support to the idea that the Euclidean distance from the hyperplane could be used as a measure of the confidence of a classification. To investigate this idea further, the results would require replication in a larger dataset and also the approach used to calculate the Euclidean distance would need to be modified to be able to apply it to non-linear kernels.

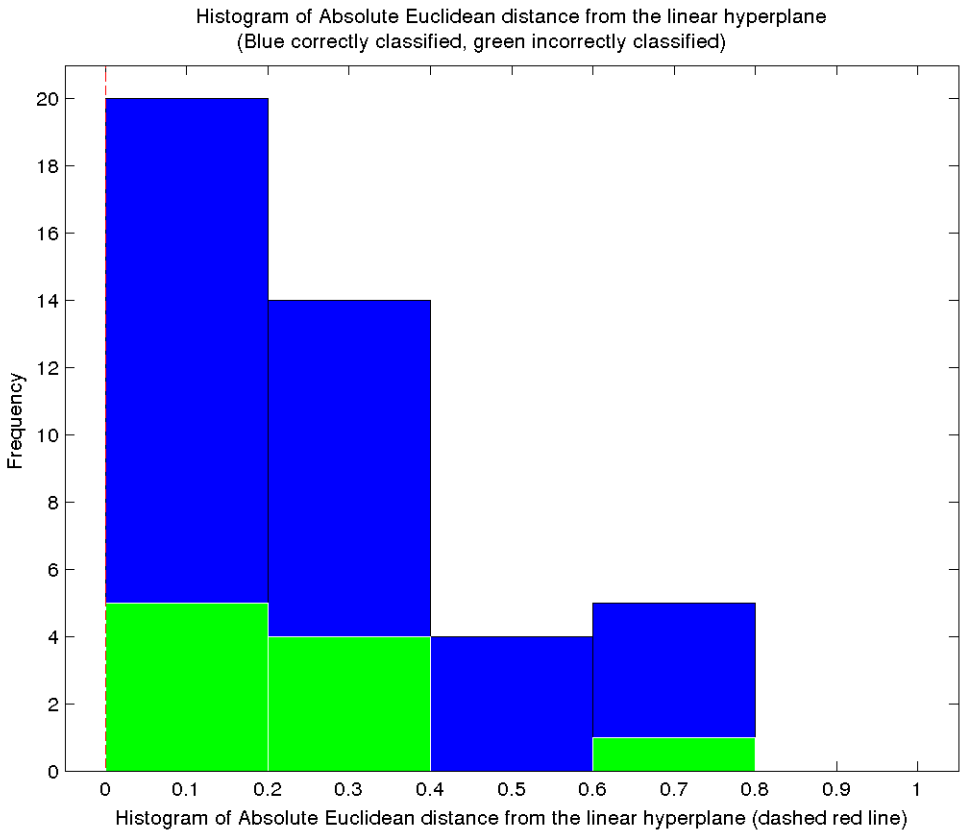

Figure 9: Histogram displaying individual subject's absolute Euclidean distance from the linear hyperplane in the classification that achieved 76.7%. The blue bars show the correctly classified subjects and the green bars show the incorrectly classified subjects.

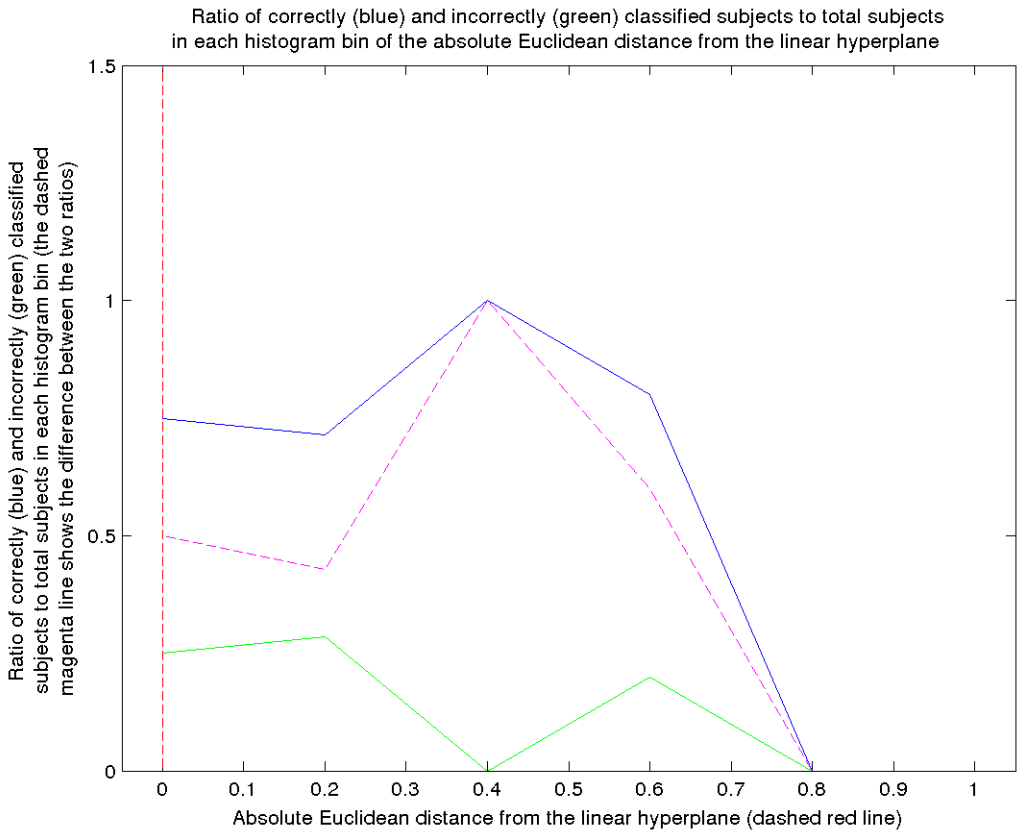

Figure 10: A plot of the ratio of correctly (blue) and incorrectly (green) classified subjects to total number of subjects within the histogram bins displayed in Figure 9. The dashed magenta line shows the difference between the ratio of correctly classified subjects (blue) and the ratio of incorrectly classified subjects (green).

4.4 Discussion

Variables which were not selected during classification are not necessarily without predictive value. For example, if variables are strongly correlated then some may be removed during feature selection, as multiple variables containing similar information do not improve classification (Guyon and Elisseeff, 2003). An uncorrelated set of variables is best for multivariate analyses. For example, the type 1 Go/NoGo task was selected more often in the prediction of MPH response than the type 2 Go/NoGo task. As 9/13 of the variables failed the Shapiro-Wilk test for normality, which is required for parametric statistical testing (e.g. Pearson's coefficient of correlation), a non-parametric Spearman's rank-order (Spearman's rho) correlation analysis was performed. This revealed that the RTT and the ERD scores from each of the two types of Go/NoGo task were significantly correlated with the corresponding variable in the other task type (RTT: $\rho = 0.76$, $p<<0.001$, ERD: $\rho = 0.78$, $p<<0.001$). Therefore the selective inclusion of type 1 tasks is most likely due to the strong correlation between these variables. Similarly, the unexpected omission of the DMtS score may be explained by the strong and significant correlations with the Parents ADHD Conners' ($\rho = 0.44$, $p = 0.004$) and pattern recognition ($\rho = 0.35$, $p = 0.023$) scores.

Achieving a highly significant prediction of 77% for MPH response is an encouraging step towards a reliable method which could allow children to avoid a trial of medication which would prove to be ineffective, as defined by the Jacobson-Truax method, or cause significant side-effects. However, it should be noted that the Jacobson-Truax criteria may be considered too stringent when describing clinical response. Further investigation is required in order to determine the boundary between response and non-response to treatment. The link between the Euclidean distance from the hyperplane and confidence of classification also merits further investigation. Nevertheless, if more sociodemographic, clinical and neuropsychological measures were available it may be possible to increase classification accuracy further. It is more likely, however, that further improvement could be obtained by combining the best sociodemographic, clinical and neuropsychological measures with genetic and/or neuroimaging data, as suggested by Denney and Rapport (1999).

Chapter 5: ADHD diagnostic classification using structural MRI data

5.1 Introduction

Being able to classify ADHD patients vs. healthy controls using MRI scans would be extremely valuable because, at present, diagnosis remains an entirely subjective clinical discipline, as there are no reliable biomarkers of ADHD to aid clinical practice. The brain regions that consistently differ between groups can point to biomarkers of the disorder and potentially elucidate the mechanisms behind ADHD.

The main pharmacological treatment for ADHD is MPH, although many other medications exist such as dextroamphetamine and atomoxetine. All of these medications act to increase the release of the chemicals dopamine and noradrenaline. The dopaminergic and noradrenergic systems are commonly reported to be abnormal in ADHD (Del Campo *et al.,* 2011). The dopamine system is most commonly found to be abnormal, and the dopamine rich basal ganglia is consistently identified to have decreased volume in children with ADHD (Ellison-Wright *et al.,* 2008; Frodl and Skokauskas, 2012; Nakao *et al.,* 2011). The basal ganglia have also been reported to be abnormal in adult ADHD as Volkow *et al.* found reduced dopamine release and reduced D_2 receptors in the caudate (Volkow *et al.,* 2007b) and ventral striatum (Volkow *et al.,* 2007a). Although the therapeutic mechanisms of the medications used to treat ADHD are not yet fully understood, it has been postulated that they improve behavioural and cognitive abnormalities by correcting for an underlying hypo-dopaminergic disorder (Del Campo *et al.,* 2011).

In addition to the more frequently reported and studied dopamine abnormalities, potential abnormalities of the noradrenaline system have been investigated (Arnsten, 1998; Arnsten *et al.,* 1996; Del Campo *et al.,* 2011; Levy and Swanson, 2001). Noradrenergic cell bodies are primarily located in the brainstem locus coeruleus and send axonal projections to the prefrontal cortices, supporting cognitive functions (e.g. response inhibition, working memory) aspects of neuropsychological functioning which are regularly reported as abnormal in ADHD (Arnsten and Li, 2005; Del Campo *et al.,* 2011; Seidman *et al.,* 2005). For example, the fact that a popular treatment, atomoxetine, is a selective blocker of the noradrenaline transporter emphasises that noradrenaline dysfunction may be a key

component of ADHD (Chamberlain *et al.*, 2006), however, to date, there have been no imaging studies of the noradrenaline system in ADHD. It is feasible that atomoxetine-increased noradrenaline function works by correcting a hypo-noradrenergic abnormality, in addition to the suggested hypo-dopaminergic disorder.

The main noradrenergic nuclei and some of the main dopaminergic nuclei are contained within the brainstem. As above, dopamine and noradrenaline abnormalities have been reported to be abnormal yet no neuroimaging studies have investigated whether brainstem abnormalities are present in ADHD without comorbidity. One study that investigated children with both epilepsy and ADHD reported a brainstem volume reduction compared to both epilepsy alone and healthy controls (Hermann *et al.*, 2007). The lack of anatomical brainstem studies in ADHD can be explained by the fact that the brainstem is a difficult brain region to image and analyse and it has been suggested that it requires specialised methods (Diedrichsen, 2006; Diedrichsen *et al.*, 2009). A recent study by Lim *et al.* (2013) identified the grey matter within the brainstem as a significant region when distinguishing children and adolescents with ADHD from healthy controls but this was closer to the midbrain than the dopaminergic and noradrenegic nuclei in the brainstem.

As mentioned in Chapter 2, the sophisticated image processing algorithm, DARTEL, has been developed recently. It has been suggested that DARTEL provides more accurate whole brain normalisation than standard VBM when pre-processing whole brain structure. A study has compared an atlas created using DARTEL against the SUIT atlas (a spatially unbiased, high-resolution atlas template of the human cerebellum and brainstem) (D'Agata *et al.*, 2011). In this manuscript the authors state that the DARTEL-created atlas performs equally well when compared with the specialist cerebellum and brainstem atlas. Notably, the SUIT atlas has recently been updated using DARTEL (http://www.icn.ucl.ac.uk/motorcontrol/imaging/suit.htm), adding support to the accuracy of this method in the brainstem.

The majority of structural brain imaging studies use conventional VBM or volumetric analyses, which test for hypothesised *group* level (e.g. ADHD vs. control) structural and functional brain abnormalities. However, whilst these differences may be reasonably reproducible at a group level, they are subtle and inter-individual variation is substantial, therefore they cannot address how specific any abnormalities are for individual patients. Consequently, it has long been established that it is not

possible to use traditional radiological (qualitative) categorisation of *individual* scans as an aid to diagnosis of these disorders. These limitations lead naturally toward the introduction of new methods, such as combining machine learning techniques with automated selection of informative brain regions through feature selection, to train diagnostic classifiers. These methods have been reported to make highly accurate predictions in adults with Major Depression (Mwangi *et al.*, 2012a), Alzheimer's Disease (Klöppel *et al.*, 2008) and Autism Spectrum Disorder (Ecker *et al.*, 2010a). The only study to date (excluding studies which used the ADHD-200 dataset due to the data quality concerns, discussed in Chapter 3) which has managed to apply machine learning methods to childhood ADHD structural MRI data was published recently (Lim *et al.*, 2013). In this study, the grey matter component of structural MR images of children and adolescents with ADHD, ASD and healthy controls were successfully classified (79% accuracy when predicting between ADHD vs. controls). The study did not investigate the white matter differences, which may have improved the classification accuracy further.

At a group level, there are now extensive data that indicate subtle differences in brain structure between subjects with ADHD and typically developing controls. Total brain volume has been reported as 'slightly but significantly smaller' (Kelly *et al.*, 2007; Seidman *et al.*, 2005). A number of studies have investigated the dopamine rich basal ganglia in ADHD and several meta-analyses reported reduction in the volume of the putamen, caudate and pallidum in ADHD (Ellison-Wright *et al.*, 2008; Frodl and Skokauskas, 2012; Nakao *et al.*, 2011). Interestingly, Castellanos *et al.* (2002) reported that the caudate abnormality may normalise as a child matures towards adulthood, which may be clinically relevant, as the caudate is associated with motor activity and there is often a relative reduction in hyperactivity later in development.

The other major brain region that is often assumed to play a prominent role in the development of ADHD is the prefrontal cortex (PFC). Again, reductions in volume have been described for the PFC (Seidman *et al.*, 2005). More recently, evidence for significant reductions in volume in other regions have been described including the vermis of the cerebellum (Berquin *et al.*, 1998; Bussing *et al.*, 2002; Castellanos *et al.*, 1996; Hill *et al.*, 2003; Mostofsky *et al.*, 1998), and the temporal, parietal and occipital lobes (Castellanos *et al.*, 1996). Each of these regions are associated with important neuropsychological functions which have been reported to

be compromised in many individuals with ADHD (Coghill *et al.*, 2005). Amygdala volume has been reported to show no abnormalities (Castellanos *et al.*, 1996; Filipek *et al.*, 1997) but more recently a decrease in volume (Plessen *et al.*, 2006). Regions which had been studied previously but not found to exhibit significant abnormalities include the insula (Filipek *et al.*, 1997; Hynd *et al.*, 1990) and hippocampus (Castellanos *et al.*, 1996; Filipek *et al.*, 1997). However, these two regions were identified by Lim *et al.* (2013) as relevant to their ADHD vs. controls classification study. Group level differences between white matter are infrequently reported (Hermann *et al.*, 2007). Seidman *et al.* (2005) reported white matter reduction in the corpus callosum and Carmona *et al.* (2005) reported no differences in white matter volume.

Feature selection is an important aspect of the present study as there are many brain regions that do not provide useful information for diagnostic prediction. Inclusion of these regions impairs the accuracy of prediction (Johnston *et al.*, 2012). Automated feature selection identifies brain regions supporting high accuracy individual classification, and therefore localises structurally abnormal brain regions.

The cross-validation procedure used in this analysis was LOOCV. This approach is ideal for clinical use, as it maximises the available data for 'training', whilst not assuming prior knowledge of diagnostic status for the 'left out' test subject. In cross-sectional studies, LOOCV is repeated with a different subject left out until all scans have been predictively classified; in longitudinal studies the process is repeated as new subject data are acquired and the outcomes of previous predictions become known.

The present study used DARTEL and feature selection with SVM and LOOCV, to develop a method for predicting, with best accuracy, *individual* diagnostic status (ADHD vs. controls) using T_1 weighted structural MRI scans. Predictive classification has previously been successfully applied to individual resting state fMRI scans (Zhu *et al.*, 2008; 2005) and to functional connectivity data extracted from resting state fMRI scans (Liang *et al.*, 2012) in ADHD subjects. T_1 weighted imaging has similar advantages to resting state fMRI in not requiring comprehension and cooperation with a paradigm, but also has the additional advantage of being more readily available at scanning centres and to provide better anatomical localisation than fMRI. The main hypothesis was that high accuracy classification would be achieved using brain regions automatically selected during

feature selection such as the brainstem and basal ganglia, given the common pharmaceutical action of medications used to treat ADHD, and the anatomical locations of dopaminergic and noradrenergic nuclei.

5.2 Methods

5.2.1 Subjects

Structural T_1 weighted scans were acquired by Dr Kerstin Konrad's group at the Research Centre in Juelich, Germany, and the Department of Child and Adolescent Psychiatry of Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Germany, from subjects participating in neuroimaging studies. Informed consent was obtained from all volunteers and their parents according to the Declaration of Helsinki. The study protocols were approved by the Ethics Committee of the RWTH, Aachen University Hospital, Germany. Volunteers were compensated for participation in the study.

Of the thirty-five males with a diagnosis of ADHD who were recruited from the Department of Child and Adolescent Psychiatry and Psychotherapy in Aachen, thirty-four were included in this analysis. The subject which was removed from the analysis was excluded to ensure there were no significant differences in age or IQ between groups. Initial diagnosis was made by experienced clinicians according to the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) (American Psychiatric Association, 2000) criteria and confirmed by an independent rater using a semi-structured diagnostic interview: either the Kiddie-Sads-Present and Lifetime Version (K-SADS-PL) (Kaufman *et al.*, 1997) or "Diagnostische Interview bei Psychischen Störungen im Kindes- und Jugendalter" (K-DIPS) (Schneider *et al.*, 2009). All parents were asked to complete a German Questionnaire on ADHD symptoms, the FBB-HKS (Döpfner and Lehmkuhl, 1998), which includes DSM-IV and International Classification of Diseases-10 (ICD-10) items for ADHD diagnosis. Three subjects in the ADHD group fulfilled additional criteria for Externalising Disorders (oppositional defiant disorder and conduct disorder) and one subject had comorbid Dyslexia. Exclusion criteria included potentially confounding diagnoses such as Psychosis, Mania, Major Depression or Substance Misuse. Ten ADHD participants were being treated with either short- or long-acting MPH (Ritalin,

Concerta or Equasym) which was stopped at least 48 hours before scanning. None were taking any other psychotropic drugs.

Fifty-five male typically developing controls were recruited from local schools and underwent an extensive psychiatric examination using the same standardised, semi-structured interviews as the ADHD volunteers. To make the groups balanced and to ensure there were no significant differences in age or IQ between groups thirty-four of the recruited controls were used in the analysis. None of these controls had a history of current or past psychiatric or neurological disorder and none were taking medication. The ADHD and control volunteers had an Intelligence Quotient above 80 as assessed by either the Culture Fair Intelligence Test 20 (Weiß, 1998) or the Wechsler Intelligence Scale for Children (WISC version III or IV) (Wechsler, 1991; Wechsler, 2004). Handedness was assessed using the Edinburgh Handedness Inventory (EHI) (Oldfield, 1971). Apart from two left-handed subjects in the ADHD group and one ambidextrous subject in the control group, all subjects were right-handed.

Age and IQ did not differ significantly (t-test, $p>0.1$) between groups. The ADHD group mean age was 12.5 years (standard deviation 2.3) mean IQ was 99.8 (standard deviation 11.5). The control group mean age was 13.2 years (standard deviation 1.0) and the mean IQ 103.7 (standard deviation 10.0).

Of the thirty-four ADHD children and adolescents included in this study, five had inattentive-type ADHD, one had hyperactive-impulsive-type ADHD and the remaining twenty-eight had combined-type ADHD.

The FBB-HKS (Döpfner and Lehmkuhl, 1998) questionnaire provides a syndrome severity score. These symptom severity scores can be categorised as raw inattention, hyperactivity and impulsivity scores, ranging from 0-25, 0-20 and 0-11 respectively, with higher scores indicating increased symptom severity. The total scores are obtained by summing individual scores allowing total percentiles to be calculated.

The mean inattention, hyperactivity and impulsivity scores for the ADHD group were 17.1, 13.1 and 6.9 respectively, and for the control group 1.9, 1.2 and 1.1 respectively. The mean total scores for the ADHD and control groups were 37.2 and 4.2 and the mean percentiles were 94.4 and 24.4 respectively.

A particular strength of the study is that the dataset is relatively heterogeneous – particularly with the comorbidity, medication history and fairly

wide age range during a time of dynamic brain development. If high classification accuracy can be achieved using this dataset then it gives more confidence that the technique would achieve similar results in the general population.

5.2.2 Image Acquisition

For each participant structural whole-brain images were acquired using a 1.5T Siemens Sonata scanner (Siemens, Erlangen, Germany) using an isotropic T_1-weighted MP-RAGE (magnetisation-prepared rapid acquisition gradient echo) sequence with the following parameters: TR (repetition time) = 2200 ms, TE (echo time) = 3.93 ms, flip angle = 15°, FOV = 256 mm, matrix = 180 x 256, 160 slices, voxel size 1x1x1 mm, slice thickness 1 mm.

5.2.3 Image Pre-processing

In order to check that the data quality was of an acceptable standard, all scans were visually inspected for artefacts and particular care was taken to identify motion artefacts which appear as blurring or 'ghosting' (McRobbie *et al.*, 2010). No scans showed blurring, ghosting or other gross artefacts. No scans were excluded from analysis.

Due to the hyperactive and inattentive symptoms of ADHD, a potential limitation of the study is that the classifier predicts on the basis of movement during the scan, rather than syndrome linked structural brain differences. However, Yerys *et al.* (2009) suggested that the age of the subject may be more significant than diagnosis when investigating fMRI scan success rates as subjects both on and off MPH and healthy controls aged 10-12 achieved increased scan success rate compared with their respective 7-9 age groups (Yerys *et al.*, 2009). In the present study there were no significant differences in age between groups.

To exclude the possibility that more motion artefact is present in the ADHD group's scans than the control group's, spatial autocorrelation (Slotnick and Schacter, 2006) was investigated (https://www2.bc.edu/~slotnics/scripts.htm). Spatial autocorrelation determines the level of correlation between pixels (or voxels in the 3D case) in an image. When an object is smoothed, the level of spatial autocorrelation increases as the voxels become more correlated (i.e. with a larger smoothing radius). As mentioned previously, motion during a structural MRI scan

will appear as blurring or ghosting (McRobbie *et al.*, 2010). Although there were no visible artefacts present, if more subtle systematic blurring or ghosting occurred then the level of spatial autocorrelation would be higher. The reasons for this are because a blurry region within an image would be equivalent to smoothing and ghost images are highly correlated with the original image.

Therefore, to check that classification was not based on a systematic motion artefact in the ADHD group rather than genuine structural brain differences, the spatial autocorrelation values (in each of the three dimensions) were calculated from the raw structural MR images from both groups. No significant differences were identified between the two groups (x-direction: $p = 0.57$, y-direction: $p = 0.50$ and z-direction: $p = 0.97$).

Pre-processing was performed using the DARTEL toolbox (Ashburner, 2007) as implemented in SPM8 (http://www.fil.ion.ucl.ac.uk/spm). As described in Chapter 2, the DARTEL procedure involves segmentation of T_1 weighted images into separate grey matter, white matter and CSF compartment images and the creation of a study-specific anatomical template for spatial normalisation. Creation of a study-specific template was important in this study as participants were at an earlier stage of development than the adults who contributed to the default SPM8 anatomical template. The DARTEL procedure included modulation to control for potential spatial normalisation rescaling problems (Ashburner, 2007). The resultant images were smoothed with an 8 mm full-width at half-maximum (FWHM) Gaussian kernel.

5.2.4 Group Level Comparisons

For a conventional *group* level VBM analysis, the null hypothesis of no difference in brain structure between ADHD participants compared to controls was tested using an unpaired t-test, as implemented in SPM8. Significance was defined as $p<0.005$ and was implemented using a simultaneous requirement for a voxel threshold of $p<0.005$ and clusters to exceed 139 contiguous voxels. These parameters were identified using a customised version of a popular Monte-Carlo neuroimaging algorithm (Slotnick *et al.*, 2003).

The customisation to the Monte-Carlo neuroimaging algorithm was required as the original code was designed to deal with lower resolution images such as fMRI rather than high resolution structural scans. The modifications included decreasing the resampling resolution (to decrease the computational strain during the Monte-

Carlo calculations) and increasing the range of values which the cluster size could take in the raw images (to allow for higher resolution input images). These modifications were approved by the author of the original Monte-Carlo neuroimaging algorithm, Scott Slotnick *(personal communication)*.

5.2.5 Individual Scan Classification

Machine learning allowing individual predictions was implemented in Matlab (The Mathworks Inc.) using an SVM toolbox (Schwaighofer, 2001) and custom Matlab scripts. As described previously, SVM analysis consists of two stages: training the classifier, then testing the accuracy using data not used for training.

To maximise the size of the training data, LOOCV (Cristianini and Shawe-Taylor, 2000) was used. This procedure involves removing a subject (from either group) and using the remaining subjects as the SVM training set. This process is repeated until each subject is left-out once. It is important to ensure that no information is leaked from the training data to the testing data.

As expected, a linear SVM classifier which used voxels from the whole brain achieved poor predictive accuracy, therefore a 'feature selection' method, which selected localised regions of the brain for SVM analysis in an automated manner (Bray *et al.*, 2009; Johnston *et al.*, 2012; Klöppel *et al.*, 2008) was used. Poor predictive accuracy when using whole brain data is unsurprising because, when a large number of voxels are used with an SVM, most of these voxels are redundant (Dash and Liu, 1997). Feature selection can be very successful as it excludes many voxels that confer no useful information for prediction, but introduce 'noise' and correlated information, so degrading classifier performance. Feature selection was applied to a linear and a non-linear (Gaussian) SVM to investigate whether this improved predictive accuracy.

A simple feature selection approach was used: the 'mean-threshold' method (described in Chapter 2). In order to ensure no prior information about the left out subject was 'leaked to testing' during feature selection, feature selection was performed during the parameter selection stage (inner LOOCV procedure) only. This ensured the features (brain regions) selected for classification were entirely independent from the 'held-out' image which was predictively classified.

In order to identify the optimal threshold for each leave-one-out loop, a broad range of potential values were investigated. Within each (outer) leave-one-out loop the starting range varied from 0.05 to the maximum absolute difference between groups in the training set (a range chosen to be large enough such that it would always include the optimal cut-off value). As it would take an unfeasibly long time to tune the parameters (2 for the linear SVM - the soft-margin parameter and the optimal cut-off value - and 3 for the Gaussian SVM - the soft-margin parameter, the 'kernel width' parameter and the optimal cut-off value) for such a wide range of thresholds, the soft-margin parameter (and the 'kernel width' parameter in the non-linear SVM) was initially set to unity in order to narrow down the wide range of potential thresholds. This range was reduced by performing an inner LOOCV with the SVM parameters fixed, identifying the thresholds which achieved a high training accuracy, centring the narrowed range on these thresholds, and then reducing the step size to investigate more thresholds within the new range. To demonstrate this range reduction technique, Figure 11 displays the average training stage accuracies of all subjects at each cut-off value over the large range of potential values for the white matter prediction. Figure 12 then shows the average training accuracy in the narrowed range of cut-off values being investigated (again for white matter prediction). As the narrowed threshold differs in each LOOCV loop, the vector elements (1 to 12) were used to create the average training stage accuracies across the narrowed ranges. This demonstrates that despite the differing choices of narrowed ranges, the peak accuracy was generally central within the range investigated. There is a clear peak in accuracy of prediction which the cut-off range centres on in each case. Once the range was sufficiently narrow, the SVM parameters and the narrowed threshold range were all tuned using a second inner LOOCV and a 2/3 variable grid search procedure (depending on whether a linear/non-linear kernel was used).

The combination of parameters which achieved the highest training stage accuracy was then applied during *testing* (i.e. when classifying the novel 'held-out' subject), to assess classifier performance: accuracy, sensitivity, specificity and chi-square significance of classification. To demonstrate this grid search procedure, Figure 13 displays how the average training stage accuracy varied with soft-margin parameter and a narrowed cut-off range (the kernel-width parameter was fixed to unity for illustration purposes - for the white matter prediction). A summary of the

SVM technique is shown in Figure 14 with parameter and feature selection highlighted in more detail in Figure 15 and Figure 16. The soft-margin parameters investigated ranged from 1 to 5 and the kernel-width parameters ranged from 1 to 10.

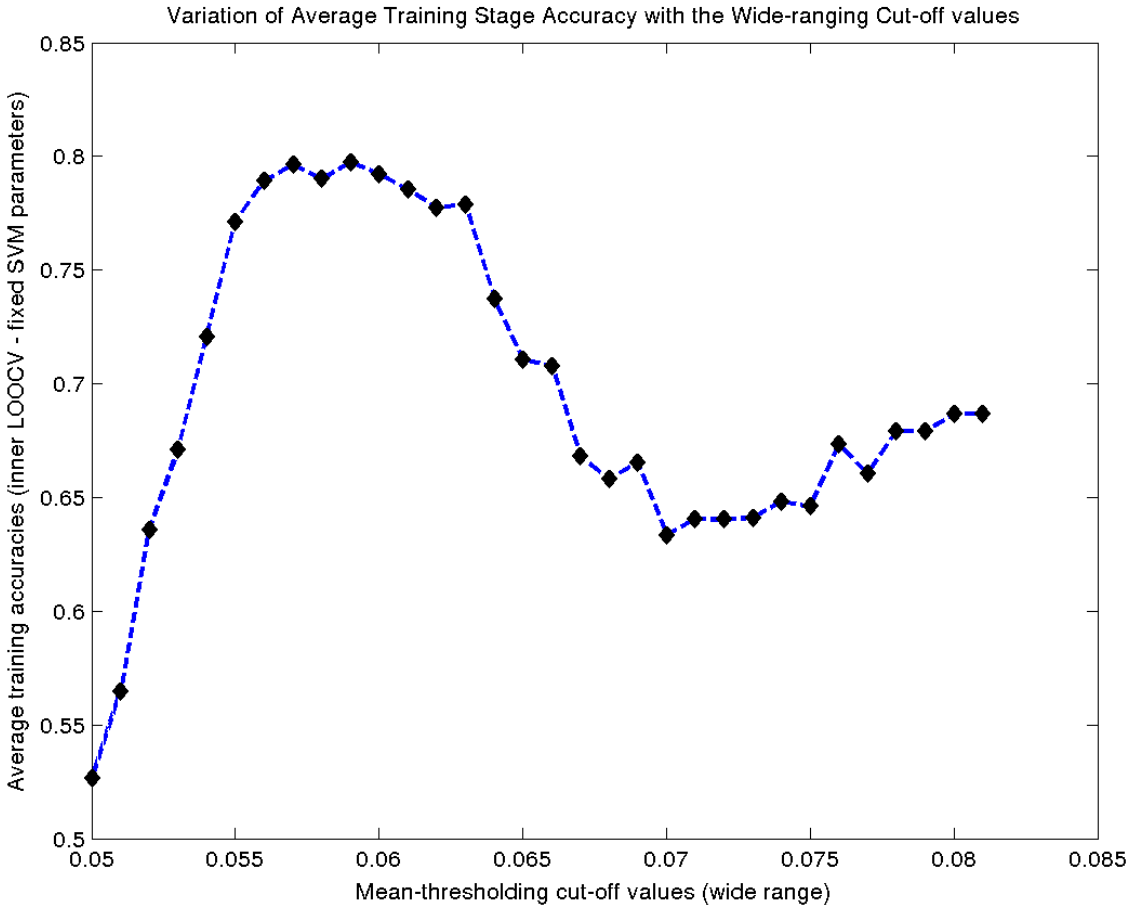

Figure 11: Gaussian SVM white matter prediction, average training accuracy as a function of the wide-ranging cut-off values (used to identify the narrow range of cut-off values for the 3-variable grid search – the soft-margin and kernel width parameters are fixed).

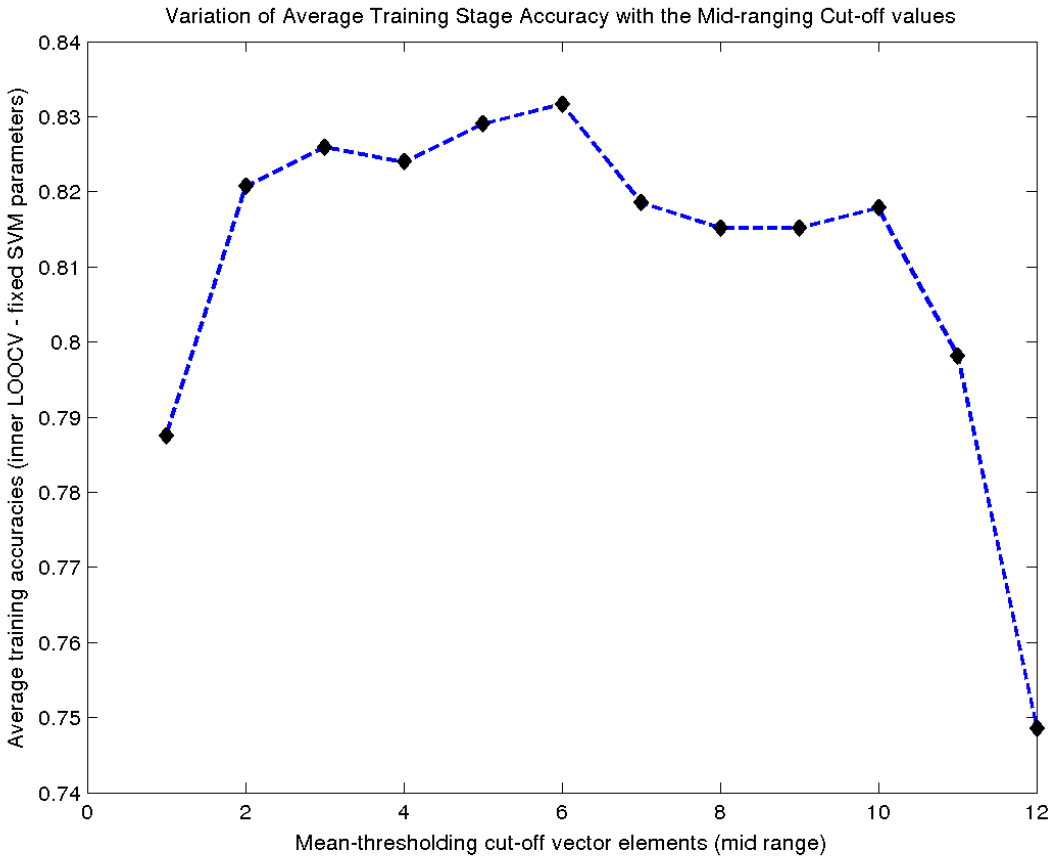

Figure 12: Gaussian SVM white matter prediction, average training accuracy as a function of mid-ranging cut-off values (the second iteration of the range reduction technique, vector elements are used instead of cut-off values as each subject differs on the range selected), used to identify the narrow range of cut-off values for the 3-variable grid search – the soft-margin and kernel width parameters are fixed.

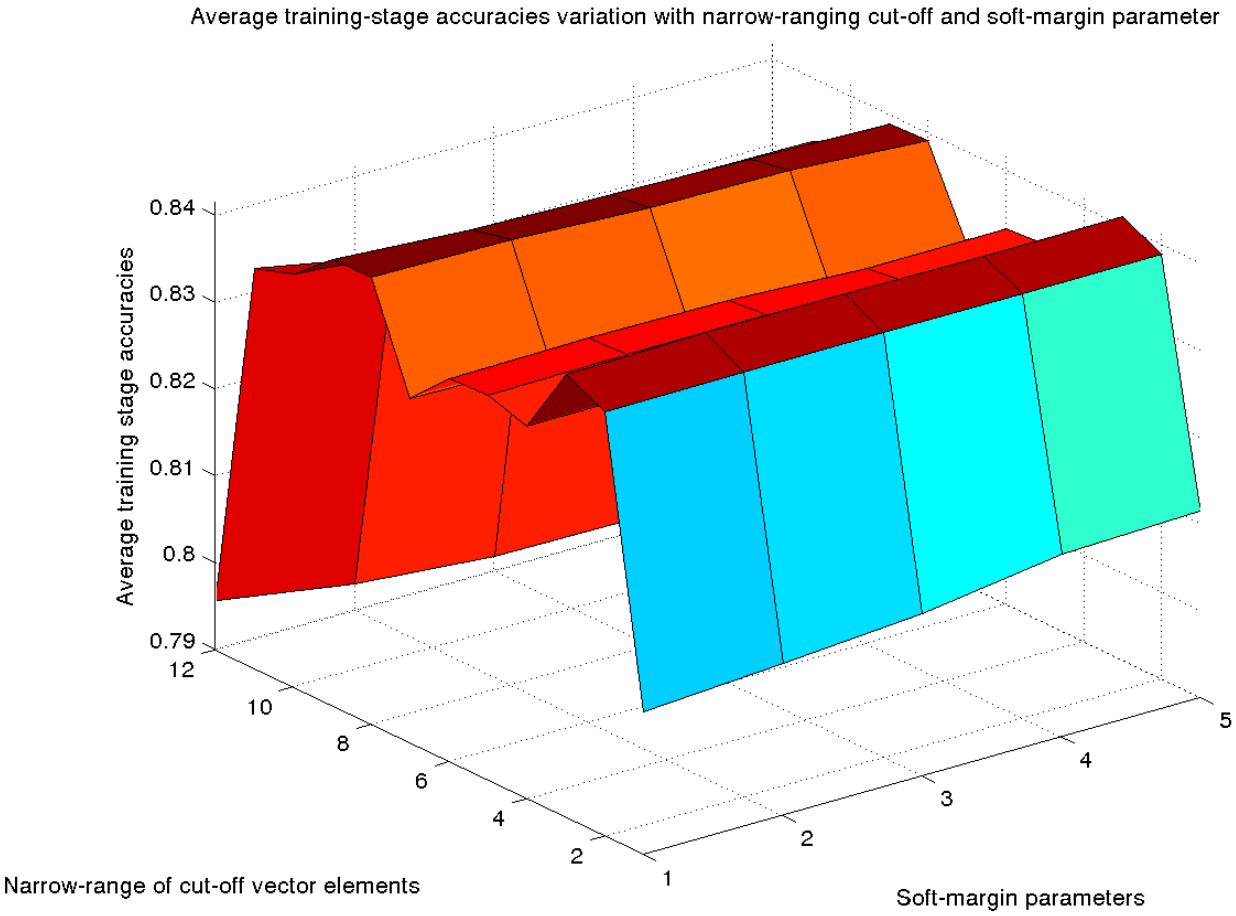

Figure 13: Gaussian SVM white matter prediction, average training accuracy versus variable soft-margin parameters and the narrowed cut-off ranges (vector elements are used instead of cut-off values as each subject differs on the range selected), in a 2-variable grid search (the kernel width parameter was fixed for illustration purposes).

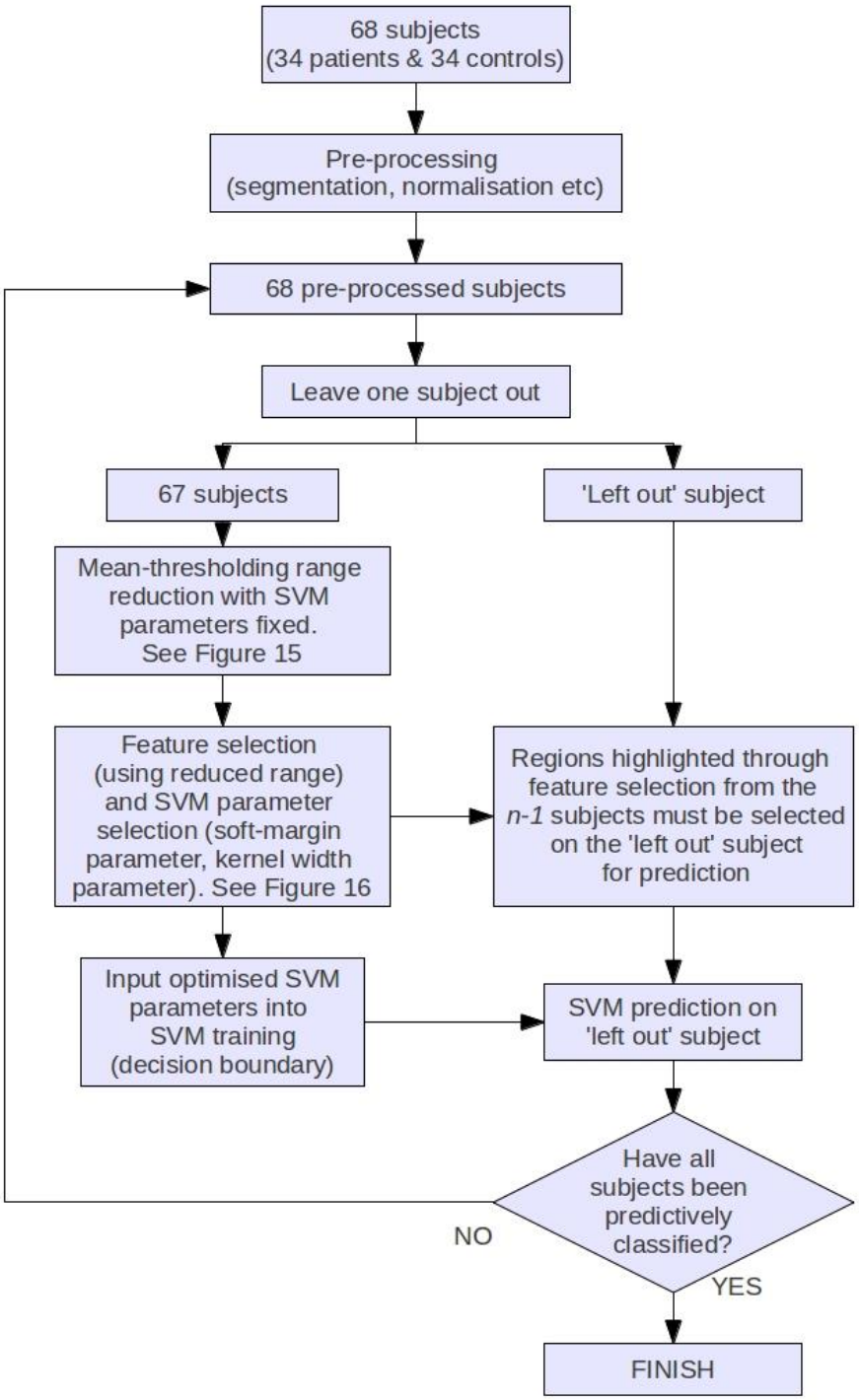

Figure 14: A flow diagram outlining the primary LOOCV procedure. The general procedure involved applying LOOCV on the pre-processed images, applying the 'mean-threshold' method of feature selection and SVM parameter tuning (shown in more detail in Figure 15 and Figure 16) to the training data, training the SVM using the training data and the optimised parameters, then making a prediction for the left-out subject.

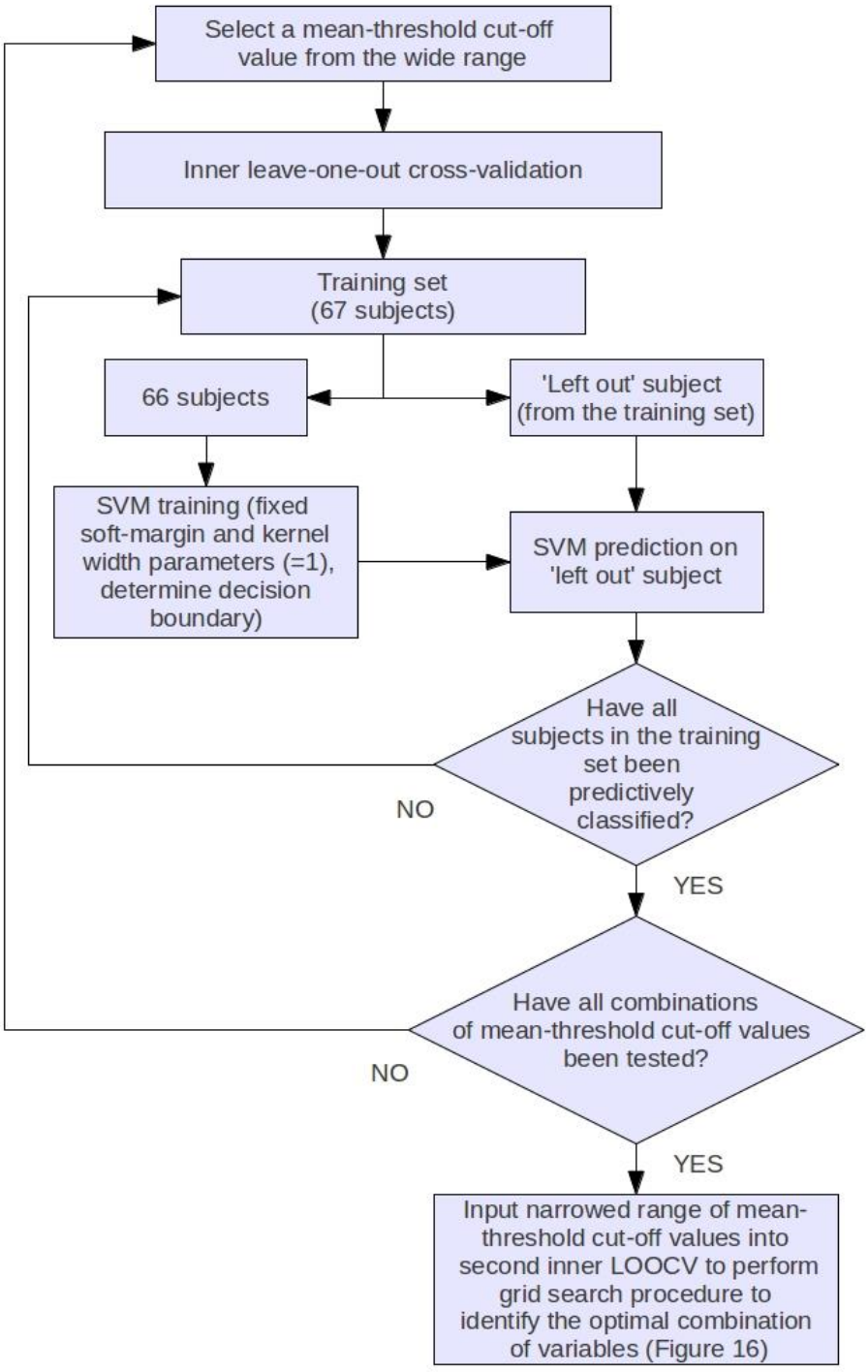

Figure 15: Parameter range reduction stage of the ‘mean-threshold’ feature selection method. In order to speed up the processing time, the range of threshold cut-off values was narrowed down by testing an arbitrarily wide range using an inner LOOCV procedure with fixed SVM parameters (fixed to unity). The selected range was based on the threshold cut-off values which achieved the highest training stage accuracy.

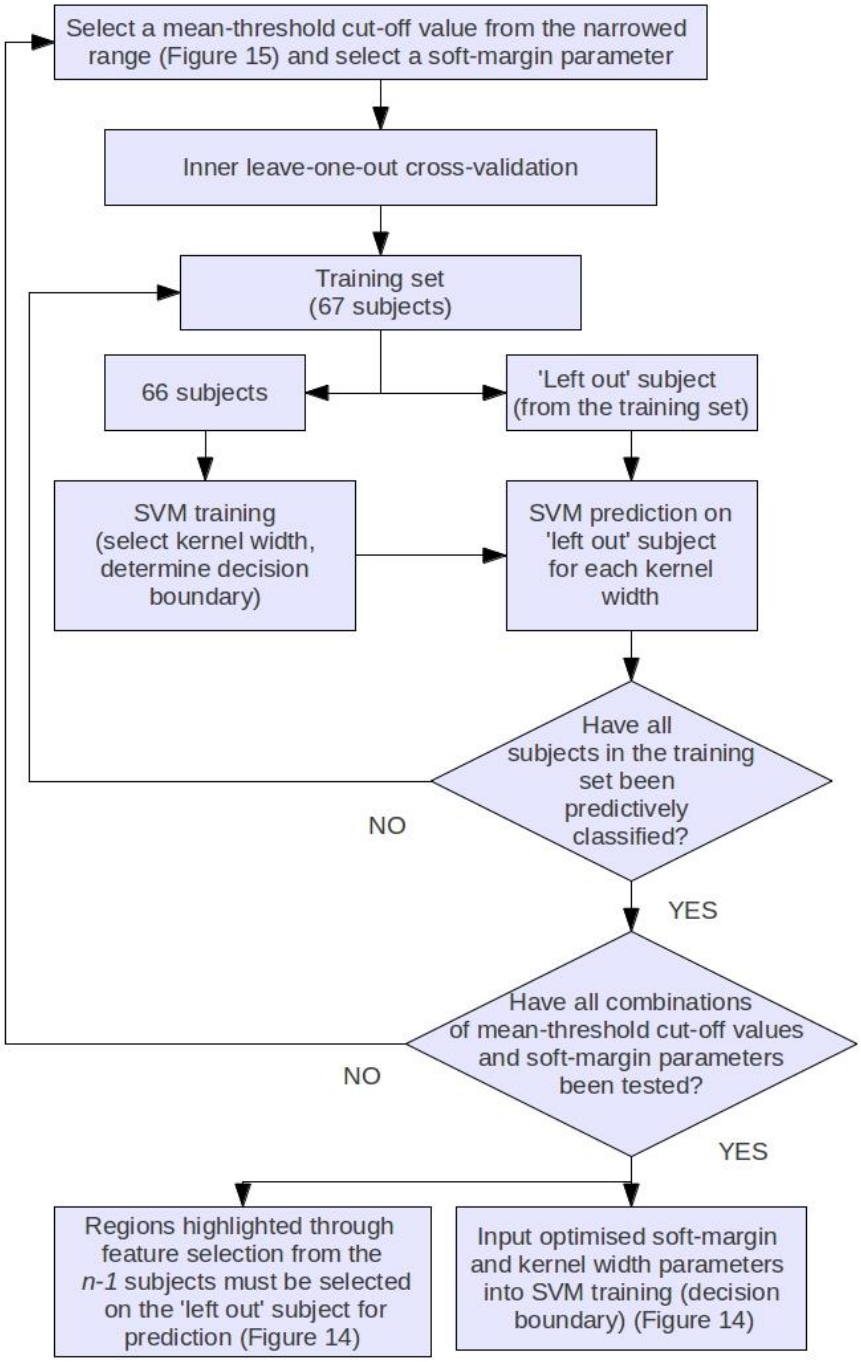

Figure 16: The grid search procedure for parameter selection (including the optimal mean-threshold cut-off value). A second inner LOOCV procedure is performed whereby all combinations of mean-threshold cut-off values (from the narrowed range identified in Figure 15) and SVM parameters are tested. The combination which achieves the highest training stage accuracy are selected for training the SVM on the full training set for the classification of the left-out subject (Figure 14).

5.2.6 Multivariate Feature Selection

As an alternative to the mean-threshold feature selection technique, a multivariate feature selection method which has also been outlined in Chapter 2: RFE was also investigated. This was to test whether the accuracy associated with using a univariate feature selection method (mean-thresholding) could be improved by a multivariate feature selection method (RFE), as the univariate mean-thresholding process ignores possible multivariate interactions between voxels.

The general approach is very similar to that used for the mean-thresholding method. As can be seen in Figure 17, the only difference from the mean-thresholding method (see Figure 14) is that the feature selection is completed prior to the parameter selection stage. The parameter selection method (Figure 19) was identical to the grid-search technique used previously (Figure 16) with the only difference being that only the SVM parameters were optimised rather than the SVM parameters and the mean-threshold value.

The theory behind RFE is that each voxel can be ranked using the SVM training stage weights to determine its importance to the classification. There are many other multivariate feature selection methods but RFE tends to be a more popular and reliable approach than most. There are many differences in the implementation of RFE; the method decided on was an 11-fold cross-validation approach, with resampling and a backwards elimination percentage of 20%.

As the training stage data contained 67 subjects (one subject was left-out for the final classification for each LOOCV stage), it was decided that 11-fold cross-validation was a sufficient compromise between balanced groups (11 folds of 6 with one subject ignored (to account for class imbalance)) and the length of time required to run the RFE code. To ensure the results were replicable and to maximise the training data, the data were resampled three times with a different subject left out each time with the collated training stage accuracy recorded to provide a robust value.

The feature selection component is performed over many iterations. The process involved beginning with the whole brain image, performing the 11-fold cross-validation and identifying the training stage accuracy (in fact the collated training stage accuracy after resampling) and the voxel weights (which determine the relative ‘value’ of that voxel towards the classification). These voxel weights were

ranked in order of importance and the lowest 20% of these voxels were removed from the analysis. The 11-fold cross-validation was then repeated with the remaining 80% of the data with the process repeating until there were less than a thousand voxels remaining (arbitrary termination criteria). The iteration which achieved the highest collated training stage accuracy was then selected as the combination of voxels which would be used in the outer LOOCV process.

5.2.7 Calculating the number of Voxels in each cluster

In order to compare the brain regions identified in the main analysis with the regions identified in a conventional VBM *group* level analysis, the location of the clusters were compared. As it can be seen in Figure 23 that far fewer voxels were used in the main analysis than were identified in the VBM analysis, the size of each cluster was also investigated. SPM outputs the number of voxels in each cluster during VBM analyses (shown in Table 3 and Table 4) but each cluster identified in the main analyses had to be identified so that the cluster sizes could be compared (shown in Table 5 and Table 6).

The author created an algorithm which could identify all the clusters of voxels used during classification and output the size of each cluster alongside other relevant information related to the regions used during classification. At each voxel selected after feature selection, the six neighbouring voxels (in three dimensions) were inspected to see if they were also selected. All the voxels' neighbours that were included in the prediction were identified as belonging to the same cluster. Each identified cluster was checked by overlaying each individual cluster over the full image containing all the brain regions identified in the main analysis (the red regions in Figure 21) with all clusters correlating exactly.

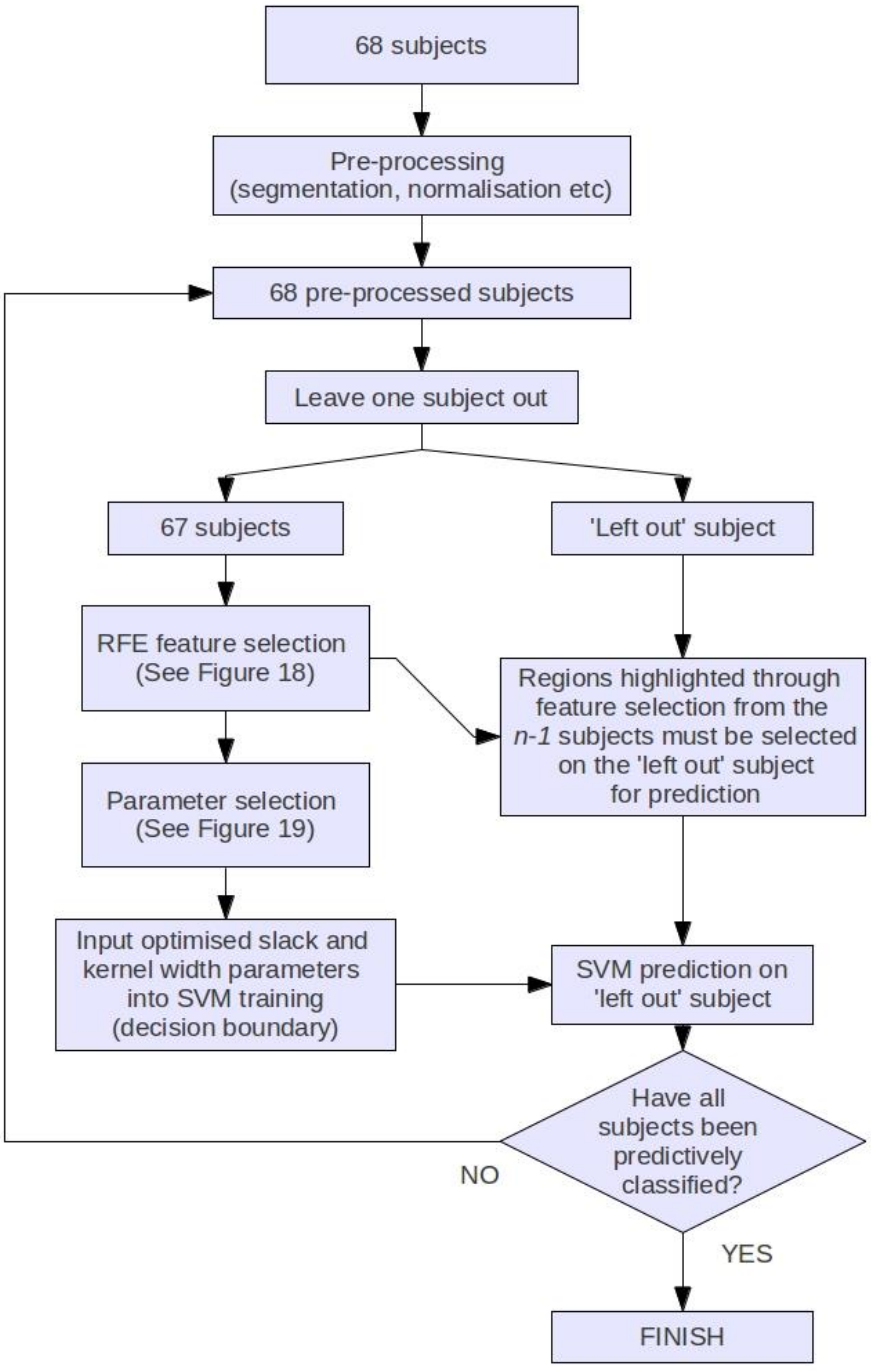

Figure 17: RFE method of feature selection with SVM.

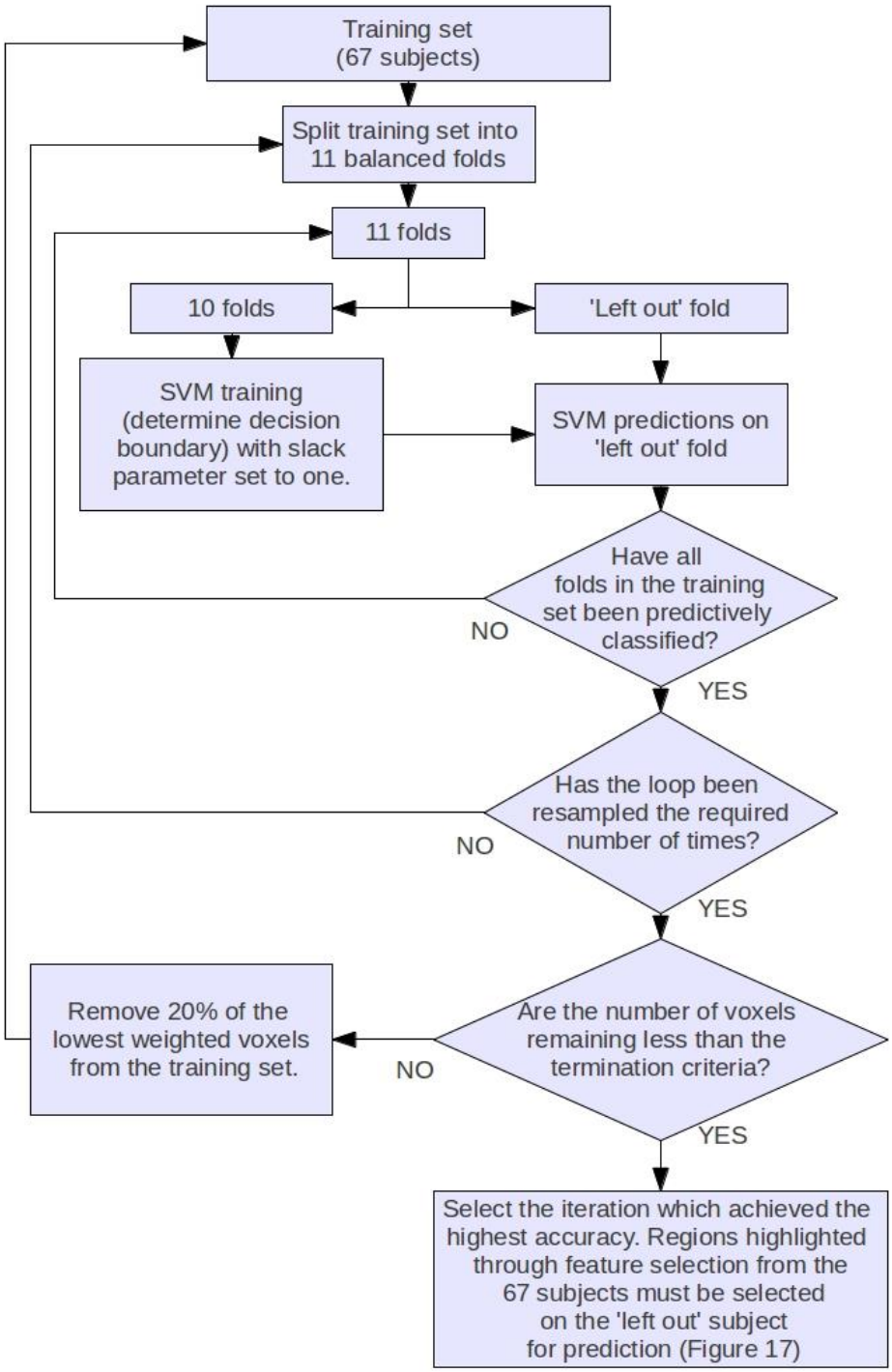

Figure 18: RFE feature selection details for Figure 17

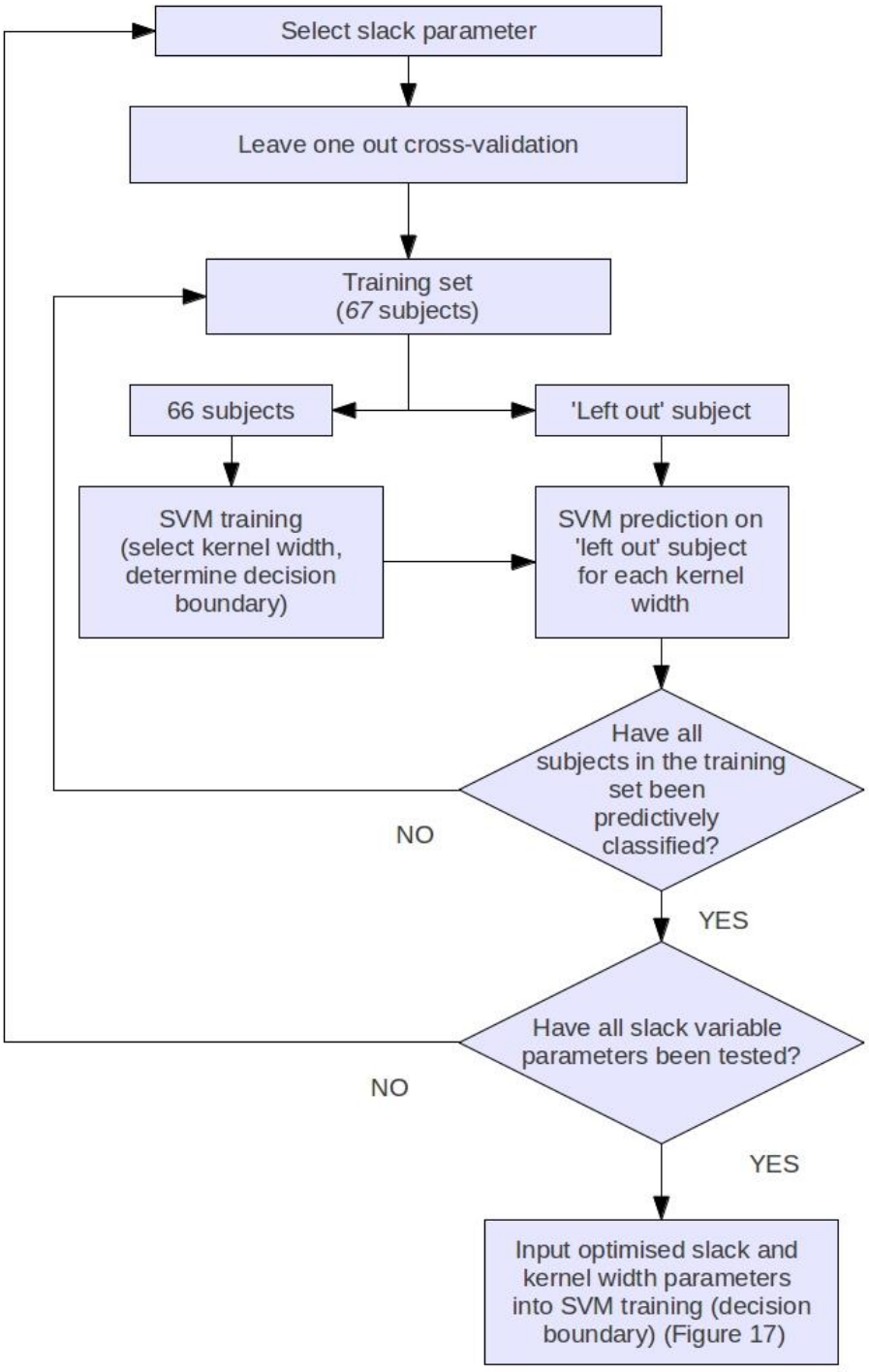

Figure 19: Parameter selection for RFE method in Figure 17.

5.3 Results

5.3.1 VBM Analysis

Compared to controls, ADHD was associated with significant *decreases* in grey matter volume in the bilateral putamen, bilateral superior cerebellum, amygdala, superior hippocampus, superior temporal gyrus, medial orbitofrontal cortex, bilateral precentral sulcus, inferior longitudinal fasciculus/lateral hippocampus and middle frontal gyrus. These are shown in Figure 20 (left) with details provided in Table 4.

Compared to controls, ADHD was associated with significant *decreases* in white matter volume which were most prominent in the brainstem, which includes the pons-midbrain junction. *Decreases* were also found in the medial superior cerebellum, pyramidal tracts, frontal medial and occipital lobe. These are shown in Figure 20 (right) with details provided in Table 3.

No grey or white matter volume *increases* in ADHD compared to controls were found.

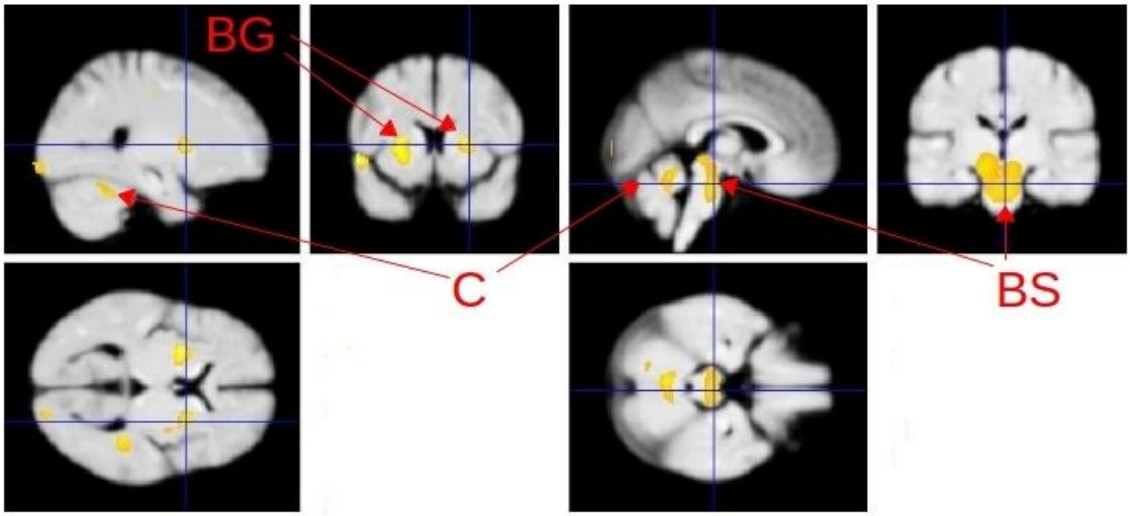

Figure 20: VBM Group Level results. Significantly ($p<0.005$) reduced grey matter (left) and white matter (right) volume in ADHD. Reduced grey matter volume in the basal ganglia (BG - putamen) and cerebellum (C) and significantly reduced white matter volume in the brainstem (BS) and cerebellum (C).

Table 3: MNI coordinates of each cluster of grey matter volume decreases.

Region	MNI coordinates			Resampled voxels per cluster	T-score
	x	*y*	*z*		
bilateral putamen	-27	6	4.5	901	3.45
	25.5	7.5	6	998	3.17
bilateral superior cerebellum	12	-39	-19.5	1021	3.18
	-24	-46.5	-28.5	756	3.18
Amygdala	-24	-3	-12	901	3.14
superior hippocampus	-30	-19.5	-6	279	3.58
superior temporal gyrus	-57	6	-3	201	3.72
medial orbitofrontal cortex	1.5	49.5	-25.5	210	3.18
bilateral precentral sulcus	31.5	-10.5	39	1268	4.44
	-39	-12	30	202	2.96
inferior longitudinal fasciculus/lateral hippocampus	40.5	-36	0	654	3.24
middle frontal gyrus	-37.5	30	22.5	145	3.11

Table 4: MNI coordinates of each cluster of white matter volume decreases.

Region	MNI coordinates			Resampled voxels per cluster	T-score
	x	*y*	*z*		
Brainstem	3	-21	-22.5	2817	2.98
medial superior cerebellum	-3	-55.5	-18	959	3.22
pyramidal tracts	-31.5	16.5	31.5	940	3.42
frontal medial	12	54	6	140	3.07
occipital lobe	6	-97.5	1.5	759	3.82

5.3.2 Individual Subject SVM Predictions

During a preliminary investigation, using voxels from the whole brain (i.e. no feature selection) and a linear SVM resulted in a classification accuracy of 57% (sensitivity 62%, specificity 53%, $\chi^2 = 1.5$, $p = 0.22$). Using the 'mean-threshold' feature selection method with an identified optimal threshold of 0.061 and a linear SVM, resulted in a marginally increased accuracy of 59% (sensitivity 62%, specificity 56%, $\chi^2 = 2.1$, $p = 0.14$). Feature selection therefore minimally improved the predictive accuracy of the *linear* SVM. However during this investigation it was discovered that applying feature selection in conjunction with a non-linear kernel such as the Gaussian kernel significantly improved classification accuracy.

A Gaussian SVM was used to analyse the 34 structural MRI images of children satisfying DSM IV criteria for ADHD and 34 structural MRI images of control subjects. Feature selection was implemented using a 'mean threshold' procedure which selected voxels (training-data only) which differed between the ADHD and control groups by more than a given threshold. The analysis was done using; the grey matter compartment of T_1 weighted images only, white matter compartment images only, and combined grey and white matter images.

The analysis using white matter images alone resulted in an individual subject predictive accuracy of 93% (sensitivity 1.0, specificity 0.85, $\chi^2 = 50.6$, $p << 0.0001$). The analysis using grey matter images alone resulted in an accuracy of 63% (sensitivity 0.68, specificity 0.59, $\chi^2 = 4.8$, $p < 0.028$), and that with grey and white matter images combined an accuracy of 81% (sensitivity 0.74, specificity 0.88, $\chi^2 = 26.5$, $p << 0.0001$). Consequently, the most accurate predictions were supported by white matter images alone. Predictions using both grey and white matter images did not improve the accuracy of prediction.

5.3.3 Brain Regions identified using Feature Selection

When only white matter images were used for analysis, the largest volume of voxels selected during feature selection which supported the 93% accuracy of prediction were located in the brainstem. As shown in Figure 21 (a) and listed in Table 5, this comprised a large region in the central pons with a small extension to midbrain, and a smaller bilateral region within the midbrain. For illustration, Figure 22 shows the

locations of the locus coeruleus (Keren *et al.,* 2009) and ventral tegmental area nuclei (Guitart-Masip *et al.,* 2012; Mai *et al.,* 1997), in relation to the white matter region used for classification. This white matter abnormality may involve the axonal projections to and from the locus coeruleus and ventral tegmental area. Smaller regions in the bilateral frontal pole white matter deep to Broadmann's Area (BA) 10 and pyramidal tract were also identified, which might be related to prefrontal and motor abnormalities.

When only grey matter images were used for analysis, regions supporting individual prediction at accuracy of 63% were identified in the dopamine rich putamen, bilateral frontal pole grey matter (BA 10) and bilateral inferior parietal lobule. Grey matter regions are shown in Figure 21 (b) and listed in Table 6.

5.3.4 Comparison between VBM analysis and Classification

Brain regions identified using feature selection were compared with the results of a VBM group level analysis ($p<0.005$, whole brain level significance). In the VBM analysis, only white and grey matter volume *reductions* were identified in ADHD subjects. As shown in Figure 23, white matter regions were identified in the brainstem and grey matter regions in the putamen, both of which overlapped with feature selection identified regions, indicating that prediction was based on significant white and grey matter volume reductions in the ADHD subjects.

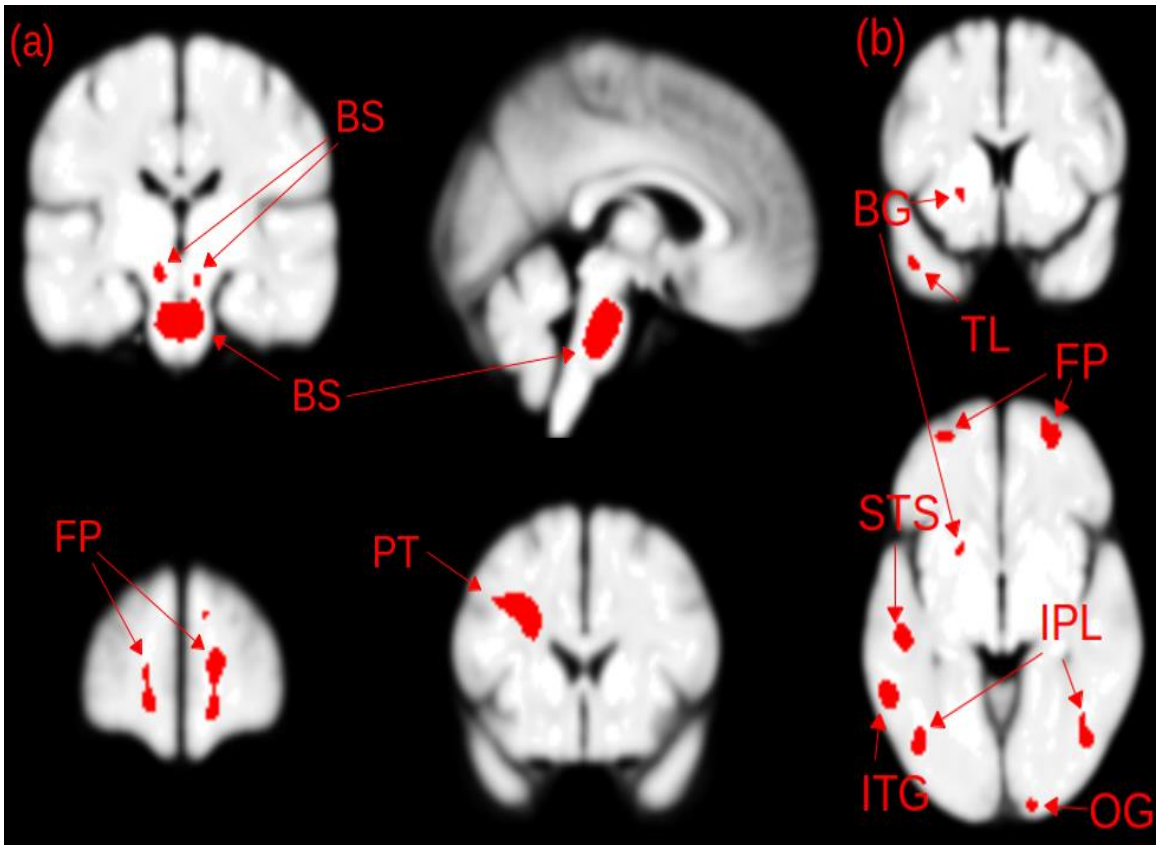

Figure 21: (a) Feature selection (Gaussian SVM) identified brain regions in *white* matter. BS – brainstem regions comprising a lower region in the pons and smaller bilateral region in the mid-brain; FP - frontal pole white matter; PT - pyramidal tracts (b) Feature selection (Gaussian SVM) brain regions identified using *grey* matter. BG – basal ganglia; FP – frontal pole; STS – superior temporal sulcus; IPL – inferior parietal lobule; ITG – inferior temporal gyrus; TL – temporal lobe; OG – occipital gyrus.

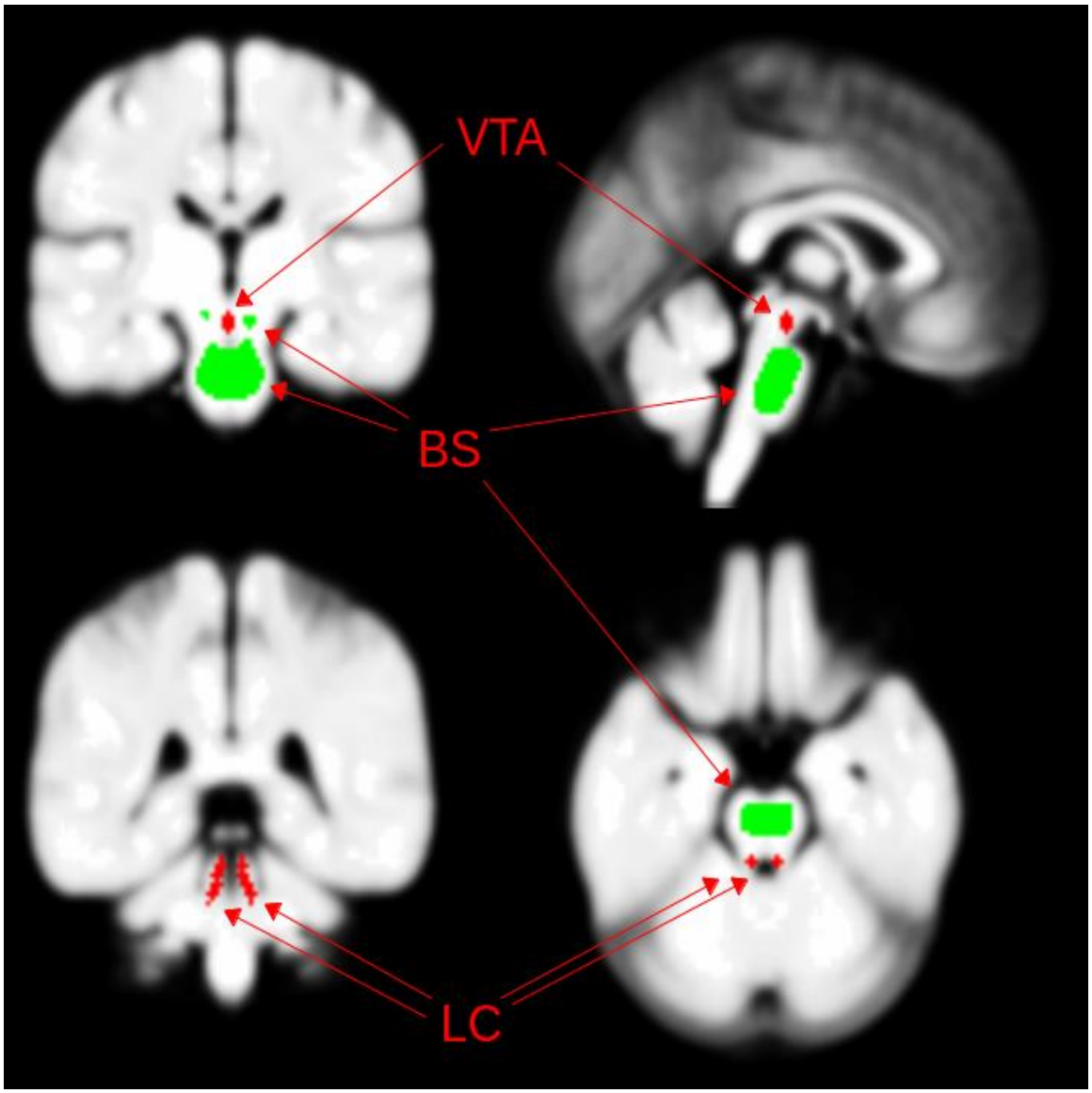

Figure 22: Locations of the noradrenergic locus coeruleus (LC) and dopaminergic ventral tegmental area nuclei (VTA), in relation to the brainstem (BS) white matter region used for classification. LC and VTA locations from previous studies (Guitart-Masip *et al.*, 2012; Keren *et al.*, 2009; Mai *et al.*, 1997).

Table 5: MNI coordinates of each cluster of white matter identified using mean-threshold feature selection with the Gaussian SVM. The number of resampled voxels contained in each cluster was calculated as discussed in Chapter 2.

Region	MNI coordinates			Resampled voxels per cluster
	x	*y*	*z*	
brainstem (pons)	0	-25	-32	1581
bilateral brainstem (midbrain)	-9	-20	-10	47
	9	-21	-13	15
frontal pole white matter	18	52	5	441
	-14	58	-5	166
pyramidal tracts	-26	18	27	571
white matter deep to cingulate gyrus	19	39	23	322
white matter deep to superior temporal gyrus	48	-49	9	267
inferior longitudinal fasciculus	34	-64	18	243
	-27	-60	18	34
white matter deep to cuneus	14	-93	12	193
white matter deep to middle frontal gyrus	34	43	14	127
	36	4	41	70
	38	30	26	117
white matter deep to medial orbitofrontal cortex	-28	40	-2	18
white matter of superior parietal lobule	37	-65	38	40
white matter of inferior parietal lobule	40	-55	46	34
white matter deep to medial frontal cortex	12	53	26	27
white matter deep to lingual gyrus	21	-87	-2	15
corpus callosum	-20	32	8	26

Table 6: MNI coordinates of each cluster of grey matter identified using mean-threshold feature selection with the Gaussian SVM. The number of resampled voxels contained in each cluster was calculated as discussed in Chapter 2.

Region	MNI coordinates			Resampled voxels per cluster
	x	*y*	*z*	
basal ganglia (putamen)	-21	9	-3	30
bilateral frontal pole	26	59	-3	278
	-27	54	0	97
bilateral superior temporal sulcus	50	-38	5	532
	-48	-27	-5	210
bilateral inferior parietal lobule	45	-59	38	336
	-40	-66	38	126
bilateral inferior temporal gyrus	44	-66	-9	284
	-60	-52	-9	258
bilateral inferior frontal sulcus	40	35	14	312
	-39	30	20	301
	-37	41	9	64
middle frontal gyrus	-35	15	30	55
	-36	4	48	70
medial temporal gyrus	-40	-70	18	214
superior parietal lobe	32	-44	43	213
occipital gyrus	17	-100	-6	182
	-42	-72	-6	120
	22	-95	10	95
	33	-80	18	151
superior temporal gyrus	-47	-52	18	100
precentral gyrus	38	-13	42	78
middle temporal gyrus	-47	7	-31	67
inferior cerebellum	-29	-44	-54	54
postcentral gyrus	47	-23	42	36

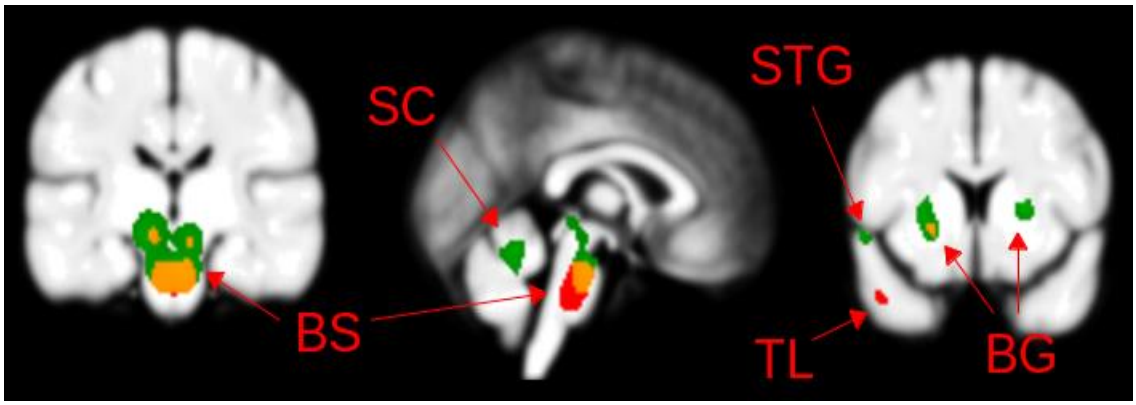

Figure 23: Brain regions identified using feature selection (red), voxel based morphometry (green), and regions common to both analyses (orange). BS – brain stem; SC – superior cerebellum; BG – basal ganglia; TL – temporal lobe; STG – superior temporal gyrus

5.3.5 Multivariate Feature Selection

The investigation into multivariate feature selection techniques was intended to demonstrate that a significant classification accuracy could be achieved using an alternative feature selection method. The combined grey and white matter image classification approach was used because it was hoped that this new feature selection method could improve upon the 81% classification accuracy achieved using the mean-thresholding method. The combined grey and white matter images were entered into the wrapper technique using 11-fold cross-validation as described, the accuracy achieved was 63% (sensitivity 0.68, specificity 0.59, $\chi^2 = 4.80$, $p = 0.0284$). Although significant, this *testing* stage accuracy is lower than the accuracy achieved using the univariate feature selection method and therefore this method was not investigated further.

5.3.6 Group level differences between previously medicated and unmedicated ADHD subjects

To confirm that the classification wasn't influenced by the 10 ADHD subjects that were being medicated around the time of the scan (but with the medication withheld 48 hours prior to scanning) a group level analysis between the 10 medicated subjects and 10 of the medication naive ADHD subjects, matched on the basis of age, IQ and FBB-HKS, was performed.

This analysis identified that there were no grey matter regions which were significantly decreased in the medicated group compared to the unmedicated group. Regions which showed an increase in the medicated group's grey matter volume included BA20 (inferior temporal gyrus (46,-34,-18)), insula (-49, 16, -4) and (-48, -10, 9), inferior frontal gyrus (-51, 34, -6), midbrain (-8, -9, -21), frontal pole (-21, 46, 6) and medial temporal lobe (-30, 12, -30). There was a slight overlap in the grey matter used during the classification and the medicated/unmedicated group level analysis in the frontal pole.

Increases and decreases were identified when comparing the white matter volume in the medicated and unmedicated ADHD groups. The increases (medicated greater than unmedicated) occurred in the white matter deep to the lateral orbitofrontal cortex (39, 34, -16) and the uncinate fasciculus (-40, 4.5, -18) and one

region showed a reduction in the white matter deep to BA7 (parietal lobe (48, -42, 55)). Figure 24 shows the grey matter volume increases and white matter volume reductions and increases between the two groups.

The brain regions which were identified in the white matter medication VBM analysis did not overlap any regions identified in the white matter prediction, increasing confidence that the classification is not directly influenced by the differences between medicated and unmedicated children. There was a small overlap between the grey matter regions but as the overlap was minimal (only included three voxels), it is likely that this is a result of analysing and classifying smoothed images. Figure 25 shows the small overlap between the regions used in the grey matter prediction and the regions which had significantly increased grey matter in the medicated group.

Use of stimulant medication has been reported to be associated with normalisation of grey matter volume reduction, which would make predictive classification more difficult, although this effect has yet to be confirmed in longitudinal studies (Frodl and Skokauskas, 2012; Nakao *et al.*, 2011).

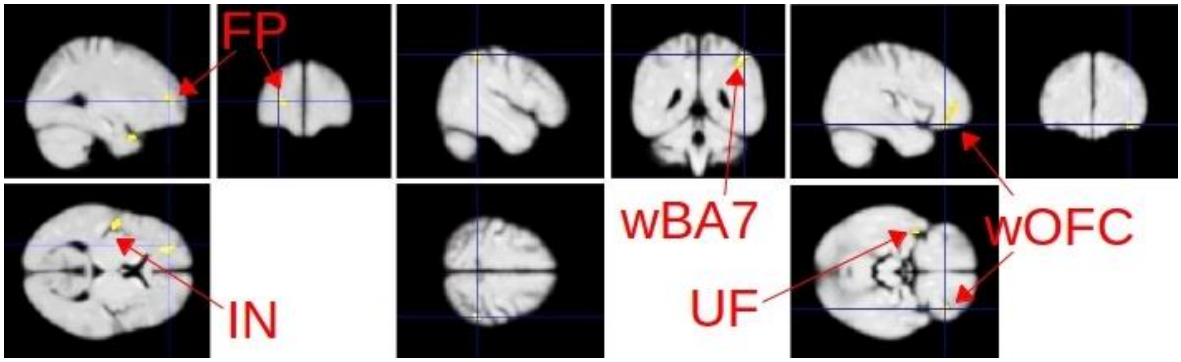

Figure 24: Group Level (VBM) significantly ($p<0.005$) increased grey matter volume (left), decreased white matter volume (centre) and increased white matter volume (right) in medicated versus unmedicated ADHD patients. Medicated patients with ADHD had significantly increased grey matter volume in BA20 (inferior temporal gyrus), insula, inferior frontal gyrus, midbrain, frontal pole and medial temporal lobe compared to unmedicated ADHD patients. A white matter volume decrease was identified in the white matter deep to BA7 (parietal lobe) and increased white matter volume was identified in the white matter deep to the lateral orbitofrontal cortex and the uncinate fasciculus.

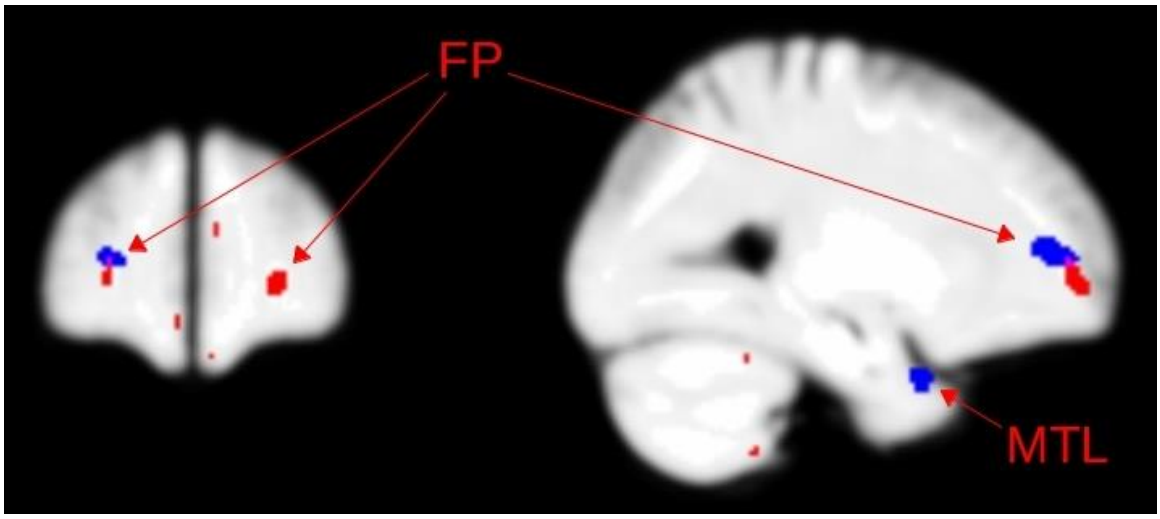

Figure 25: The overlap between the regions used in the grey matter prediction and the regions which had significantly increased grey matter volume in the medicated group. Blue: significant regions identified in the group level analysis (Figure 24) but not used in the grey matter classification (Figure 21 (b)). Red: regions used in the Gaussian SVM (Figure 21 (b)) but not identified as significant in the group level analysis (Figure 24). Purple: the region (frontal pole -FP) which was both identified in the group level analysis (Figure 24) and used in the grey matter prediction (Figure 21 (b)).

5.4 Discussion

Using a Gaussian SVM in conjunction with anatomical feature selection, it was possible to classify individual scans from young males with ADHD and healthy controls, using white matter compartment of T_1 weighted scans, with 93% accuracy. This is the highest accuracy reported for a predictive classification study in ADHD using 'structural' (T_1 weighted) brain scans alone; most have used resting state fMRI though a few have combined resting state fMRI with T_1 data for prediction, and none have used white matter images.

The only other attempts to predict diagnostic status between children and adolescents with ADHD and healthy controls using only structural MR images achieved classification accuracies of 67% (Chang *et al.,* 2012) and 79% (Lim *et al.,* 2013), although two other studies combined resting state fMRI and structural MRI to the same classification problem, achieving accuracies of 58% (Eloyan *et al.,* 2012) and 76% (Cheng *et al.,* 2012). The accuracy of the grey matter prediction in this study was a comparable figure, 63%, however, it is important to note that the 93% predictive accuracy was obtained using the white matter component of the structural images – images that none of the above studies investigated.

The largest cluster of voxels which were relevant to the prediction which achieved 93% accuracy was in the brainstem. As mentioned previously, this region is adjacent to noradrenergic and some dopaminergic nuclei. The noradrenergic locus coeruleus nuclei lie in the posterior brainstem lateral to the periaqueductal grey matter and the dopamine ventral tegmental area (ellipsoid structures with an axis in the direction between the midbrain and pons) is located anteriorly in the ventral-medial aspect of the brainstem (Afshar *et al.,* 1978; Mai *et al.,* 1997). The white matter region could therefore contain axonal connections between the locus coeruleus / ventral tegmental area nuclei and rest of the brain, raising the possibility of 'catecholamine dysconnection' (abnormality in connection (Stephan *et al.,* 2009), in contrast to 'disconnection') contributing to the ADHD syndrome. If such a dysconnection exists in ADHD, this could provide a plausible explanation why medications which enhance dopamine and noradrenaline function are able to reduce associated behavioural abnormalities.

Several reviews of the neural substrates of attention relevant to ADHD, have implicated a distributed network of regions including the brainstem Reticular

Activating System (RAS, which includes the locus coeruleus and ventral tegmental area nuclei), ascending white matter pathways from the RAS (mediating arousal), and descending pathways from the prefrontal cortex via the thalamus to the RAS (mediating inhibition), and basal ganglia / frontal lobe abnormalities (e.g. Riccio *et al.* (2002) and Voeller (1991)). It has been argued that disruption at any level of this system could lead to a behavioural phenotype that resembles ADHD (inattention, difficulty concentrating, distractibility, impulsivity, hyperactivity) (Riccio *et al.,* 2002; Voeller, 1991). A number of the ADHD and control subjects who contributed T_1 weighted images to the present study also took part in an fMRI study of attention (Konrad *et al.,* 2006). During the alerting component of an attention task, ADHD subjects showed abnormally increased activation at the midbrain-pons junction at a posterior brainstem region, which the authors suggested was the locus coeruleus (Konrad *et al.,* 2006). Abnormal functional activity of the locus coeruleus could be linked to decreased white matter connections with the rest of the brain.

In addition, Lim *et al.* (2013) reported that the brainstem was a relevant feature during classification of grey matter images which yielded 79% accuracy – although this region appears to be more posterior when compared with the white matter region identified in this study. The brain regions which were most relevant to this grey matter-based prediction included the caudate, ventral striatum/putamen, insula, brainstem, thalamus, hypothalamus, precuneus/cuneus, hippocampus, amygdala, cerebellar vermis and inferior and superior parietal regions (Lim *et al.,* 2013). A number of these regions were also identified in the current study.

The frequently reported decrease in cerebellar grey matter volume (Berquin *et al.,* 1998; Bussing *et al.,* 2002; Castellanos *et al.,* 1996; Hill *et al.,* 2003; Lim *et al.,* 2013; Mostofsky *et al.,* 1998) was replicated in this study. Additionally, this study identified a decrease in cerebellar white matter volume. The decrease in grey matter volume identified in the temporal lobes has been reported by Castellanos *et al.* (1996). Whilst the amygdala has previously been reported to show no significant changes in grey matter (Castellanos *et al.,* 1996; Filipek *et al.,* 1997), it has also been reported to show a decrease in volume (Lim *et al.,* 2013; Plessen *et al.,* 2006) as found here.

Reduced grey matter and white matter deep to BA 10 were identified in the relative classifications. BA 10 functions are diverse, including episodic memory retrieval and ‘multitask’ information processing, with evidence for lateral-medial and

rostral-caudal functional gradients, implying BA 10 is not a functionally homogeneous region (Gilbert *et al.,* 2006). The frontal pole cortex has been found to develop late into childhood/adolescence which may increase susceptibility to developmental disorders (Tsujimoto *et al.,* 2011). Partial disruption of BA 10 and its connections could therefore have widespread effects on cognition. A *decrease* in ADHD grey matter volume in the basal ganglia (putamen) was identified compared to controls which concords with previous studies (Ellison-Wright *et al.,* 2008; Frodl and Skokauskas, 2012; Nakao *et al.,* 2011) and further suggests dopamine dysfunction in ADHD. In addition, lesions of the basal ganglia of experimental animals can result in behavioural change reminiscent of some aspects of the ADHD syndrome (Alexander *et al.,* 1986).

Group level abnormalities in white matter are rarely investigated (Hermann *et al.,* 2007). Whilst no differences in white matter volume were reported in one study (Carmona *et al.,* 2005), another, larger study, described significantly reduced total white matter volume and significant reductions in the frontal, parietal, temporal and occipital lobes in ADHD (Castellanos *et al.,* 2002). It is important to note that the decreased region in the brainstem, identified using both the VBM analysis and the classification algorithm, requires further investigation using DTI. When using the normalisation technique, as described, it is unclear whether the brainstem reduction results from a reduced volume or reduced white matter integrity in children and adolescents with ADHD. Reduced brainstem volume has been reported previously in subjects with both ADHD and epilepsy compared with both a healthy control group and subjects with epilepsy alone (Hermann *et al.,* 2007).

An important aspect of the VBM group level analysis in the present study and analyses reported in the literature is the use of DARTEL. This means a study-specific template, created during the DARTEL process, is used to realign and warp the images to a standardised anatomical space. This is especially important here as the brains of children and adolescents are quite different from those of the adults that were used to create the default SPM templates. Importantly, in comparison to the traditional SPM VBM method, DARTEL has also an improved method for warping the MRI images towards the aforementioned study-specific template, resulting in more accurately aligned images across subjects (Ashburner, 2007), decreasing inter-subject variance and therefore increasing the power of subsequent statistical analyses. As a DARTEL-created template has been shown to perform as well as a

specialist atlas of the cerebellum and brainstem (SUIT - http://www.icn.ucl.ac.uk/motorcontrol/imaging/suit.htm)(D'Agata *et al.,* 2011) it is considered to perform accurate normalisation in these regions. The SUIT atlas has since been updated to use the DARTEL approach "for more accurate results". Whilst several of the abnormal regions identified have been reported previously, some have not, and it is possible this is in part a consequence of using DARTEL.

A potential limitation to the present study is stopping medication two days or more before scanning in a minority (29%) of the ADHD subjects. Medication might be associated with structural brain change. To investigate whether this impacted on the results, ten previously medicated ADHD subjects were matched with ten of the medication naive ADHD subjects, on the basis of average age, IQ and FBB-HKS scores. White matter differences in previously medicated ADHD subjects were identified. However, none of the medication related regions overlapped with the white matter regions used for predictive classification and none were found in the brainstem.

Another possible limitation is movement during the image acquisition. It's unclear if movement would make classification more accurate or less accurate, but as described in section 5.2.3, a range of methods were used to exclude significant movement effects. It is important to emphasise that the high classification accuracy achieved here is only relevant for scans obtained from the same MR scanner. If images from a MRI scanner were classified using an algorithm developed using images from a different scanner, it is unlikely that the scans would be classified to the level of accuracy reported here. This is due to subtle differences in images obtained from different scanners (Moorhead *et al.,* 2009). Work on possible scanner related confounds to prediction is required. Whilst the reported method achieved high predictive accuracy, it is important to note that this was in the context only of scans from volunteers with ADHD and controls. Further work would be required to establish the accuracy of the technique if scans from other diagnostic categories were included, and scans from subjects with comorbidities. Whilst the ADHD dataset includes subjects with comorbid disorders, there is insufficient data to conduct sub-group analyses.

In summary, it was possible to predictively classify scans from individual children and adolescents with ADHD to an accuracy of 93% using the white matter compartment of T_1 weighted images alone. This is of comparable diagnostic

accuracy to that reported for general adult psychiatric syndromes (Klöppel *et al.*, 2008; Koutsouleris *et al.*, 2011; Mwangi *et al.*, 2012a). The grey matter-based prediction achieved a lower, yet comparable, accuracy to similar studies in the literature (Chang *et al.*, 2012; Lim *et al.*, 2013). In addition, a number of *group* level reports of grey and white matter abnormalities in young males with ADHD were replicated. However, structural brain abnormalities which had not been reported previously were also identified, perhaps due to the use of DARTEL and in particular a large region of reduced white matter volume in the brainstem.

Given the possible heterogeneity of the ADHD syndrome (Fair *et al.*, 2012a), the results are encouraging for the identification of consistent imaging biomarkers, that can inform future work into the aetiology, pathophysiology and clinical management of ADHD. To the author's knowledge, brainstem white matter volume has not been specifically investigated in previous studies of ADHD. The brainstem region identified here may constitute a biomarker for ADHD, although independent studies are required to replicate these findings, investigate the nature of the white matter abnormality using DTI, explore issues of diagnostic syndrome specificity and possible scanner related confounds to prediction.

Chapter 6: The iBOCA study

6.1 Introduction

The original goal of this PhD was to investigate whether it would be possible to apply machine learning methods to the iBOCA (Imaging Brains of Children and Adolescents) study. iBOCA is an ongoing study which involves scanning medication-naïve children and adolescents who have been diagnosed with ADHD, healthy sibling controls and volunteers with no family history of ADHD. After the scan, the children and adolescents with ADHD began a trial of MPH.

The iBOCA study was designed to be analysed with three different aims. The first was to replicate the accurate diagnostic classification between patients and controls (discussed in Chapter 5). If the work was able to be replicated it could increase confidence in the robustness of the method and results.

The second and main aim was to attempt to classify between responders and non-responders to MPH. This has the obvious advantages of potentially providing a reliable predictor of treatment response prior to exposure of the medication. Furthermore, it has the potential to increase the understanding of the mechanisms which underlie MPH response.

As discussed in Chapter 3, prediction of treatment response is a developing research area with a small number of studies reported in the literature. Chapter 4, which described the prediction of medication response using sociodemographic, clinical and neuropsychological measures, demonstrated that it is possible to predict MPH response in children and adolescents with ADHD. However, it is hoped that structural and functional MRI data, in addition to sociodemographic measures and clinical scores, could improve prediction accuracy further and feature selection techniques could reveal potential biomarkers of MPH response.

Finally, the third aim of the study was to test the dopamine transfer deficit (DTD) theory of altered reinforcement mechanisms in ADHD as suggested by Tripp and Wickens (2008). This is explained in more detail in section 6.8.

Structural MRI was the primary imaging modality that was intended to perform the classification analyses for several reasons. First, as structural MRI had a shorter duration than the fMRI scans it was less likely to suffer from motion artefact. Second, structural MRI has been reported to allow accurate predictions of diagnosis,

as described in Chapter 5 and in the literature. However, if a high enough quality fMRI dataset was obtained, it was planned that the classification analyses could be repeated using various contrast images, or a combination of fMRI contrasts and structural images.

6.2 Recruitment Criteria

Subjects were recruited through participation in an existing EU FP7-funded pharmocovigilance study (ADDUCE) at Ninewells Hospital and Medical School in Dundee, UK. Informed consent was obtained from all volunteers and their parents. The study protocols were approved by the local Ethics Committee.

Initial diagnosis was made by experienced clinicians according to the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) (American Psychiatric Association, 1994) criteria using the KIDDIE SADS semi-structured clinical interview. The inclusion criteria included children and adolescents aged between 10 and 17 years and an IQ>70 (as assessed by the BPVS).

Exclusion criteria included potentially confounding diagnoses – any other psychiatric disorder, including autism spectrum disorder, schizophrenia, bipolar disorder, depression, Tourette's or major neurological disorder. All subjects were required to be medication-naïve.

The primary outcome measure is clinical response, measured using the ADHD rating scale with responder status defined according to the methods of Jacobson and Truax (Jacobson and Truax, 1991) to define whether there has been clinically significant change and clinically meaningful response after six months of MPH treatment.

Typically developing controls were recruited from healthy siblings of children with ADHD and volunteers with no family history of ADHD. All controls underwent psychiatric screening using the same interviews as the patient volunteers. None of the controls had a history of current or past psychiatric or neurological disorder and none were taking medication. Volunteers were compensated for participation in the study.

6.3 fMRI paradigm

An event-related fMRI instrumental learning task involving reward processing was performed. The paradigm used was a modified version of the Pessiglione task (2006). This modified task has been previously applied in studies into MDD (discussed in Chapter 8) and drug addiction (Gradin *et al.,* 2013).

The aim of the task was that participants were required to attempt to win as many vouchers as they could whilst avoiding losing as many vouchers as possible. The volunteers were informed that they would be given a gift voucher with an amount dependent on how well they performed the task during the scan.

The paradigm incorporated rewarding, neutral and aversive events into one task. One pair of novel fractal images was associated with each stimulus type (e.g. see Figure 26 where the reward pair is represented by a pair of square shaped images, the neutral – a pair of circular shaped images and loss – a pair of triangular shaped images). To remove the possibility that the fractal shapes were introducing a bias, the associations between the events and the pairs of fractal shapes were randomised across subjects.

During reward trials there are two possible outcomes – either winning a voucher or there is no change in the number of vouchers. Conversely, the loss trials had two possible outcomes – either losing a voucher or there was no change in the number of vouchers. A neutral condition was also included whereby the number of vouchers would not be altered irrespective of the volunteer's choice. The possible outcomes from each stimulus type are outlined in Figure 27.

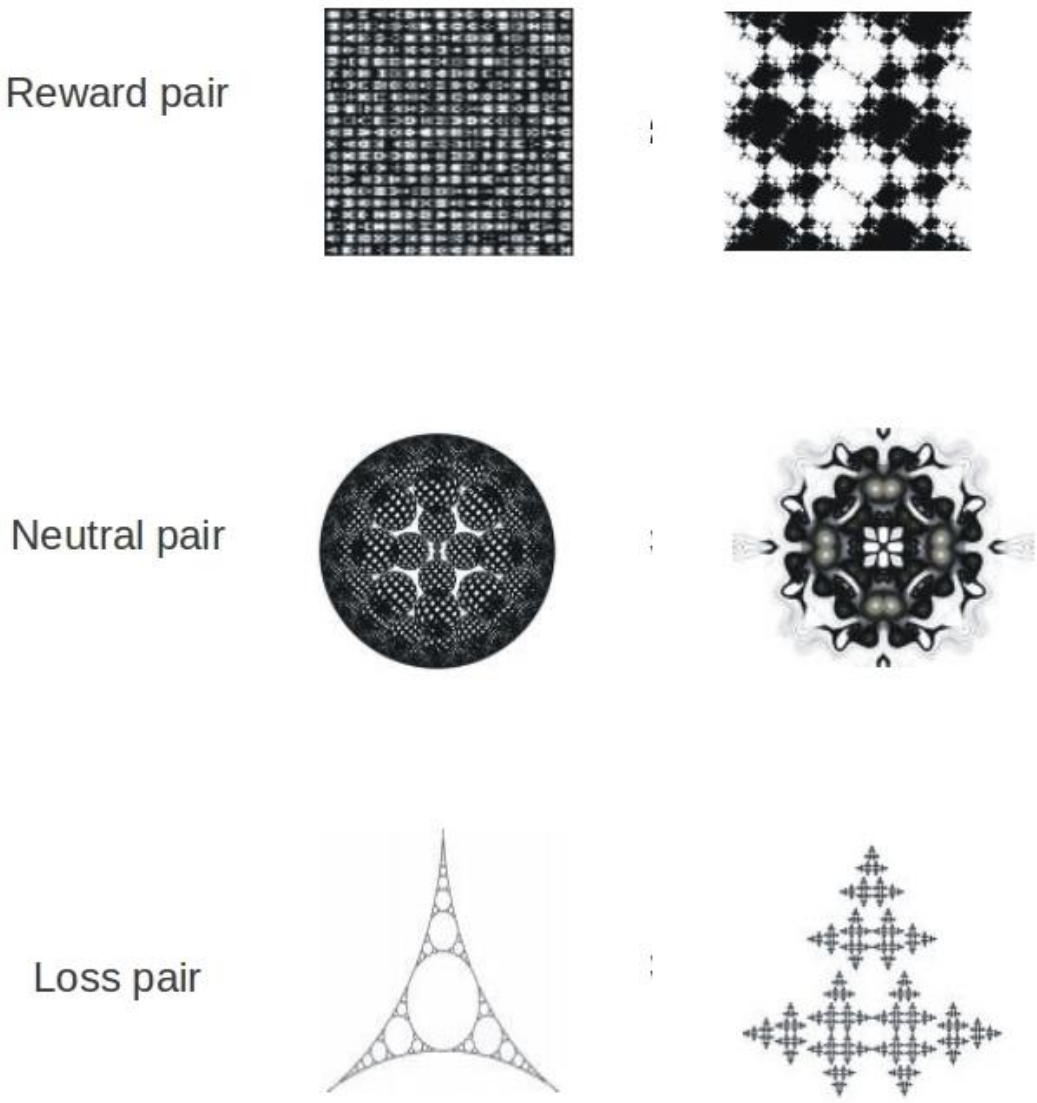

Figure 26: The stimuli displayed during the modified version of the Pessiglione task. In this example the reward pair is represented by square-shaped fractal images, neutral by circular-shaped fractal images and the loss pair by triangular shaped fractal images but the shapes are randomly assigned to the 3 stimulus types.

Reward pair possible outcomes	You Win Voucher	Nothing
Neutral pair possible outcomes	No change Voucher	Nothing
Loss pair possible outcomes	You Lost Voucher	Nothing

Figure 27: The two possible outcomes for each stimulus type. The subjects were informed during task training that there was no difference between “No change in vouchers” and “Nothing”.

The full paradigm contained 240 trials in total (80 of each stimulus type). To reduce scanning time and increase tolerability for the children, the total number of trials was reduced to 180 trials (60 of each stimulus type). The task was separated into 3 scanning sessions of 60 trials (20 of each stimulus type) to allow volunteers to have a short break. Volunteers were informed during training that there would be no changes to the task between scanning sessions. Each session took approximately 15 minutes to complete.

At the beginning of each trial one pair of fractal images was presented, with the order of the fractal images being randomly assigned to the left or right of the screen. This was to ensure any potential bias towards a favoured hand or visual field did not have an effect on the overall results. The volunteer was required to choose between the left or right image using triggers in their hands. Once the choice was made, a red circle appeared around the selected fractal image. Subjects were instructed that there were no differences between images appearing on the left or the right and that the task is based on trial and error – therefore there is no way to get the favourable outcome on every trial.

After three seconds from the beginning of the trial the images were replaced with a fixation cross, a small black "+" in the centre of a white background. If the subjects did not respond within the first three seconds then no feedback was given, and the fixation cross remained in place for the following four seconds of the trial. However, it was made clear to the participants that they should always make a choice (to avoid tactical play whereby non-response during the lose trials would result in a good outcome). In trials where subjects responded within the time limit, the fixation cross was shown for three seconds and the corresponding feedback was given for the final second of the trial.

The scan was optimised for detection of the neural signals of interest for event-related fMRI, such that the inter-trial interval varied over the course of the paradigm using a program called optseq2 (http://surfer.nmr.mgh.harvard.edu/optseq/). During the inter-trial interval a fixation cross was displayed for a variable amount time, ranging between 3 and 13.75 seconds. Figure 28 displays an example trial which illustrates the sequence of images and the task timing.

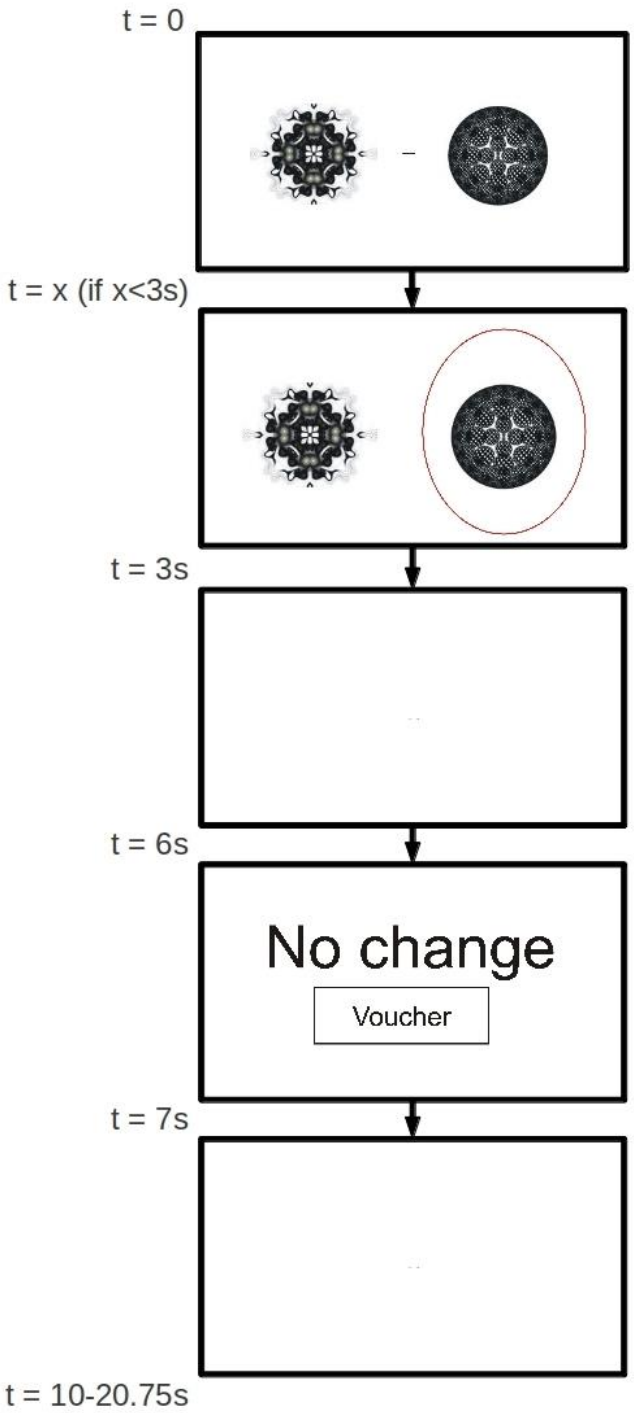

Figure 28: An example trial to illustrate the sequence of images and the task timing. In this example the circular fractal pair is the neutral pair (as in Figure 26). This illustration displays the example when an image was selected within the time limit of 3 seconds (precisely after 'x' seconds in this example). As this was a neutral pair there were two possible outcomes: "No change in vouchers" and "Nothing" (which are essentially the same thing – see Figure 27 for the possible outcomes for each trial type). Each trial took between 10-20.75 seconds depending on the inter-trial interval.

The participants were not informed before undertaking the task that each pair of images had one image which had a high probability (0.7) of giving a favourable reward (winning a voucher or avoiding losing one) and a lower probability (0.3) of delivering an unfavourable reward (not winning a voucher or losing one) whilst the other image had the reversed outcome probabilities. The Pessiglione task, on which this paradigm was based, used a 0.8/0.2 ratio (Pessiglione *et al.*, 2006) but it was decided that decreasing this to 0.7/0.3 would ensure that the subjects would not be able to identify the pattern too early in the task. To successfully complete the task, the subjects had to learn, by trial and error, which images were most likely to give the favourable outcome.

6.4 Personal input towards the fMRI paradigm

The paradigm used in the previous studies was written and optimised for event-related fMRI in Presentation (http://www.neurobs.com/menu_presentation/menu_features/features_overview) by a former Research Assistant and Professor Steele. It is necessary to include details of the task in this thesis: a) to describe the basis for testing the DTD theory of Tripp and Wickens, and b) the author re-implemented the task using Psychtoolbox (http://psychtoolbox.org/HomePage) which runs in Matlab (http://www.mathworks.co.uk/products/matlab/).

Re-implementing the paradigm required understanding the processes used in the original 'Presentation' coded version of the task, identifying relevant Psychtoolbox functions and then translating each component of the task into Psychtoolbox format. A particular issue was the recording of the fMRI scanner pulses at the same time as the behavioural data. This is an essential component of all fMRI paradigm data recordings as the task is required to start exactly after the 'discard acquisitions'. In Dundee, this corresponds to the fifth scanner pulse (the first four brain images acquired are discarded due to scanner transients). This issue was resolved by writing code which could receive the two different data streams simultaneously via different ports (the behavioural information was received through a USB port and the scanner pulse timings were received through a serial port).

A major benefit of the rewritten paradigm was that the overall paradigm timing was better than the old 'Presentation' version. This paradigm timing

improvement is easily seen when each version of the paradigm were compared by the time taken to complete the same task with the same responses for one session. By the end of one session, the Presentation code overran the intended timing of the task by six seconds. In comparison, the Psychtoolbox code remained within 0.39 seconds of the correct full session timing. This was possible due to a custom "timing correction function" which was created to dynamically (as the task was running) adjust for any time delays. As can be seen in Figure 29, at all time points, the Presentation code took longer than intended whereas the Psychtoolbox based code varied between taking more and less time for each picture presentation (including the fixation cross image). This led to the mean deviation from the requested timing for each picture appearance in the Presentation code being 0.019 seconds in comparison with 0.001 seconds in the Psychtoolbox code.

This improved overall timing came at a slight cost, however, as the maximum range of timing deviations over the whole session was larger in the Psychtoolbox code than in the Presentation code (0.70 and 0.03 seconds respectively). This larger range is due to a small number of larger corrections being required during the "timing correction function" in order to "catch up" with the correct timing. This is not a significant concern though because it can be argued that having a larger range of timing deviations and more accurate timing overall is more important than the converse situation.

In addition to writing the paradigm, the author was also able to resolve a problem in the scanner which was identified in the Psychtoolbox code but not in the old Presentation code. The issue which occurred was that there were two types of pulses being received in the logfiles, the correct ones from the scanner and some "phantom pulses" which occurred randomly throughout a number of fMRI sessions. As the Psychtoolbox code identified this randomly occurring "phantom pulse", it prompted an investigation to identify the potential source of this issue. It was discovered that the issue came from a connection between the hardware which receives the behavioural data and the hardware which receives the scanner pulses. This connection is required for another experimental fMRI set-up but it is unnecessary for the configuration in this study. Once this cable was disconnected the issue was resolved. In order to prevent the same issue from re-occurring, custom code to check for such phantom pulses was written.

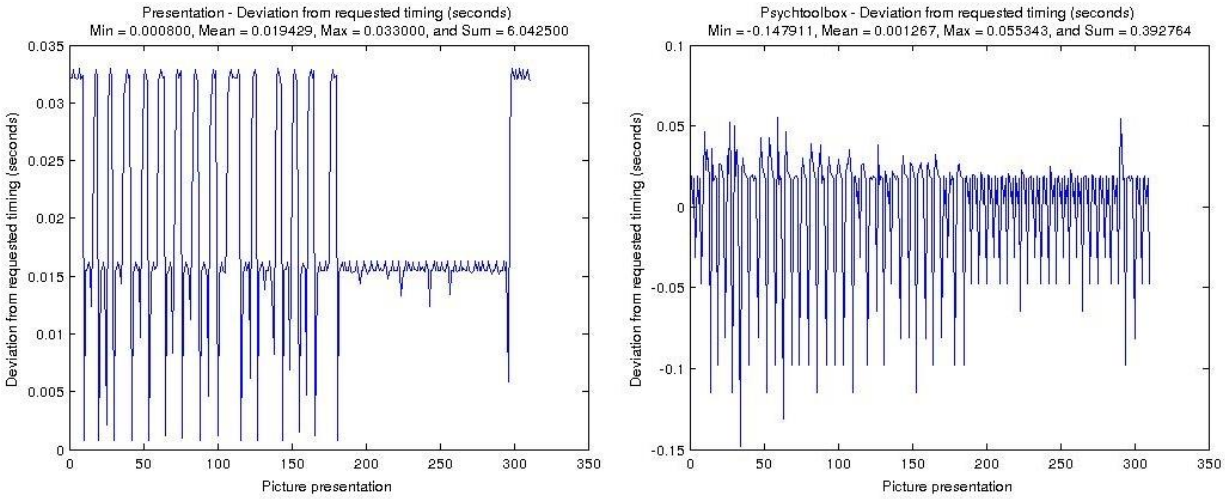

Figure 29: Comparison between the timing of the Presentation code (left) and the Psychtoolbox code (right). The Psychtoolbox code (written by the author) shows an improved overall timing over the same session with identical responses.

6.5 Review of the Practicalities of Scanning Children and Adolescents

As this was the first child and adolescent neuroimaging study this research group had performed, the practicalities of scanning children and adolescents, particularly studies into childhood ADHD, were investigated and reviewed in the literature before the study began. Typically it takes about 7 min to obtain a high spatial resolution T_1 weighted 'structural' Magnetic Resonance scan. From a subject's perspective, the main requirement for the acquisition of high quality T_1 data is to remain still during the scan, to avoid motion artefacts. In contrast, paradigm based fMRI is much more demanding for participants. Task-based fMRI scans require both comprehension of, and cooperation with, a paradigm whilst still ensuring head motion is minimised. Paradigm based fMRI studies of children and adolescents can therefore provide an indication of practical limits of the study design for this thesis. Findings from the fMRI results such as: success rates in completing the planned scans to an acceptable quality, typical total number of patients scanned in a study, length of time each subject was in the scanner, and the effect of medication on reducing motion artefact in the case of ADHD can provide useful information for potential success rates, study sizes, scan time tolerance rates and the effect of medication on scan success in future studies.

Most children have difficulty in remaining motionless for prolonged periods and it would be reasonable to assume that this represents a greater challenge for children with disorders such as ADHD, where over activity is by definition, common (Banaschewski *et al.*, 2010). During fMRI scanning, small movements of more than a single voxel of the head can render scan data uninterpretable by introducing movement artefact. It is therefore relevant to consider to what extent children are able to remain motionless and therefore, typically how long they would likely be able to tolerate scanning. It is also important to determine whether children can maintain task performance without head movement. Within fMRI, a 'successful scan' has been defined as "*completion of the fMRI run with acceptable head motion and adequate task performance*" where the criterion for "acceptable head motion" is less than one voxel of movement in any direction during a run (Yerys *et al.*, 2009).

In a study by Yerys *et al.* (2009), the success rate of completing single fMRI sessions across all subjects, all sessions for all subjects, and at least one session was investigated for children and young people with ADHD, ASD and epilepsy as well as

matched, typically developing controls. The authors concluded that, in comparison to controls, there was a significantly lower success rate for completing a single fMRI session in ADHD children. Specifically, the mean success rate for completing a single session for ADHD children (both on and off MPH) was 78%, whereas the mean success rate for age, IQ and gender matched controls was 96%. The lowest success rate for completing a single session was with children and adolescents with ASD where a 70% successful scan rate was achieved. Yerys *et al.* (2009) reported that, on average, only 50% of all ADHD patients (both on and off MPH) were able to complete an entire fMRI study, compared to an 88% completion rate for controls. They also reported that 95% of both the medicated and unmedicated ADHD subjects successfully completed at least one session in an fMRI battery. Furthermore, the percentage of ADHD children who successfully completed at least one session in an fMRI battery was higher than both the epilepsy group (93%) and the ASD group (81%).

The main conclusion drawn by the authors was that in scanning children and adolescents with disorders such as epilepsy, ADHD and ASD, an extra 20-30% more patients should be recruited into studies, to compensate for anticipated failure to acquire successful scans. The authors also recommended recruiting an additional 10-20% healthy control participants (Yerys *et al.*, 2009). For more information relating to these important findings the reader is referred to Yerys *et al.* (2009).

There is, however, some evidence that an additional 20-30% of patients might not always be sufficient to compensate for failed scans. Durston *et al.* (2003) had to exclude 50% (7/14) of ADHD patients' fMRI data due to excessive head motion artefacts. Whilst such a low scanning success rate does not appear to be common in the literature, the exclusion of patient data due to artefacts is often not clearly discussed within study reports. Of those studies that did discuss the exclusion of data, the next largest percentage of data which had to be excluded (due to excessive head motion) were from three studies with exclusion rates of ~30% (Wang *et al.* (2009) - 34.5% (10/29), Fassbender *et al.* (2009) - 29.4% (5/17), and Zhu *et al.* (2008) - 25% (3/12)).

A particularly important factor affecting scan success rate is the age of the subjects. Indeed, Yerys *et al.* (2009) suggested that effect of age may in fact be as significant as the disorder. For example, ADHD patients aged 7-9 years completed a single session only ~70% of the time (both on and off MPH) whereas those aged 10-

12 years achieved a success rate of ~82% (both on and off MPH). The success rate of these age groups completing an entire fMRI study was also consistent with this. On average, 43% of 7-9 year olds with ADHD (both on and off MPH) completed an entire fMRI study and in comparison an average of 54% of 10-12 year olds with ADHD patients (both on and off MPH) completed an entire study. The association with increased fMRI scanning success rate with age was also found with the other clinical groups (epilepsy and ASD) and the healthy controls.

With respect to medication effects, Yerys *et al.* (2009) reported no significant difference in fMRI task success rate between medicated and unmedicated subjects with ADHD. There was an equal (50%) success rate in completing an entire fMRI study and the success rate of single runs was 79% and 77% for medicated and unmedicated ADHD patients respectively, with 48% of ADHD patients failing at least one session while on or off MPH.

Excessive head motion is the most common cause of failure of fMRI scanning. Yerys *et al.* (2009) reported that medicated ADHD patients had the lowest percentage of failed runs due to excessive head motion, including when compared to healthy controls. Unfortunately, the reason why the medicated ADHD patients' overall success rate was equal to the unmedicated ADHD patients' overall success rate is not clear, as medicated ADHD patients failed a larger number of sessions classified by the authors as due to "*other*" reasons.

When reviewing the literature of fMRI studies in ADHD it was discovered that 16.4 children and adolescents with ADHD and 15.3 controls (including only those used in the analysis) were included, on average, in the thirty-two studies investigated (a full list of the ADHD subjects scanned in the thirty-two fMRI studies is shown in Table 7). The largest number of ADHD patients scanned in a study was 52 (Yerys *et al.*, 2009), the smallest number of ADHD patients scanned (not including patients that were scanned but later excluded from the study) was 7 (Durston *et al.*, 2003). Desmond and Glover (2002) suggest that including 12 subjects in an fMRI scanning study would be adequate to find voxel differences at a low significance threshold of $p<0.05$, however this depends entirely on the effect size of interest (smaller numbers of subjects result in a lower power to detect differences that are actually present, therefore an increased risk of type II errors). They also suggest the number of subjects must be doubled if a higher level of significance is required. As the average number of ADHD patients and controls in the thirty-two

studies reviewed was 31.7 subjects then it is clear that some studies may have required more subjects to obtain higher statistical power (Cohen, 1977). Consequently studies that have a low number of subjects and find negative results are of less interest as a null result could be due to a lack of statistical power. However, a study that rejects the null hypothesis, regardless of the number of subjects, is of more interest, as the result is significantly different from what would have been expected by chance.

Statistical power can be difficult to estimate for neuroimaging studies as it varies from brain region to brain region; some brain regions (e.g. medial orbitofrontal cortex, inferior temporal lobes) adjacent to air filled spaces in the head (e.g. nasal sinuses, ear canals) are additionally affected by signal dropout (the 'susceptibility' artefact), and statistical power is further affected by other factors such as poor image quality. For *individual* subject MVPA studies it is generally thought that even more subjects are required to achieve high accuracy and generalisability of predictions. In all fMRI studies discussed in this brief review, the total duration for an fMRI study was restricted to 30 minutes or less.

In summary, when scanning children and adolescents, the time taken for scanning should be kept to less than 45 min and short breaks between sessions may help to minimise head movement and maintain task performance. This is consistent with Ernst and colleagues who suggest that studies should last ≤ 30 minutes for children younger than 8 years old but that children up to 12 years old may manage a scan of around 45 minutes (Ernst *et al.*, 2003). Ernst *et al.* also recommend that children should receive training on how and when to remain still (Ernst *et al.*, 2003). The literature therefore recommends that the planned number of subjects recruited for studies should be increased by 20-30% to compensate for expected failed scans (Yerys *et al.*, 2009) if scanning children older than 12 years, or increased by ~50% if predominately scanning children in the 7-12 age range.

Table 7: The number of ADHD subjects in a selection of fMRI studies.

fMRI Study	Number of ADHD patients included in the analysis	Number of controls included in the analysis
Adler *et al.* (2005)	11 (bipolar + ADHD)	11 (bipolar only)
Anderson *et al.* (2002)	10	6
Booth *et al.* (2005)	12	12
Brotman *et al.* (2010)	18	37
Cao *et al.* (2008)	12	13
Cao *et al.* (2009)	19	23
Durston *et al.* (2003)	7	7
Durston *et al.* (2006)	11	11
Durston *et al.* (2007)	22	22
Epstein *et al.* (2007b)	20	9
Epstein *et al.* (2009)	10	14
Fassbender *et al.* (2009)	12	13
Hoekzema *et al.* (2010)	19	0
Kobel *et al.* (2009)	14	12
Konrad *et al.* (2006)	16	16
Mostofsky *et al.* (2006)	11	11
Passarotti *et al.* (2010)	11	15
Peterson *et al.* (2009)	16	20
Pliszka *et al.* (2006)	17	15
Rubia *et al.* (2009b)	20	20
Rubia *et al.* (2009a)	13	13
Rubia *et al.* (2010a)	20	20
Rubia *et al.* (2010b)	14	20
Shafritz *et al.* (2004)	15	14
Solanto *et al.* (2009)	20	0 (comparing ADHD subtypes)
Suskauer *et al.* (2008)	25	25
Vaidya *et al.* (2005)	10	10
Van ’t Ent *et al.* (2009)	27	27
Wang *et al.* (2009)	19	20

Yerys *et al.* (2009)	52	32 matched with the ADHD group (137 total)
Zang *et al.* (2007)	13	12
Zhu *et al.* (2008)	9	11

6.6 Practical Aspects of Preparing Volunteers for Scanning

In order to prepare the volunteers for the scan, they required training to be able to play the task. The same fMRI paradigm was used for the task training but with different stimuli used, so children had not learned about the scanner stimuli contingencies before scanning. This training paradigm was also re-written in Psychtoolbox.

As mentioned above, it has been reported that scan success rate in a paediatric population (and in particular those with ADHD) is improved with the use of a mock scanner (Epstein *et al.*, 2007a; Slifer *et al.*, 2002). However as professionally made mock scanners are expensive and require a large area, a low-cost version, with the key components was created instead. The main ideas behind the use of a mock scanner are the ability to practice performing the fMRI task whilst lying down, practising the button presses without looking away from the screen and also knowing the importance of minimising head motion during the scan.

To create a mock scanner, a sturdy camp bed was used with a mirror attached to a home-made frame used to display a laptop screen positioned on a table at the top of the camp bed. As the volunteers were required to view the screen using a mirror, the computer display had to be mirrored to compensate. The design of the mock scanner is shown in Figure 30.

In order to replicate the scanner environment further, a USB PC gaming pad was partly disassembled and wires soldered to the motherboard. The other ends of these wires were soldered to basic "push to connect" buttons. To make these buttons easier to use and more "child friendly" they were attached to basic hand grips. In order to translate PC gaming pad key presses into the equivalent of keyboard presses (the method the behavioural results were output from the scanner), JoyToKey (http://www-en.jtksoft.net/) software was customised to the relevant key presses.

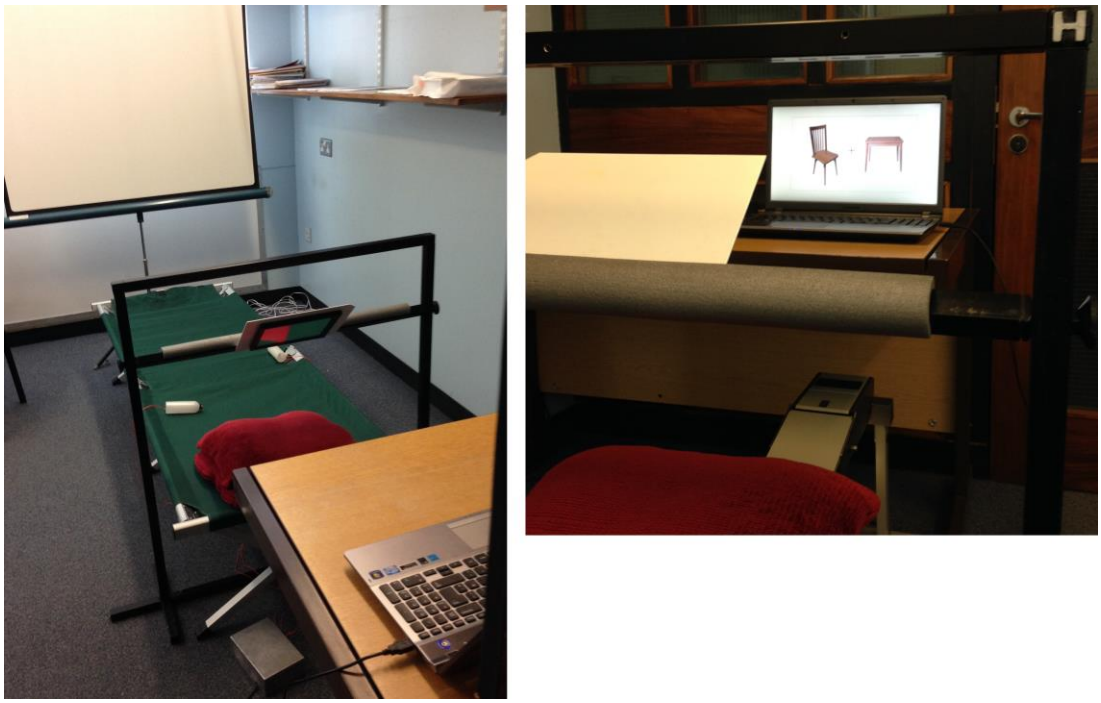

Figure 30: The mock scanner setup used to train volunteers in the iBOCA study.

Finally, in order to educate children and adolescents about the importance of minimising head motion, a bright white light was set-up at the top of the camp bed (underneath the table with the laptop and just above the volunteer's head) and a large projection screen was set-up at the foot of the camp bed. Once the volunteer became confident in performing the task they were asked to continue playing while the light was switched on. When positioned correctly, the light casts a large shadow of the volunteer's head on the projection screen, allowing them to view how much their head moved while performing the practice task. The volunteers were taught the importance of keeping their head as still as possible to get the best possible pictures and were invited to continue performing the task while monitoring their head movement. In addition the volunteers were invited to try various movements (such as kicking their legs) to see how much it might affect scan quality.

A pilot study identified an additional method that could potentially reduce anxiety in a young population prior to the scan. In addition to this basic mock scanner set-up for task training, volunteers were invited to listen to the various noises they would hear during the scan. These were 10-20 second recordings, acquired by the author, of the five scan types they would undergo. If a volunteer opted to listen to the scanner sounds the author talked them through each of the five scan types with information provided of how long each would take alongside the noises they should expect.

The author prepared the majority of the volunteers for their scans.

6.7 Practical Aspects of Scanning: Data quality checks

In addition to re-implementing the fMRI paradigm and being involved with scanning preparation, the author made sure that the data quality was also checked after each scan. This was considered crucial to identify any problem with scanning that could be rectified before the next scanning session. It involved visual inspection of the images, normalising the structural and functional MRI data, and performing 'first level analysis' (of a random effects design) on the fMRI data. The normalisation was performed as described in the Methods section. The most important checks on the data were visual inspection for gross artefacts and inspection of the SPM calculation of the amount of movement in each direction, and the amount of rotation in each direction, over all sessions.

Acceptable scans were defined as those with no gross abnormalities (as checked during visual inspection). For fMRI scans, if movement greater than 1mm or rotation greater than 1° occurred in any direction between scans, the scans were further assessed to identify any artefacts. If a relatively large amount of movement occurred, one or more images in the fMRI volume would show gross artefacts. If the movement was over a short enough time such that only one image contained an artefact, replacing the image with the mean of the two neighbouring images was used as an approximate correction. However, if there were too many images with gross abnormalities then the subject had to be excluded from the analysis.

Regularly performing data quality checks was also crucial as it determined whether a subject's images met requirements to be included in the analysis and had the potential to identify and eliminate any potential systematic errors. Checking if a subject's images meet requirements to be included in the analysis allows for the scanning success rate to be accurately evaluated as scanning proceeds, as above, highlighting if more needs to be done to improve data quality e.g. reduce head movement. Also, identifying potential systematic errors allows scanning problems to be identified and rectified much sooner, saving time and money on scanning. An example of the importance of regular quality assurance checks is the identification of the "phantom pulses", discussed in section 6.4. This allowed potential problems to be identified swiftly and was resolved before another scan took place.

6.8 Theory of ADHD Syndrome Mechanism: the Dopamine Transfer Deficit (DTD) Theory of Altered Reinforcement Mechanisms in ADHD

A number of theories of ADHD have been proposed which could account for the aetiology of ADHD (Plichta and Scheres, 2013). However, the majority of these models do not provide testable hypotheses, unlike the DTD theory (Tripp and Wickens, 2008).

The DTD theory proposes that children with ADHD have an abnormal sensitivity to positive reinforcement. This theory suggests that the mechanisms that cause this abnormal sensitivity occur due to an alteration in the magnitude and timing of anticipatory dopamine cell firing.

When an unexpected reward is received, dopamine is released shortly after the reward event. As the learning between association of making a certain response,

such as choosing one of the presented stimuli, and receiving a reward increases, the release of dopamine increases at the response time (e.g. decision time) and decreases at the time of the actual reward delivery. The DTD theory proposes that the transfer of dopamine release from the reward/feedback time towards the response/decision time is abnormal in children with ADHD. The DTD theory is illustrated in Figure 31.

The DTD theory centres around *five* key *testable* hypotheses. First, as learning proceeds, the dopamine response to the receipt of rewards is increasingly transferred to the decision/response time in healthy control children (as shown in Figure 31(A)). Second, in healthy controls, this transfer of dopamine towards the response time is maintained when the reinforcement is either delayed or intermittent. The third postulate is that for children and adolescents with ADHD, the dopamine transfer discussed in the first postulate (dopamine response at the feedback time transferring towards the decision time) is abnormal; specifically, the dopamine cell response at the response time is lower than in healthy controls due to a failure to transfer the dopamine signal from the feedback time (as shown in Figure 31(B)). The fourth hypotheses in the DTD theory is that when children with ADHD receive intermittent or delayed feedback, dopamine signalling also becomes intermittent or delayed. In other words, if an association between a response and obtaining a reward is learned and then the reinforcement becomes delayed or intermittent, then the learned association is not maintained (the dopamine signal at the response time becomes delayed or intermittent also). Finally, if the feedback which is reinforcing the dopamine signal is discontinued, healthy controls tend to maintain their response longer than children with ADHD due to their increased anticipatory dopamine release taking longer to be extinguished.

The modified Pessiglione paradigm used in the iBOCA study is well suited to test these five postulates of the DTD theory as it contains reward trials which involve learning which stimulus at the decision time would most likely provide a reward. In addition to testing the DTD theory on the reward trials, the paradigm will also test whether the theory requires additional considerations with respect to loss/aversive information.

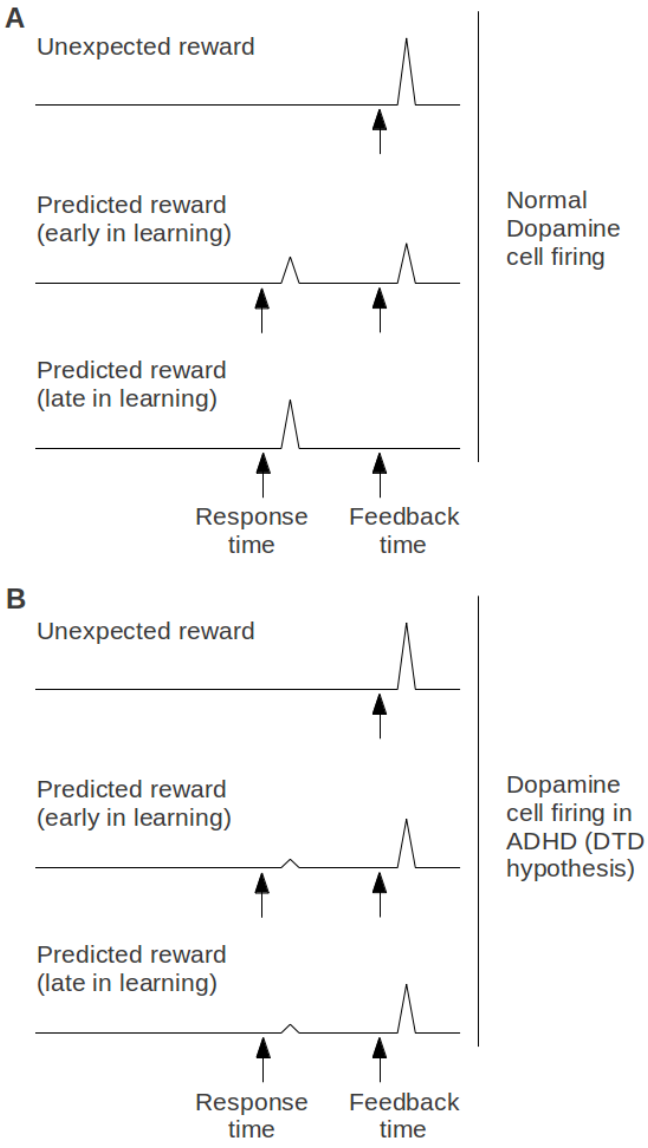

Figure 31: (A) A pictorial representation of the expected magnitude and timing of anticipatory dopamine cell firing in controls, compared to (B) the abnormal dopamine cell firing in the DTD theory. The DTD theory suggests that children with ADHD fail to correctly adjust their dopamine cell firing rate from the reward reinforcement/feedback time towards the response time in anticipation of a reward.

6.9 Discussion

The practicalities of scanning children and adolescents with ADHD were reviewed. For most children aged 12 years or older, the scanning time can be extended up to 45 min but an extra 20-30% of subjects should be recruited to compensate for expected failed scans. If mostly younger children are planned to be scanned, the scanning time should not exceed 30 min and recruitment of ~50% extra subjects should be aimed for. Pre-scan training of children is important to maximise the scan success rate and was implemented in this study.

The quality of scans is very important to allow interpretation of results. Gross scanning artefacts must be avoided, it is necessary to establish robust QA programs for neuroimaging research labs, and every scan should be visually inspected for artefacts, even within a large dataset. Therefore, data quality checks were performed after each scan to ensure the image quality was maintained at a high level throughout the study.

Although recruitment of unaffected siblings of children and adolescents with ADHD is more frequently associated with genetics studies, it has been shown that unaffected siblings share some of the anatomical differences associated with ADHD and, as such, form an intermediate group (Castellanos *et al.*, 2003; Durston *et al.*, 2004). For the analysis outlined in this chapter, unaffected siblings and healthy subjects with no family history of ADHD are combined in the healthy control group. The reason for this is that the unaffected siblings have not shown significant symptoms of ADHD despite the same genetic and environmental background of volunteers with ADHD and therefore differences in brain anatomy or function could help elucidate why the patient group developed these symptoms. Whilst this is a potential limitation of the study as fewer differences may be identified between the patient and controls groups, any differences identified may be more closely linked with ADHD symptoms. The inclusion of this group in the study also allows further investigation between the three groups, similar to the studies by Castellanos and colleagues (2003) and Durston and colleagues (2004). For example, differences identified between volunteers with a family history of ADHD and those with no family history of ADHD could be linked with risk factors of ADHD.

Unfortunately, due to a number of administrative, financial and recruitment delays in the study outwith the control of the author, there have been too few subjects successfully scanned for the planned analyses to take place.

As the prediction between ADHD patients and healthy controls using structural MRI data has been performed previously, it is anticipated that this can be replicated once enough data are acquired. In addition, the potential of the use of fMRI data in classification algorithms is encouraging due to the results from the MDD study in Chapter 8, which uses the same fMRI paradigm (for different reasons). Diagnostic classification analyses and testing of the DTD theory may increase the understanding of ADHD mechanisms. Additionally, the prediction of MPH response in children and adolescents with ADHD could have a substantial impact on the understanding of the mechanisms which support MPH response and, potentially, future clinical practice.

Chapter 7: Diagnostic Classification and Prediction of Symptom Severity in MDD

7.1 Introduction

MDD is a mood disorder which is associated with persistent and disabling symptoms of low mood, anhedonia, hopelessness, guilt, low self-worth, poor concentration, lack of energy, suicidal thoughts and altered appetite and sleep (American Psychiatric Association, 2000) with no established pathophysiological mechanisms or biomarkers. There have been a large number of studies which have reported group-level differences in brain structure between patients with MDD and healthy controls, the majority of which have reported reductions in MDD patients' grey matter volume compared with healthy controls (Fu *et al.,* 2003; Haubold *et al.,* 2012; Koolschijn *et al.,* 2009; Shah *et al.,* 1998).

Grey matter abnormalities in MDD subjects, compared with healthy controls, have been identified most consistently in the bilateral rostral anterior cingulate cortex (Bora *et al.,* 2012). Kempton *et al.* (2011) reported that MDD subjects had larger lateral ventricular and CSF volumes compared with controls. Other regions reported to have decreased grey matter volume in at least one meta-analysis include the putamen, caudate, insula, globus pallidus, thalamus, hippocampus and many areas within the frontal lobe (Bora *et al.,* 2012; Kempton *et al.,* 2011; Koolschijn *et al.,* 2009). The amygdala and thalamus volumes were reported to show no significant differences between groups by Kooschijn *et al.* (2009) but the latter was found to be decreased in MDD patients in a more recent meta-analysis by Kempton *et al.* (2011).

To the author's knowledge, no white matter (T_1 weighted MRI) VBM studies have been reported in MDD. However a number of DTI studies have been reported. All studies either found no significant differences in fractional anisotropy (FA, a measure of the level of direction/orientation of water diffusion) between groups (Kieseppä *et al.,* 2010; Korgaonkar *et al.,* 2011) or only decreases in FA (Cole *et al.,* 2012; Li *et al.,* 2007; Ma *et al.,* 2007; Steele *et al.,* 2005; Zhu *et al.,* 2011; Zou *et al.,* 2008). A decrease in the FA in the anterior limb of the internal capsule is the most consistently replicated finding (Cole *et al.,* 2012; Zhu *et al.,* 2011; Zou *et al.,* 2008). Other regions which have been identified as having reduced FA in MDD subjects include the parahippocampal gyrus, posterior cingulate cortex, corpus callosum,

superior longitudinal fasciculus, anterior corona radiate, superior and middle frontal gyri, lateral occipitotemporal gyrus, and subgyral and angular gyri of parietal lobe (Cole *et al.*, 2012; Li *et al.*, 2007; Ma *et al.*, 2007; Zhu *et al.*, 2011; Zou *et al.*, 2008). In addition, mean diffusivity (MD, a mean of the amount of water diffusion) was found to be increased in the MDD group in comparison with controls in the corpus callosum (Cole *et al.*, 2012).

Regions identified using VBM provide information about group level differences. In contrast, techniques based on machine learning such as SVM, which include additional techniques such as feature selection, can determine which brain regions consistently differ between groups in order to produce an accurate individual subject classifier. Kipli *et al.* (2013) tested four different feature selection techniques with four different machine learning methods by attempting to classify structural MRI images (using information extracted from structural MRI e.g. volumes of various structures) of depressed individuals and healthy controls. The author suggested that the Information Gain algorithm outperformed OneR, SVM (using RFE) and ReliefF feature selection methods as it achieved the highest average accuracy (72%) when applied to four different classifiers (Kipli *et al.*, 2013). As mentioned in Chapter 3, a concern is, however, that 77% (88/115) of subjects in this study belonged to the control group, and as the sensitivity and specificity of the results are not reported, it is unclear if the large class imbalance is an issue. Another study that focused on the results from various feature selection methods, attempted to predict diagnosis between bipolar disorder and healthy controls, achieving accuracies ranging between 60-90% (Termenon *et al.*, 2013). Also, Costafreda *et al.* (2009a) used machine learning to classify MDD subjects and healthy controls, achieving 68% accuracy.

A multicentre study by Mwangi *et al.* (2012a) successfully classified structural MR images between people with depression and healthy controls with the images obtained over two scanning centres. Mwangi *et al.* (2012a) implemented both an SVM and RVM approach with the latter achieving a slightly higher classification accuracy (90%). In that study, grey matter volume reductions were identified in MDD compared to controls in the dorsolateral prefrontal cortex, medial frontal cortex, orbitofrontal cortex, temporal lobe, insula, cerebellum and posterior lobe.

Mwangi *et al.* also used RVR to predict illness severity (Mwangi *et al.*, 2012b). In that study, they found that it was possible to significantly predict the BDI

scores from the whole-brain structural MRI scans, but not the HAM-D (Mwangi *et al.*, 2012b).

This is the only study which has applied machine learning to predict individual severity scores in MDD, but there are a few studies which have calculated group level correlations with severity scores. The majority of these studies have performed correlation analyses with the HAM-D score. Vakili *et al.* (2000) found that the bilateral hippocampal volume was negatively correlated in males, but not females. Studies have also identified the bilateral dorsal prefrontal, bilateral medial frontal, inferior and superior frontal, orbitofrontal and cingulate cortices, bilateral temporal fusiform gyrus, occipital lobe, inferior temporal gyrus, amygdala/parahippocampal gyrus and postcentral gyrus as brain regions which are negatively correlated with the HAM-D score in MDD subjects (Chen *et al.*, 2007; Li *et al.*, 2010). In addition, a positive correlation between the HAM-D score and the occipital cortex and cerebellum were identified (Chen *et al.*, 2007). The only study which reported significant correlations with the BDI score found that decreased grey matter volume in the right planum temporale correlated with an increase in BDI score (Takahashi *et al.*, 2010), however, Kim *et al.* (2008) could not find a significant correlation between BDI scores and volume estimates within a number of *a priori* regions of interest. To the author's knowledge, no study has performed a correlation between structural MRI and Montgomery-Åsberg Depression Rating Scale (MADRS) scores.

Again, as there are no reports of white matter structural MRI correlations with symptom severity scores, correlations identified using DTI can be used to determine changes in white matter volume. Li *et al.* (2007) found no correlation between any of the prefrontal ROIs investigated and HAM-D scores but Zou *et al.* (2008) found that the anterior limb of the internal capsule, the region that was most frequently reported to have decreased FA in MDD subjects compared to controls, negatively correlated with the HAM-D score. The same region was also identified as negatively correlating with the Center for Epidemiologic Studies Depression Scale (Zhu *et al.*, 2011). FA values within the corpus callosum and posterior tracts of subjects with MDD have also been found to negatively correlate with BDI scores (Cole *et al.*, 2012).

The main goal of the work described in this chapter was to attempt to accurately classify structural MRI scans of MDD patients and healthy controls and to

investigate whether symptom severity scores could be accurately predicted in the patient group, similar to the studies by Mwangi *et al.* (2012a; 2012b). The machine learning studies discussed all used grey matter structural MR images only. White matter images are less frequently investigated and this is the first work applying machine learning techniques to T_1 weighted white matter images from patients with MDD.

7.2 Method

7.2.1 Subjects

Structural T_1 weighted scans were obtained from subjects at the Clinical Research Centre, Ninewells Hospital and Medical School in Dundee, UK. Informed consent was obtained from all volunteers. The study protocols were approved by the local Ethics Committee.

Twenty adults with a past or present diagnosis of MDD were recruited from the Advanced Interventions Service in Dundee. Diagnoses were made by experienced clinicians according to DSM IV criteria using the MINI PLUS interview schedule (Sheehan and Lecrubier, 1992). Exclusion criteria included potentially confounding diagnoses – any other primary psychiatric disorder, substance misuse or significant head injury. 18 MDD participants were being treated with one or more anti-depressant medication (venalfaxine (6), sertraline (3), trazodone (3), citalopram (2), fluoxetine (2), isocarboxazid (2), mirtazapine (2), ltryptophan (1), phenelzine (1), tranylcypromine (1)). In addition, 7 MDD participants were being treated with anti-psychotic medications (quetiapine (6) and chlorpromazine (1)) and 3 MDD participants were being given lithium augmentation.

Twenty-one healthy, never-depressed controls were recruited mostly from partners, relatives and friends of patients and underwent psychiatric screening using the same semi-structured interview schedule as the patient volunteers. None of the controls had a history of current or past psychiatric or neurological disorder and none were taking medication.

All MDD and control volunteers had a predicted premorbid Full Scale Intelligence Quotient above 106 (one control was not assessed for IQ) as assessed by the National Adult Reading Test (NART). Handedness was assessed using the EHI (Oldfield, 1971). Apart from 2 left-handed subjects in the control group and 1 and 3

ambidextrous subjects in the control and patient groups respectively, all subjects were right-handed (although one control and one patient could not be reported due to incomplete data).

7.2.2 Image Acquisition

For each participant structural whole-brain images were acquired using a 3T Siemens Magnetom TrioTim syngo scanner using a T_1-weighted MP-RAGE sequence with the following parameters: TR = 1900 ms, TE = 2.64 ms, flip angle = 9°, FOV = 200 mm, matrix = 256 x 256, 176 slices, voxel size 0.8x0.8x1 mm, slice thickness 1 mm.

7.2.3 Image Pre-processing

All scans were visually inspected for artefacts and particular care was taken to identify motion artefacts which appear as blurring or 'ghosting' (McRobbie *et al.*, 2010). No scans showed blurring, ghosting or other gross artefacts. No scans were excluded from analysis.

Pre-processing was performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm). The procedure involved segmentation of T_1 weighted images into separate grey matter, white matter and CSF compartment images and normalisation of the grey and white matter segmented images towards to the default SPM8 anatomical template. The resultant images were smoothed with an 8 mm FWHM Gaussian kernel.

In addition to the standard segmentation, warping and smoothing steps, a modulation step was included, as recommended in the SPM manual for structural MRI normalisation. This means that if a region is increased in volume during normalisation then the intensity within the region is proportionally reduced to preserve the overall intensity and, correspondingly, if the region's volume is decreased the intensity is increased accordingly.

7.2.4 Neuroimaging data quality - Outlier analysis

As mentioned in Chapter 2, assessment of both the quality of the normalisation and the quality of each subject's images is important before drawing conclusions from

the data. An outlier analysis, defined in section 2.7, was performed on the DARTEL pre-processed and standard VBM pre-processed images in this study (both using modulation, as suggested in the SPM8 manual (Ashburner *et al.*, 2012)). This approach allows the identification of subjects with a high proportion of voxels which significantly differ from their corresponding group. As subjects would be expected to have a similar brain structure within each group, identification of many outliers in one subject suggests either the subject should not be included in the analysis due to a gross brain abnormality (which was not present in the rest of the group) or that the outliers were produced during the normalisation procedure. As all subjects were inspected for gross brain abnormalities, it is suggested that a high proportion of outliers may be due to limitations in the normalisation procedure.

In general, the "Preserve Concentrations" (no modulation) alteration seemed to reduce the number of outliers in the subjects with the highest number of outliers when modulation was performed without dramatically altering the number of outliers in the other subjects. However, as modulation is suggested for structural MR images, DARTEL and standard VBM approaches with modulation were compared.

The outlier analysis on the DARTEL processed images showed that the highest percentage of outliers in one subject was 11% for the grey matter images and 13.7% for the white matter images. Two subjects had a percentage greater than 10% of their grey matter voxels as outliers and one subject for the corresponding white matter analysis. The mean percentages of outliers were similar in both the grey and white matter analysis (2.1% and 1.9% respectively) with a higher number of outliers in the control group in both cases. Figure 32 displays the DARTEL pre-processed results for both grey and white matter images, when each diagnostic group is considered separately, as in the outlier analysis.

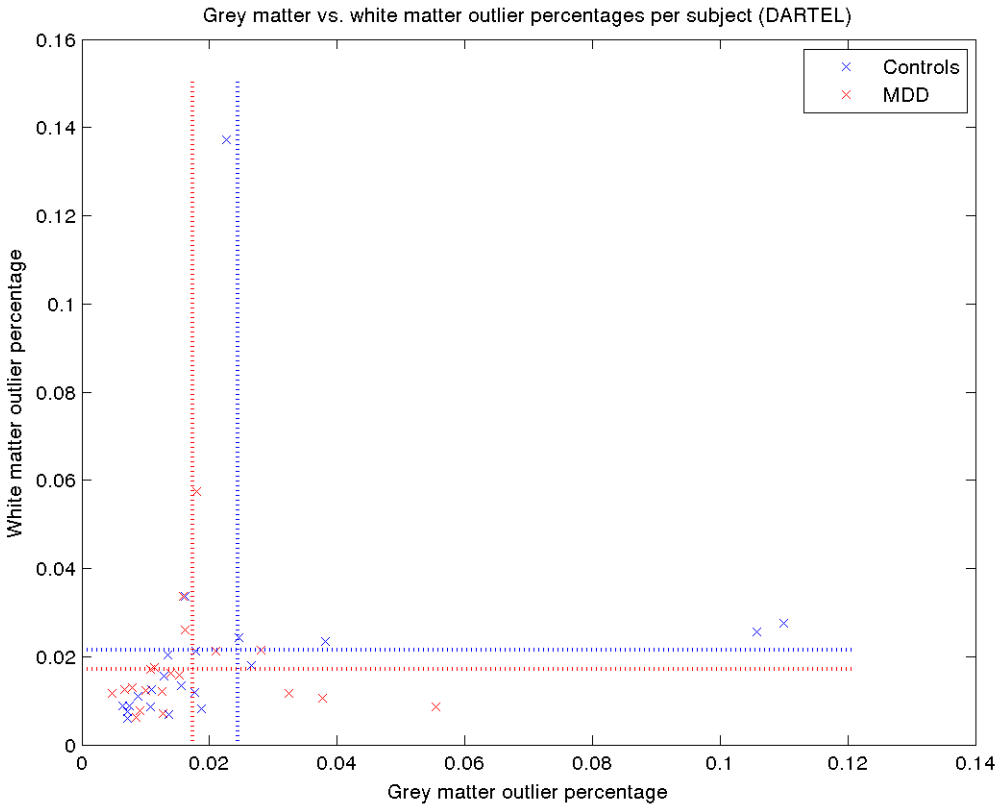

Figure 32: A plot showing the percentage of voxels considered to be outliers in each diagnostic group (blue- controls, red- MDD) when using DARTEL. The dotted lines indicate the mean values for each diagnostic group and each imaging modality.

Using standard VBM, the subject with the highest proportion of outliers was found to have 13.5% of grey matter voxels considered outliers and, similarly, 9.7% of voxels was the maximum percentage of white matter outliers. Although the maximum percentage was higher in the grey matter images using standard VBM, the overall percentage was lower due to the drastically reduced maximum percentage of white matter outliers. Only one subject had a percentage of outliers greater than 10% in either the grey or white matter images – two fewer than when using DARTEL. In addition, the mean percentages of outliers were decreased in both grey and white matter images (1.9% and 1.8% respectively). The mean percentages of outliers for controls and patients were all reduced in comparison with the DARTEL approach, with the exception of patients' white matter images. The difference in the percentage of outliers between patients and controls was also reduced using the standard VBM approach compared with DARTEL. Therefore, although the DARTEL approach has been reported to show greater pre-processing accuracy, it seems the DARTEL method may not be optimal for this dataset. Figure 33 displays the standard VBM pre-processed results for both grey and white matter images, when each diagnostic group is considered separately, as in the outlier analysis.

On closer inspection, a trend was identified between basic brain statistics and the percentage of outliers when using standard VBM. Total brain size was identified to be significantly positively correlated ($p = 0.001$) with the percentage of white matter outliers, indicating that this method tended to be worse for normalisation of larger brains. In particular, a large CSF/brain ratio positively correlated strongly with the percentage of grey and white matter outliers ($p= 0.011$ and $p < 0.001$, respectively) which suggests the method struggles to normalise subjects with large ventricles. This is a concern as Elkis *et al.* (1995) identified a correlation between ventricular enlargement and MDD during a meta-analysis. Neither of these statistics, nor any other statistics investigated, correlated with the percentage of outliers when using DARTEL, therefore no justification for the poor performance of DARTEL normalisation could be identified.

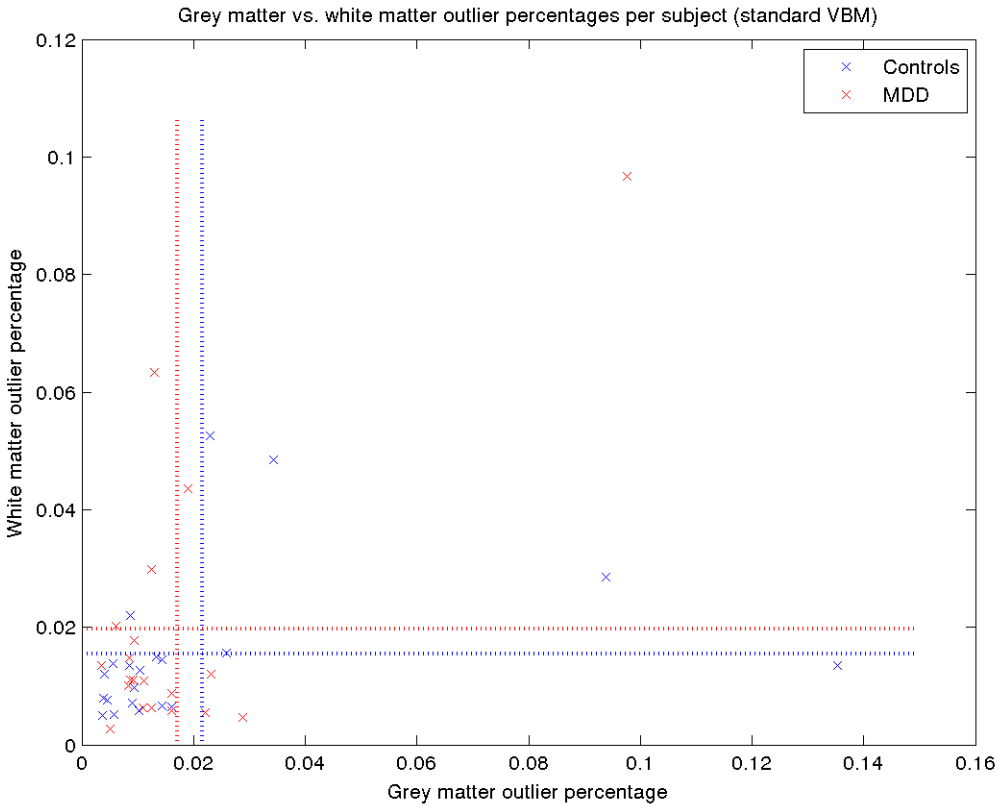

Figure 33: A plot showing the percentage of voxels considered to be outliers in each diagnostic group (blue- controls, red- MDD) when using standard VBM. The dotted lines indicate the mean values for each diagnostic group and each imaging modality.

In a further investigation into the outlier issues, a third diagnostic group was investigated; an MDD neurosurgical group (i.e. the patients had received neurosurgery as treatment for refractory depression) of fifteen subjects (not discussed further in this thesis). The inclusion of this third group increased the number of outliers and the extent of outliers in the DARTEL-processed images. One subject in the non-surgical MDD group was found to have 25% outliers within the grey matter images and 18% outliers in the white matter images when compared with the other non-surgical MDD subjects. The number of subjects that had greater than 10% of their brain considered outliers with respect to their diagnostic group doubled in the control and non-surgical MDD patients for both grey and white matter. As standard VBM pre-processes each subject individually, the inclusion of the third group did not alter the number or extent of outliers using grey or white matter in the control and non-surgical group at all. Therefore, as standard VBM seemed to perform better with respect to outliers and was more robust to new data, this technique was used for pre-processing the structural MR images.

7.2.5 Individual Scan Classification

Machine learning to allow individual predictions to take place was implemented in Matlab (The Mathworks Inc.) using an SVM toolbox (Schwaighofer, 2001), the PRoNTo toolbox for the RVR analysis (Schrouff *et al.*, 2013) and custom Matlab scripts.

SVM and RVR analysis both consisted of two stages: training the classifier, then testing the accuracy using data not used for training. In both cases, standard LOOCV was used for training with the SVM parameters being selected on the basis of training stage accuracy, whilst the RVR parameters were selected on the basis of a combination of three standard statistical variables: RMSE, MAE and Pearson's correlation coefficient (R) as calculated using the toolbox (http://www.mathworks.com/matlabcentral/fileexchange/22020-goodness-of-fit-modified/content/gfit2.m).

Feature selection was used to identify brain regions supporting predictive classification. The feature selection method chosen during the SVM classification was a standard t-test, as implemented in the SPM toolbox. A t-test was performed for each LOOCV step (with the subject being classified removed from the training

process) with significance defined as $p<0.05$ at a whole brain corrected cluster level (Slotnick *et al.,* 2003). The z-scores at each of the significant voxels were then ranked during LOOCV and the threshold, whereby voxels with z-scores above this threshold would be included in the classification, was optimised at the same stage as the SVM parameter selection, as described in Chapter 2.

The RVR procedure was performed on patients only both with and without feature selection. The feature selection method selected for the RVR prediction involved multiple linear regressions as implemented in the SPM toolbox and the method for optimising the feature selection was analogous to the classification procedure and is described in more detail in Chapter 2. In addition, multivariate feature selection (RFE, as described in Chapter 2) was also tested to see if it could improve the prediction. RVR was used to investigate whether various symptom severity scores such as the HAM-D, the MADRS and the BDI could be predicted on individual subjects. Higher scores in each of these symptom severity scores indicate more severe symptoms of depression. As discussed in section 2.12, the RVR procedure assumes the variable being predicted (e.g. HAM-D, MADRS or BDI) is normally distributed. The Shapiro-Wilk test for normality found that each of these symptom severity scores satisfied this requirement.

7.2.6 Group Level Comparisons

The *group* level analysis involved performing the same calculations as described in the feature selection (excluding RFE) except the calculations were based on all subjects rather than all but the one dataset left out for classification.

For the conventional *group* level VBM analysis, the null hypothesis of no difference in brain structure between patients and controls was tested using an unpaired t-test as implemented in SPM8. The *group* level regressions were performed as implemented in SPM8 using patients' images and various symptom severity scores such as the HAM-D, the MADRS and the BDI.

In both analyses, significance was defined as $p<0.05$ at a whole brain corrected cluster level (Slotnick *et al.*, 2003).

7.3 Results

7.3.1 Participant Characteristics

Age and IQ did not differ significantly (t-test, $p>0.1$, excluding the control subject who failed to complete the IQ test from the IQ t-test calculation) and gender was not significantly different (as assessed by a chi-square calculation) between groups. The MDD group mean age was 51.8 years (standard deviation 11.2) mean IQ was 122.8 (standard deviation 4.7). The control group mean age was 46.1 years (standard deviation 14.0) and the mean IQ 122.8 (standard deviation 5.8).

The average HAM-D, MADRS and BDI illness severity rating scores in the MDD group were 16.1, 22.5 and 32.2, indicating depression severity in the moderate range. The degree of treatment-resistance was quantified by detailed inspection of the clinical notes rated according to the Massachusetts General Hospital (MGH-S) staging method (Fava, 2003). The MGH-S takes account of the number of failed 'adequate' (i.e. exceeding a minimum dose and duration of a given medication) antidepressant treatment trials, including optimisation of antidepressant dose, antidepressant combinations and treatment augmentation. The average treatment-resistance MGH-S score was 13.3 which is similar to a previous assessment of patients attending the AIS (15.5) and significantly greater than typical secondary care psychiatric (5.3) and primary care (0.5) treatment-resistance levels (Hazari *et al.*, 2013).

These results are outlined in Table 8.

Table 8: Clinical descriptors for the MDD and healthy control groups in the structural MRI analysis. Variables are shown as mean (standard deviation). *chi-square test with other tests being t-tests.

	MDD	Controls	
Age	51.80 (11.23)	46.14 (13.97)	n.s.
IQ	122.75 (4.71)	116.95 (27.38)	n.s.
Female/Total*	15/20	15/21	n.s.
HAM-D	16.10 (5.58)	0.48 (0.93)	<0.001
MADRS	22.50 (7.97)	0.48 (1.03)	<0.001
BDI	32.20 (11.38)	0.43 (0.87)	<0.001
MGH-S	13.25 (10.49)	N/A	N/A

7.3.2 Individual Subject SVM Predictions

A Gaussian SVM was used to analyse 20 structural MRI images of adults with a past or present diagnosis of MDD and 21 structural MRI images of control subjects matched for age, gender and IQ. Feature selection was implemented using t-tests with a variable threshold which was optimised during cross-validation. The analysis was done using the grey and white matter compartment of T_1 weighted images separately.

The analysis using grey matter images alone resulted in an individual subject predictive accuracy of 85% (sensitivity 0.85, specificity 0.86, $\chi^2 = 17.7$, $p < 0.0001$).

The analysis using white matter images alone resulted in a poorer accuracy of 71% (sensitivity 0.45, specificity 0.95, $\chi^2 = 6.9$, $p < 0.0085$).

7.3.3 Brain Regions identified using Feature Selection

When grey matter images were used for analysis, the largest regions supporting individual prediction at accuracy of 85% were identified in the caudate, insula, and periventricular grey matter. All the additional smaller regions which were used in the classification overlapped with the VBM results (discussed in section 7.3.4) aside from a few insignificant regions which rose above the significance level when the test subject's image was not included in the t-test. Grey matter regions used in the classification are shown in Figure 34.

When only white matter images were used for analysis, the largest brain region supporting 71% accuracy of prediction was in the cingulate gyrus. However, smaller regions in the posterior cingulate and the white matter deep to the insula were also identified. White matter regions used in the classification are shown in Figure 35.

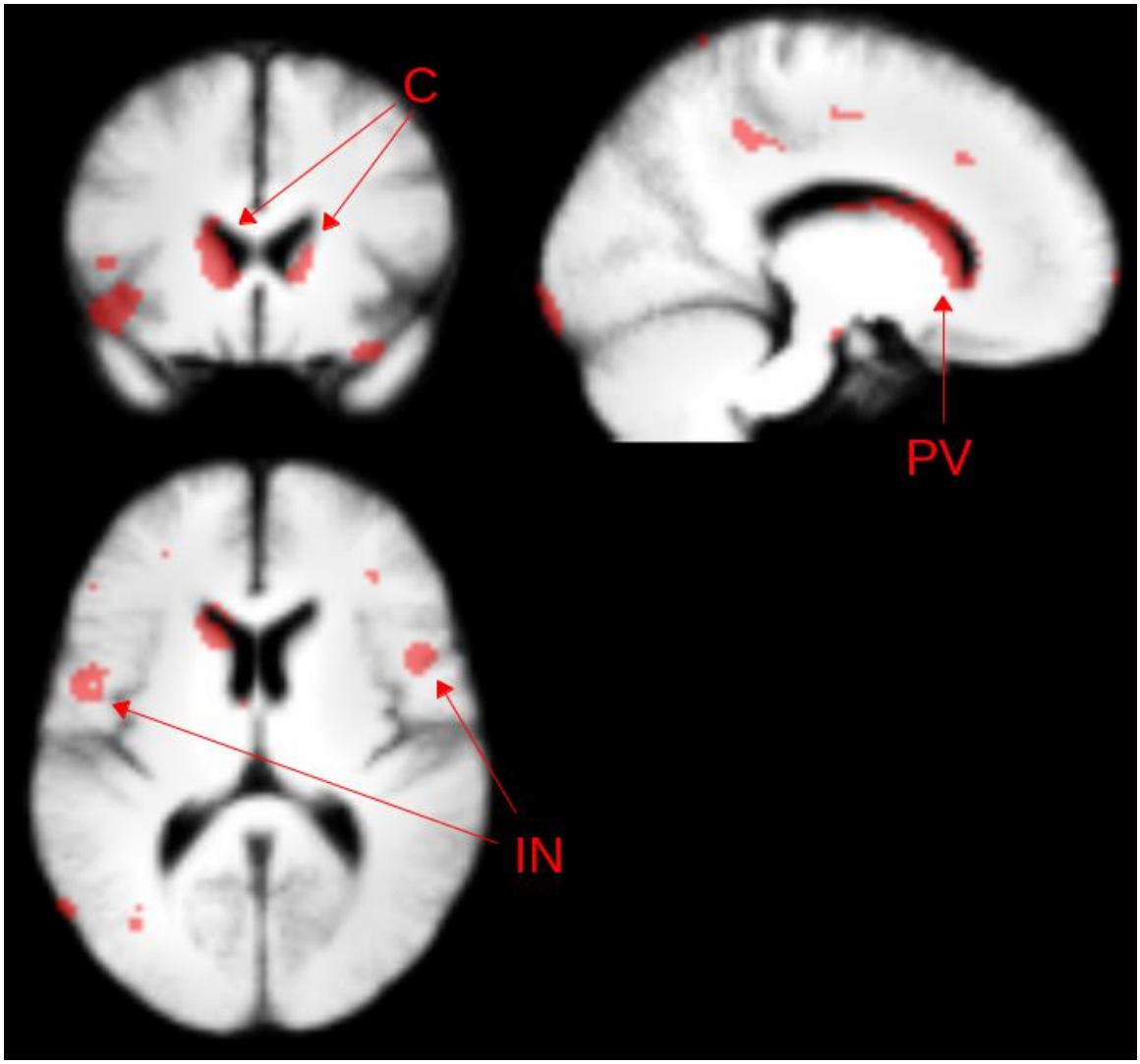

Figure 34: Feature selection (Gaussian SVM) identified brain regions in *grey* matter. PV – periventricular grey matter; C - caudate; IN - insula.

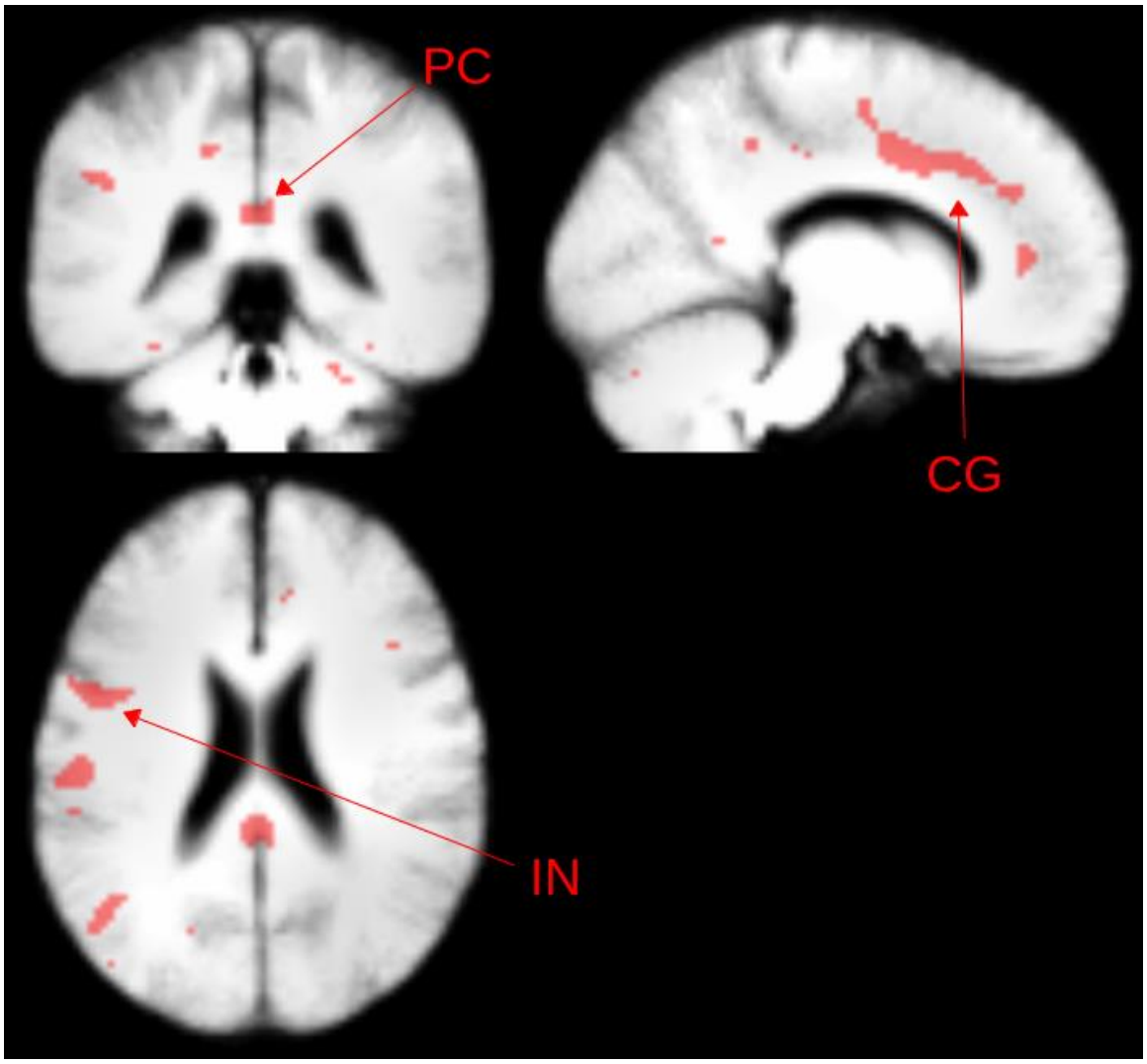

Figure 35: Feature selection (Gaussian SVM) identified brain regions in *white* matter. CG – cingulate gyrus; PC – posterior cingulate; IN – white matter deep to the insula.

7.3.4 VBM Analysis (t-test)

The brain regions identified using feature selection overlapped greatly with the results of the VBM group level analysis ($p<0.05$, whole brain level significance) as t-tests were used in both cases (although the t-tests used during feature selection did not include the subject being classified and not all significant voxels from the t-tests were used in the prediction as it was thresholded). In the VBM analysis, mostly grey matter volume *reductions* were identified in patient group but a small number of increases were also found. As shown in Figure 36, the largest grey matter volume *reductions* were found in the caudate, insula, and periventricular grey matter. In addition, patients were also found to have significantly reduced habenula volume in comparison to controls. Identification of a reduced habenula volume is interesting as deep brain stimulation has previously been attempted in the lateral habenula in treatment resistant depression (Sartorius and Henn, 2007) and an MRC funding has recently been awarded to Dr Roiser to investigate "Habenula function in major depression".

The VBM analysis showed very few white matter volume *reductions* in the patient group such as white matter deep to the insula and a small region in the frontal lobe. Unexpectedly, an *increase* in patient's white matter volume was identified in the cingulate gyrus and the posterior cingulate. Figure 37 shows the white matter VBM results.

Figure 38 and Figure 39 also show the overlap between the VBM results and the results from the feature selection for grey and white matter respectively. It is unsurprising that these overlap so strongly as t-tests were used in both cases with the minor differences being that the t-tests used during feature selection obviously did not include the subject being classified and also as the t-tests were further optimised during feature selection, a subset of the significant voxels from the t-tests were used in the prediction.

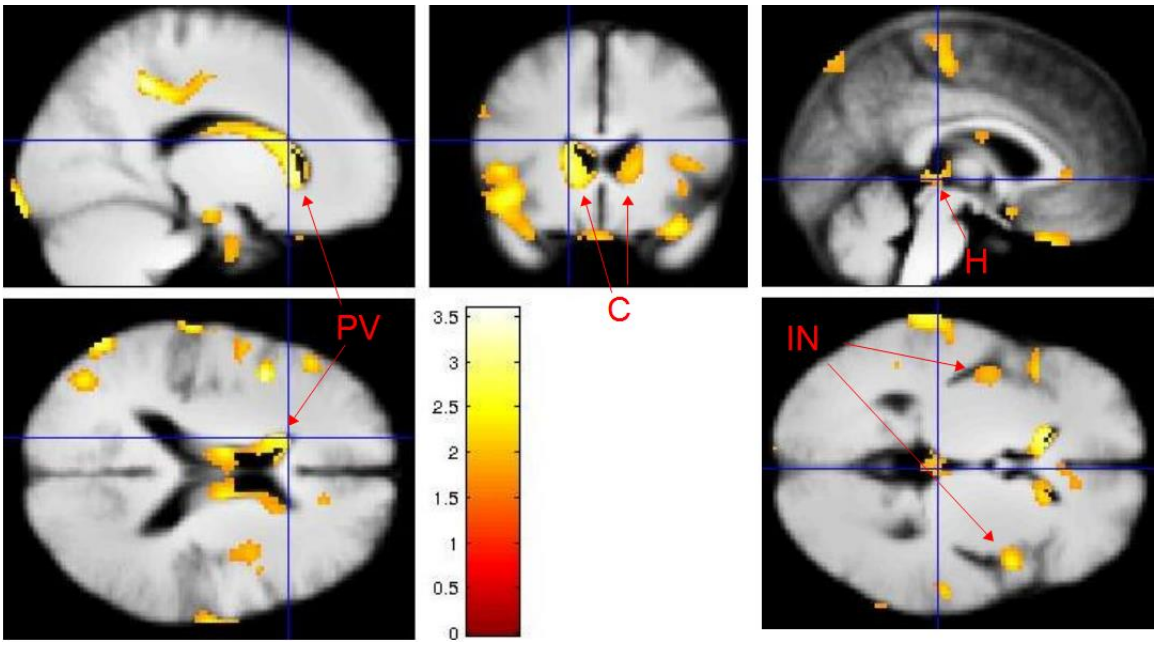

Figure 36: *Group* level *grey* matter volume *reductions* in patients with MDD compared with healthy matched controls. PV- periventricular grey matter, C – caudate reductions, H – habenula and IN – insula.

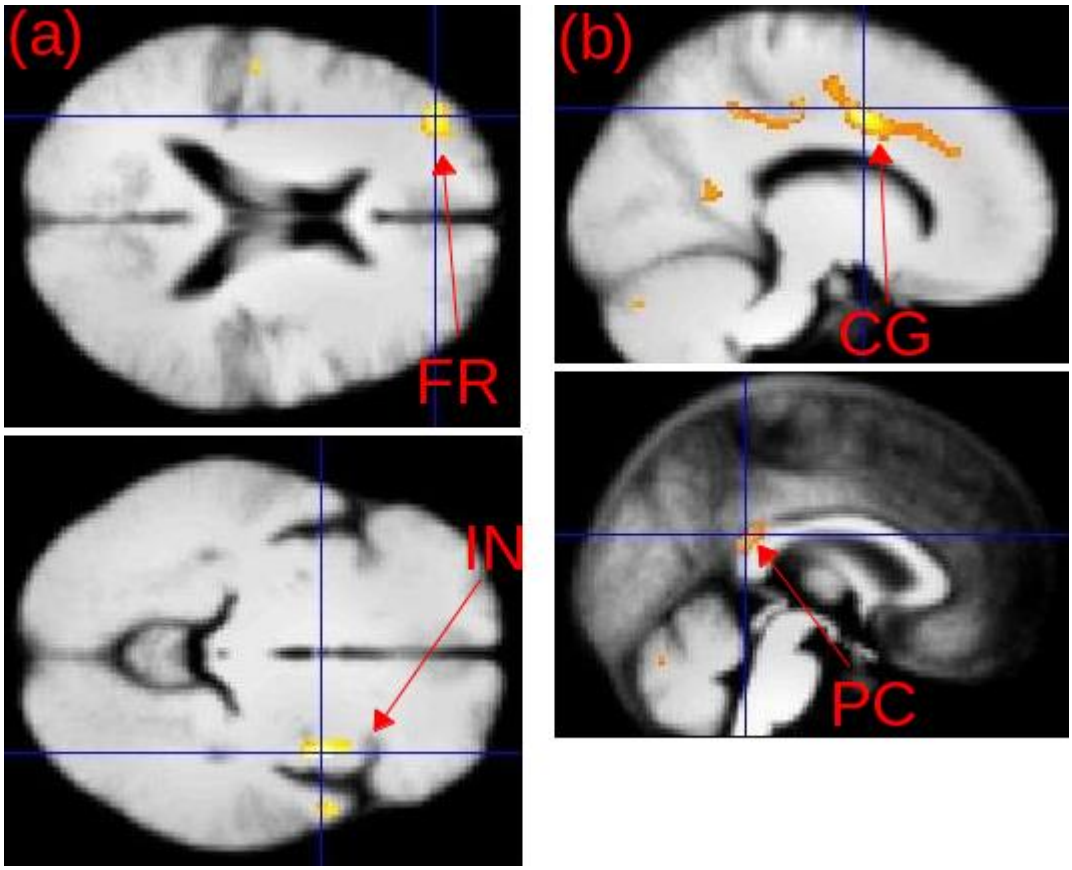

Figure 37: *Group* level (a) *reductions* and (b) *increases* in *white* matter volume in patients with MDD compared with healthy matched controls. FR- frontal region, IN – white matter deep to the insula, CG – cingulate gyrus and PC – posterior cingulate.

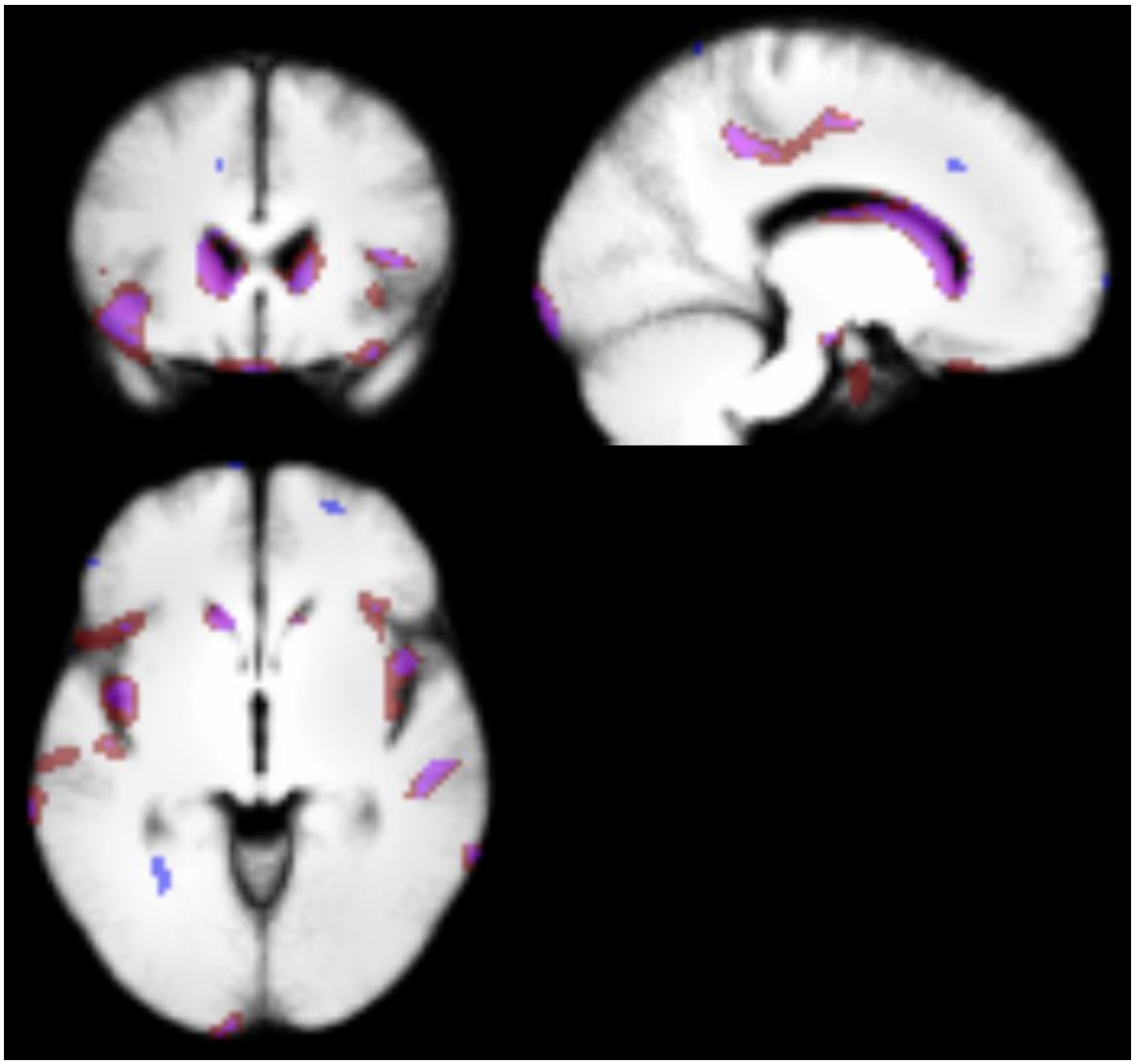

Figure 38: Overlapping grey matter regions between features selected during classification (purple/blue) and regions selected in the VBM analysis (red/purple).

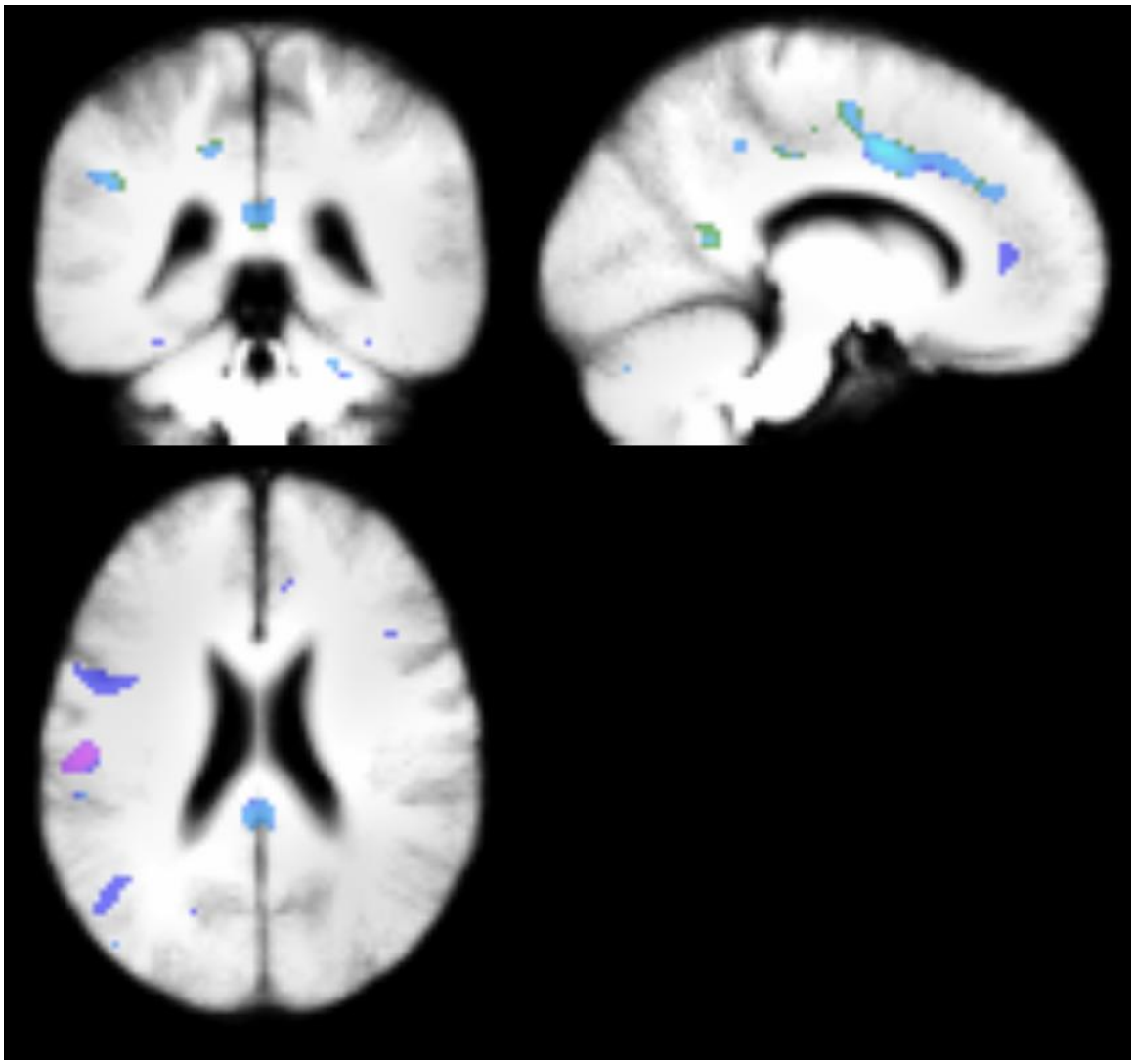

Figure 39 Overlapping white matter regions between features selected during classification (purple/blue) and regions selected in the VBM analysis (green/blue).

7.3.5 Whole Brain Individual Patient RVR Severity Score Predictions

A linear kernel RVR was used to try to predict symptom severity scores (HAM-D, MADRS and BDI) using 20 structural MRI images of adults with a past or present diagnosis of MDD. The results outlined in this section are based on whole brain images (no feature selection, using either grey or white matter images separately).

Using a linear kernel and whole brain grey matter images resulted in a significant correlation between the true and predicted HAM-D scores (RMSE = 4.6963, MAE = 3.6212, R = 0.50712, p =0.0225). A significant correlation was also identified between the true and predicted MADRS scores (RMSE = 6.8328, MAE = 5.441, R = 0.4822, p =0.0314).

Using whole brain white matter images, the correlations increased in significance between the true and predicted HAM-D (RMSE = 4.1315, MAE = 3.3398, R = 0.65662, p =0.0017) and MADRS (RMSE = 5.8122, MAE = 4.6029, R = 0.66422, p =0.0015) scores.

The best fit line between true and predicted scores for both the HAM-D and MADRS predictions, and for both grey and white matter-based predictions, is shown in Figure 40 and the brain regions which had the highest weights are shown in Figure 41.

However, no significant positive correlations were found using the BDI scores using either grey or white matter images.

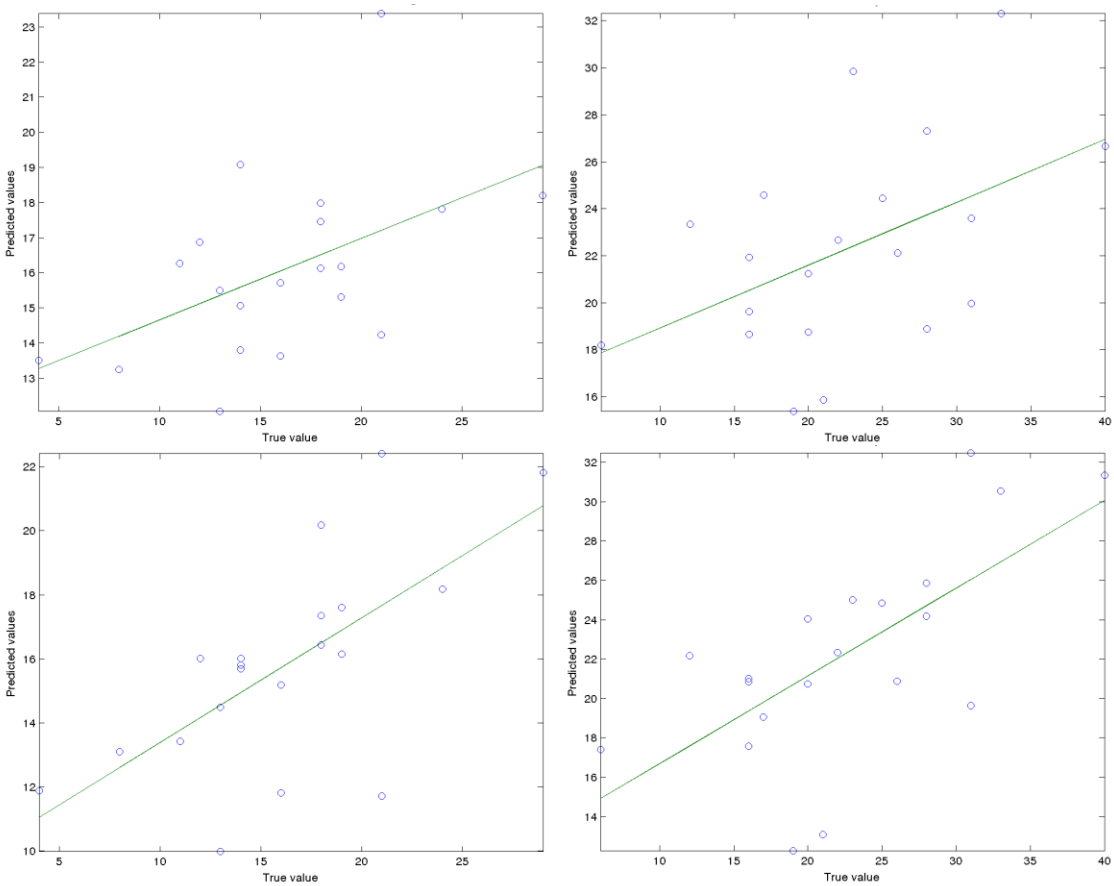

Figure 40: The best fit lines for whole brain severity score predictions (top: grey matter predictions, bottom: white matter predictions, left: HAM-D predictions, right: MADRS predictions).

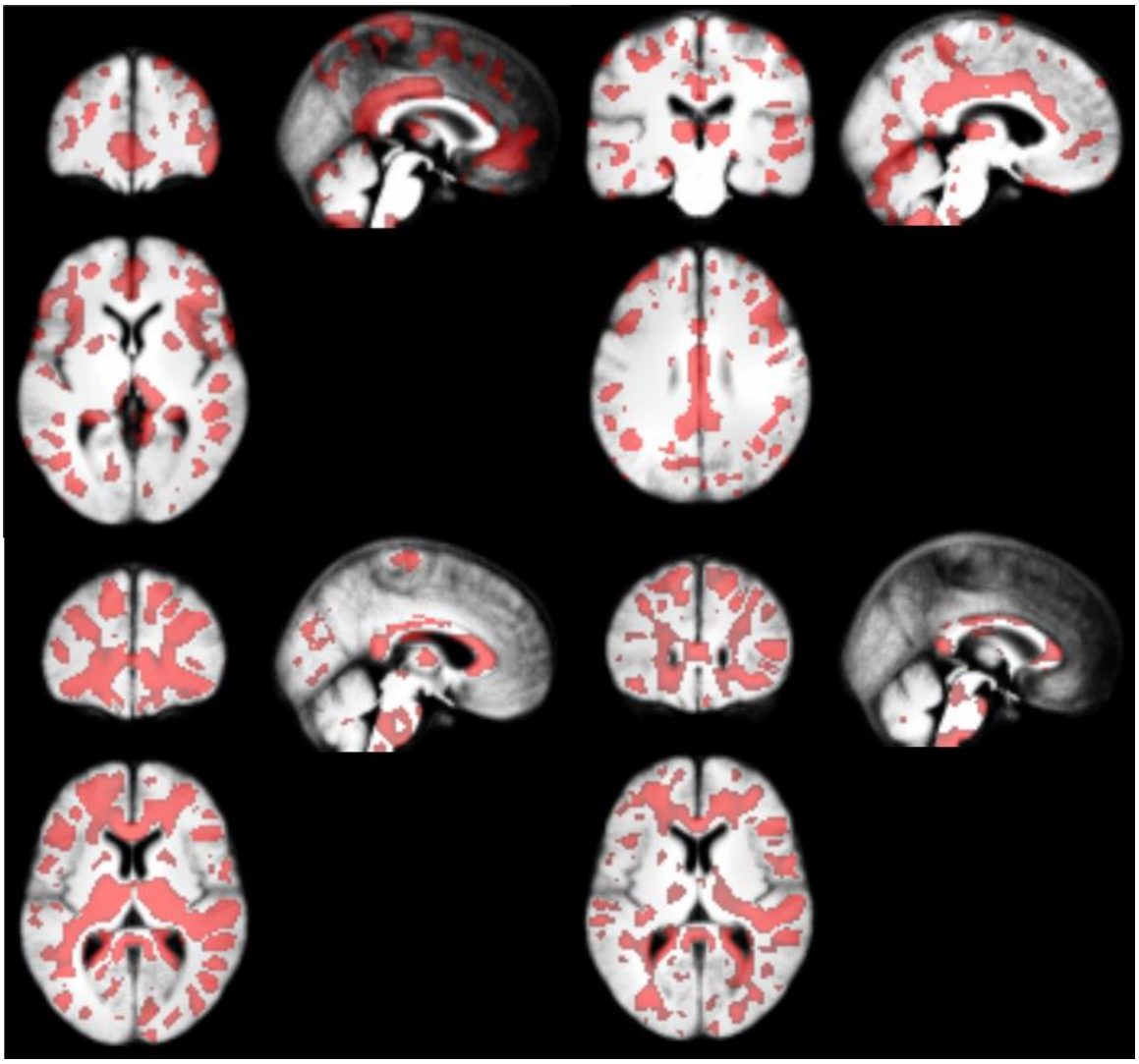

Figure 41: The brain regions which were identified as the most predictive during the whole brain severity score predictions (top: grey matter predictions, bottom: white matter predictions, left: HAM-D predictions, right: MADRS predictions).

7.3.6 Individual Patient RVR Severity Score Predictions using Feature Selection

Using feature selection to attempt to improve symptom severity score prediction over the whole brain predictions produced mixed results. When using thresholded multiple linear regression for feature selection, only the predicted BDI scores were significantly correlated with their corresponding true scores when using grey matter and a linear kernel (RMSE = 9.9218, MAE = 7.6304, R = 0.47901, p = 0.0326, shown in Figure 42).

However, when using white matter images, the BDI score, again, could not be significantly predicted, but the HAM-D and MADRS scores were. Again, using thresholded multiple linear regression for feature selection, white matter images and a linear kernel, the HAM-D (RMSE = 4.545, MAE = 3.9984, R = 0.55618, p = 0.0109) and MADRS (RMSE = 5.7784, MAE = 4.663, R = 0.68735, p = 0.0008) scores could be significantly predicted. The best fitting lines for each set of predictions is shown in Figure 43.

Although these findings were significant when using thresholded multiple linear regression to perform feature selection, the linear trends identified were not quite as impressive as those shown in Figure 40, specifically, the white matter-based predictions without feature selection, and the level of noise in the prediction (as assessed by the RMSE and MAE values) was generally higher than those obtained using the whole brain calculations. Furthermore, the number of voxels and the brain regions identified in the predictions were too sparse and inconsistent to be confident in these results. Therefore, another feature selection approach was investigated to see if it could provide more reliable findings, RFE. As described in Chapter 2, the main issue when using RFE for feature selection is overfitting. This issue was controlled by reducing the number of folds during the inner *N*-fold cross-validation process which optimises the RVR and feature selection parameters. All RFE results described below used 3-fold cross-validation during optimisation of the training data and LOOCV on the outer cross-validation loop as this does not affect overfitting and maximises the data available to the training set.

Using RFE on grey matter images provided the opposite result to those found using thresholded multiple linear regression and a linear kernel, namely, the BDI prediction was not significant and the HAM-D (RMSE = 4.5583, MAE = 3.6615, R

= 0.55464, p = 0.0111) and MADRS (RMSE = 6.6423, MAE = 5.9254, R = 0.54238, p = 0.0135) predictions were.

The results are similar when using RFE and a linear kernel with white matter images, significant predictions are obtained for the BDI (RMSE = 9.5092, MAE = 8.2958, R = 0.55068, p = 0.0119), HAM-D (RMSE = 4.7685, MAE = 3.6170, R = 0.49548, p = 0.0263) and MADRS (RMSE = 6.5554, MAE = 5.8144, R = 0.56082, p = 0.0101). However, like thresholded multiple linear regression, the linear trends identified were not as compelling as the whole brain results and the average RMSE and MAE values are increased when using RFE when compared with the mean whole brain RMSE and MAE values.

More promising results were obtained when a non-linear kernel, such as a Gaussian or RBF kernel, was used. When predicting HAM-D scores, both the Gaussian (RMSE = 3.5694, MAE = 2.7241, R = 0.76721, $p < 0.0001$) and the RBF (RMSE = 3.5715, MAE = 2.731, R = 0.764, $p < 0.0001$) kernels achieved a highly significant correlation between true and prediction scores, using RFE and white matter images. Furthermore, as seen in Figure 44, there is a clear linear trend between the scores and the RMSE and MAE scores are lower than the previous results using feature selection, demonstrating a better fit. The regions identified using RFE and non-linear kernels are shown in Figure 45. These results are less impressive for the MADRS predictions using a Gaussian (RMSE = 6.1539, MAE = 4.635, R = 0.61211, p = 0.0041) and RBF (RMSE = 6.2768, MAE = 4.7325, R = 0.59312, p = 0.0058) kernel but are still very significantly correlated.

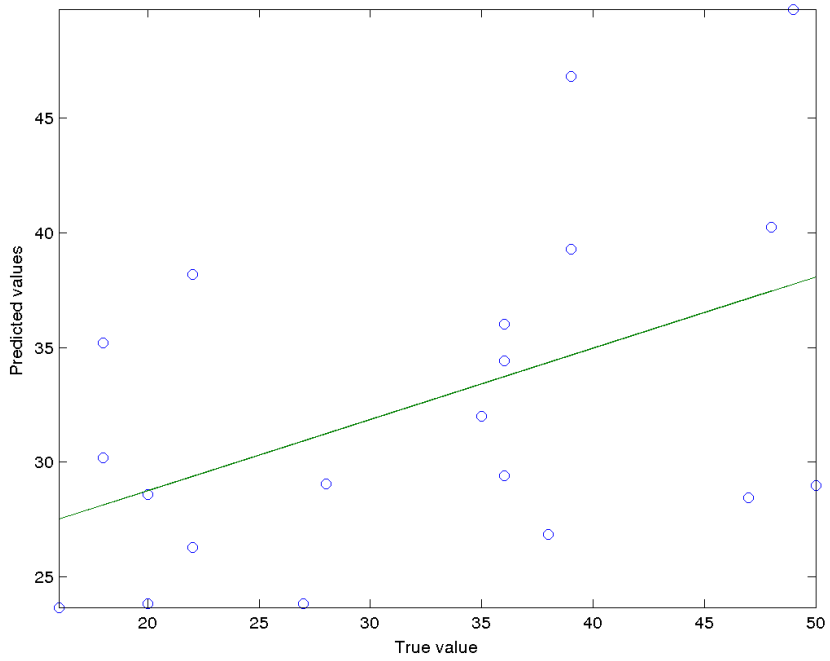

Figure 42: The best fit line for the prediction of the BDI score using thresholded multiple linear regression feature selection and grey matter images.

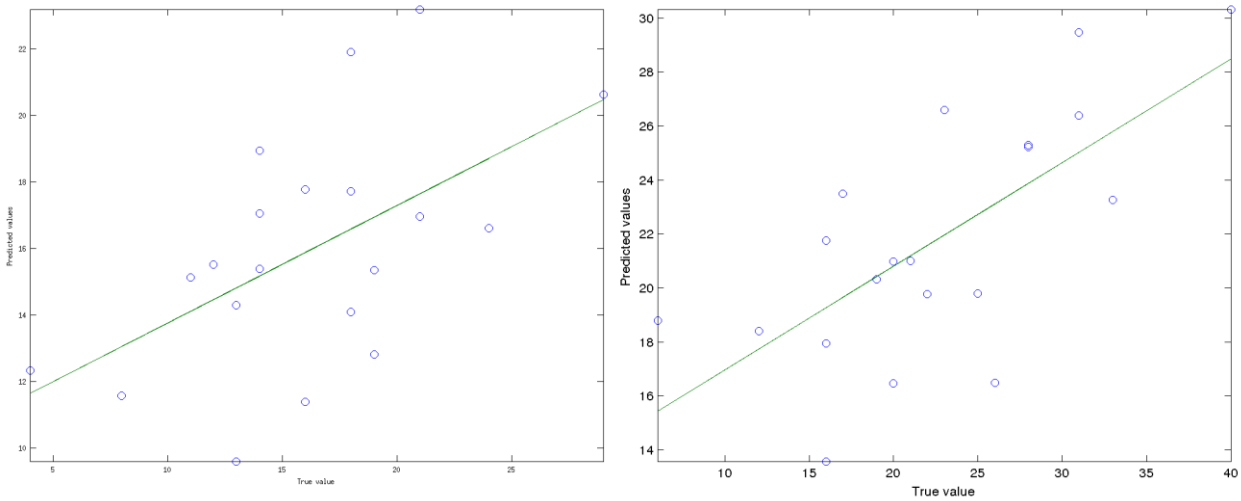

Figure 43: The best fit lines for thresholded multiple linear regression-based white matter severity score predictions (left: HAM-D prediction, right: MADRS prediction).

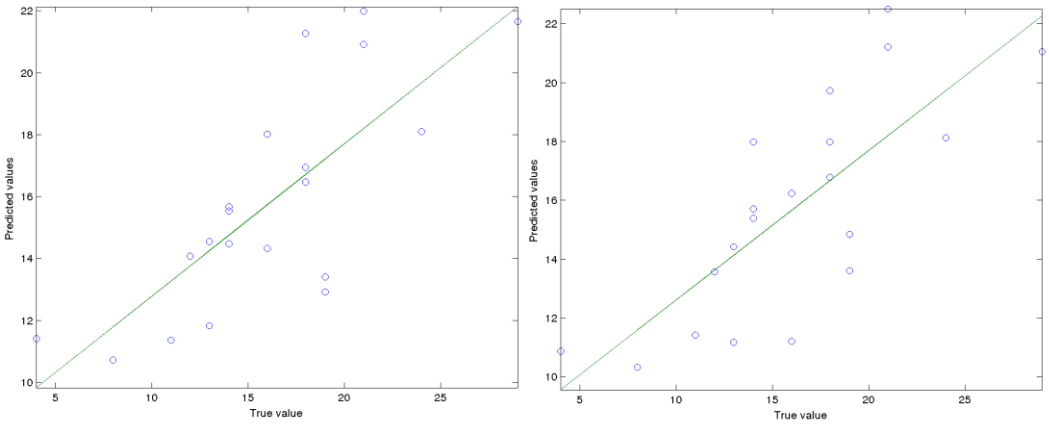

Figure 44: The best fit lines for RFE-based white matter HAM-D predictions (left: Gaussian kernel, right: RBF kernel).

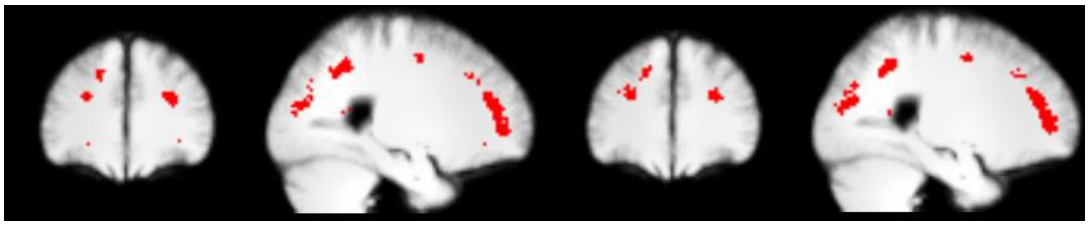

Figure 45: The regions identified using RFE-based feature selection on a non-linear kernel and white matter images to predict HAM-D scores (left: Gaussian kernel, right: RBF kernel).

7.3.7 VBM Analysis (Multiple Linear Regression)

Multiple linear regressions were performed on patients' grey and white matter images to see which regions positively or negatively correlate with symptom severity scores. As higher scores in each of these symptom severity scores indicate more severe symptoms of depression, it would be expected that negative correlations (whereby more severely depressed patients have less grey or white matter volume) would be more likely than positive correlations given the majority of the between group differences in this study and in the literature identify volume reductions compared with a control group. All results shown are $p<0.05$, whole brain level significance.

The MADRS and HAM-D regressions gave similar results for both grey and white matter. Increased grey matter with increasing severity scores were found in the posterior cingulate gyrus and thalamus (shown in Figure 46). Also the anterior cingulate gyrus and basal ganglia were identified as having increased grey matter with increased MADRS score (Figure 46). Decreases were found in the hippocampus, medial orbitofrontal cortex and periventricular grey matter (shown in Figure 47 for HAM-D and Figure 48 for MADRS).

Increased white matter with increasing severity scores were found in the posterior corpus callosum and medial corpus callosum (extending bilaterally as clearly shown in the MADRS calculation, Figure 49 (right)). Negative correlations were found in periventricular white matter, posterior brainstem, white matter deep to the putamen, various areas within the frontal lobe and for the HAM-D regression - the cingulate sulcus. These results are shown in Figure 49 and Figure 50.

The grey matter correlation with the BDI scores identified increased grey matter volume in the cingulate sulcus and the lateral orbitofrontal area. The BDI regression showed that increased BDI severity scores correlated with reduced white matter volume (negative correlation) throughout the frontal lobe, the anterior corpus callosum, white matter deep to the basal ganglia, white matter deep to the thalamus and white matter deep to the ventral tegmental area. The significant regions from the negative correlations for grey and white matter are shown in Figure 51 and Figure 52 respectively. Only a small region in the parietal lobe was found to have increased white matter with increased BDI score.

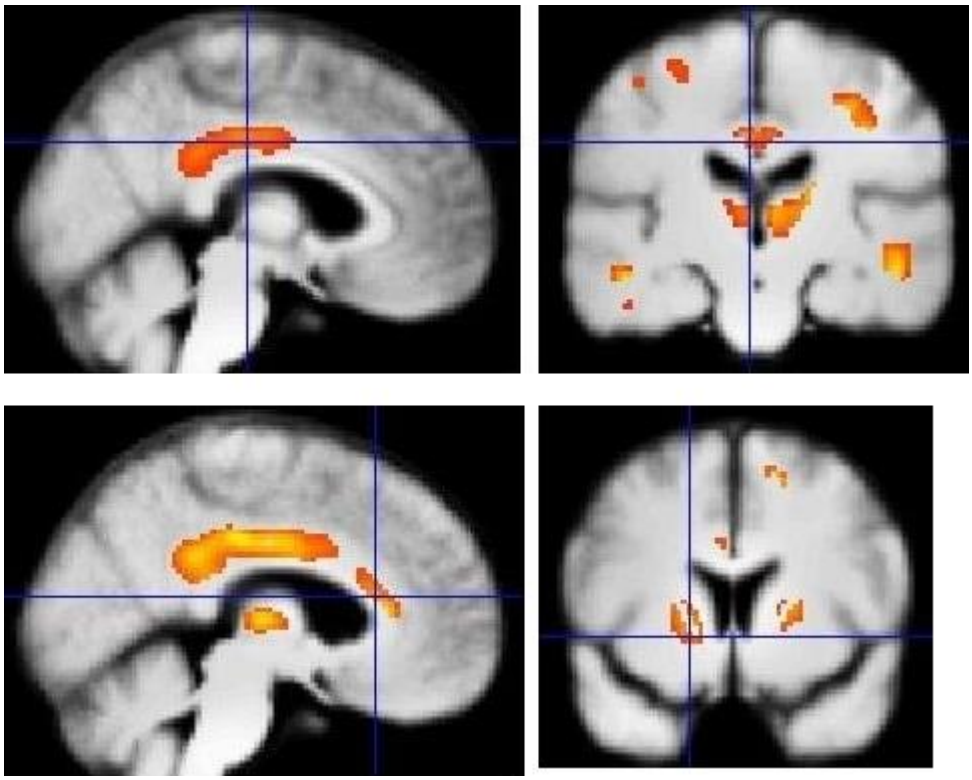

Figure 46: Group-level positive correlations between grey matter volume and HAM-D (top) and MADRS (bottom).

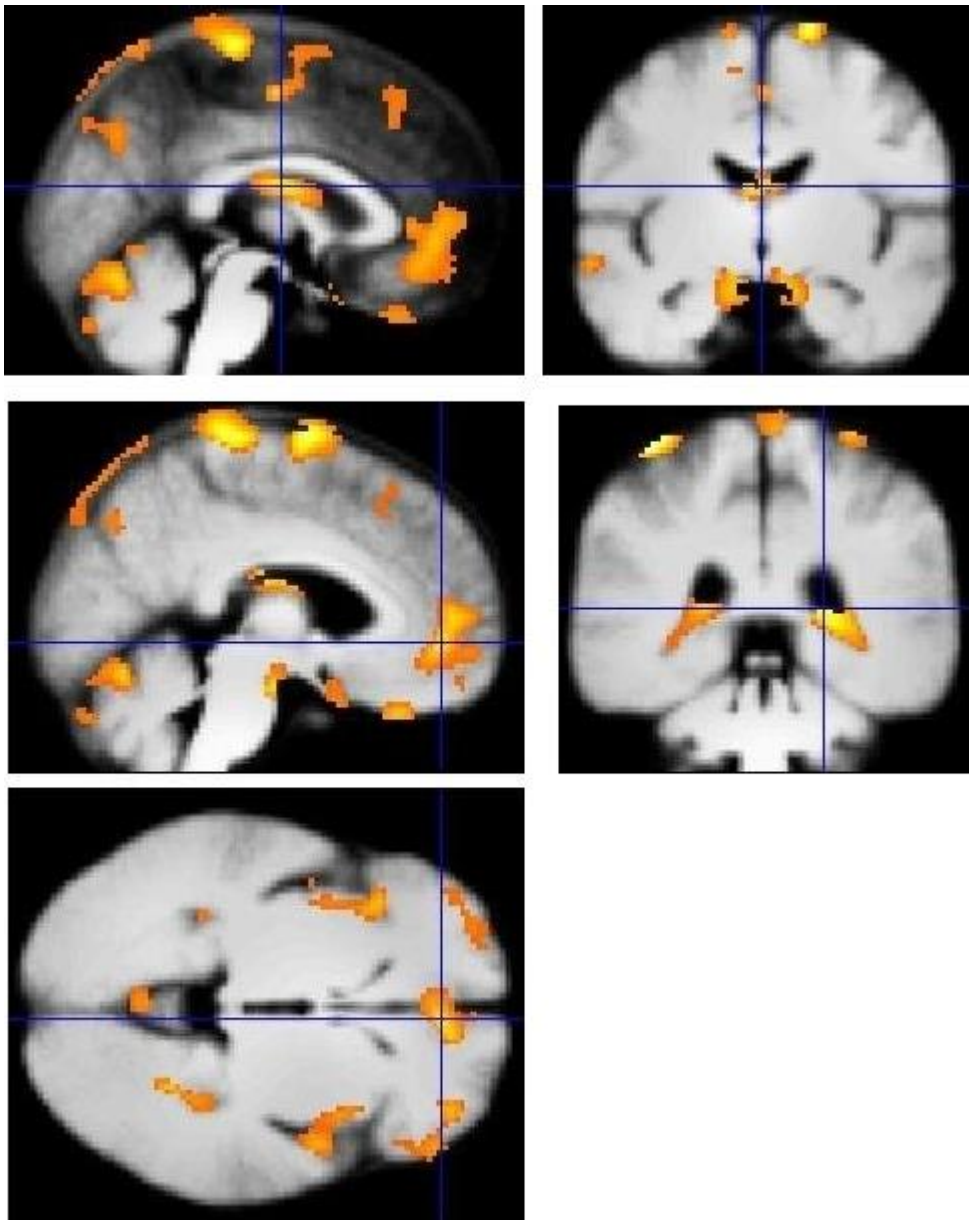

Figure 47: Group-level negative correlations between grey matter volume and HAM-D.

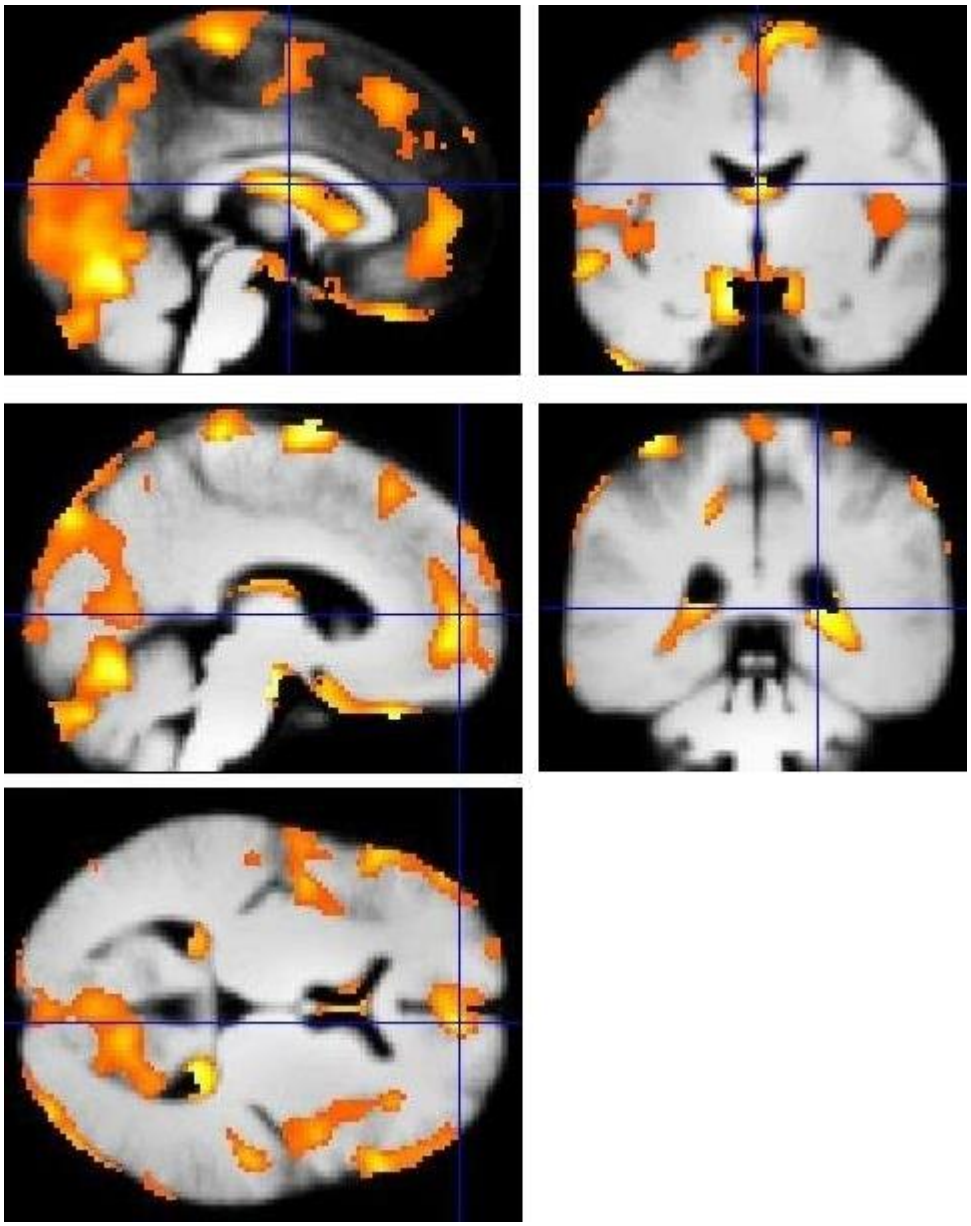

Figure 48: Group-level negative correlations between grey matter volume and MADRS.

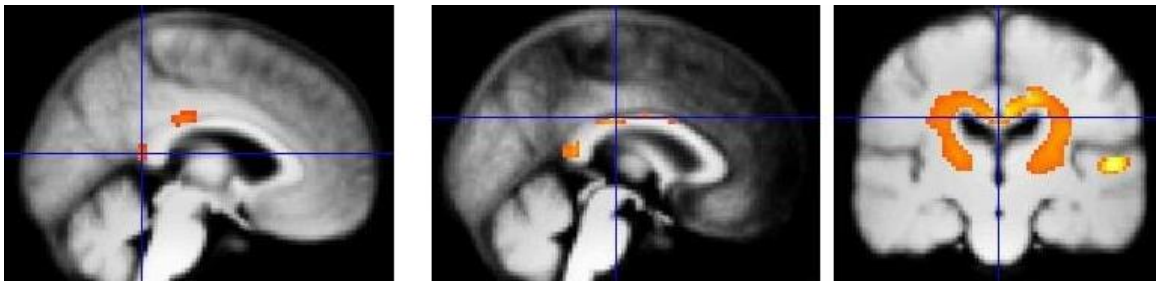

Figure 49: Group-level positive correlations between white matter volume and HAM-D (left) and MADRS (centre and right).

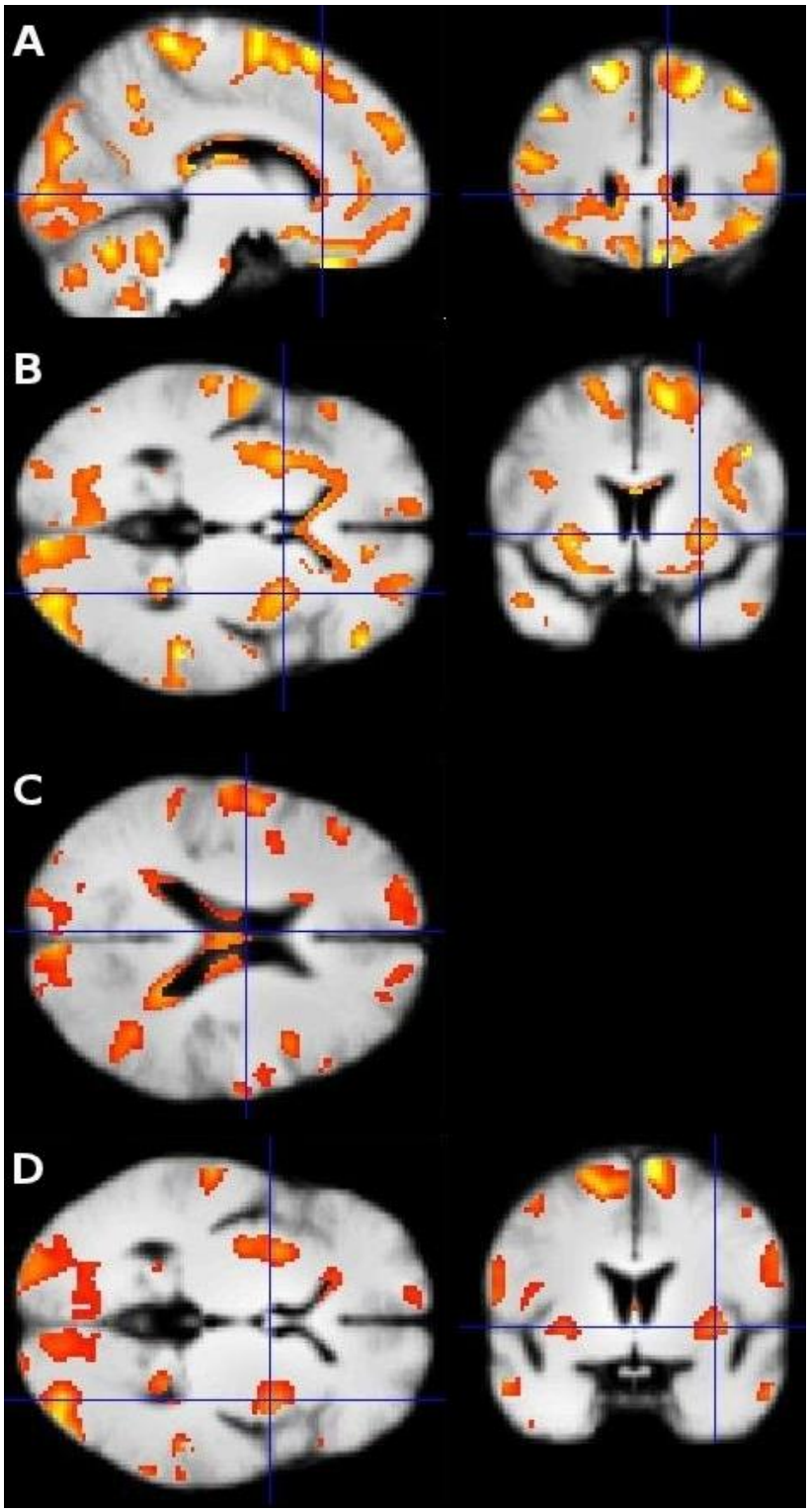

Figure 50: Group-level negative correlations between white matter volume and HAM-D (A and B) and MADRS (C and D).

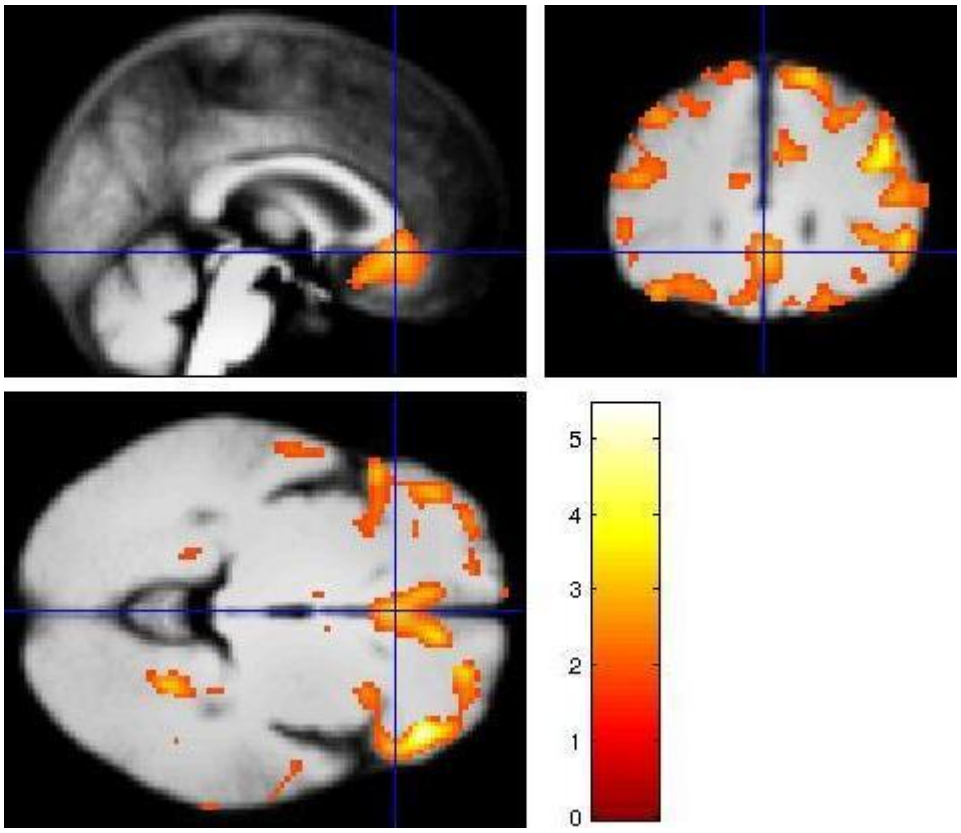

Figure 51: *Group* level *grey* matter volume *decreases* in patients with MDD with increased BDI scores.

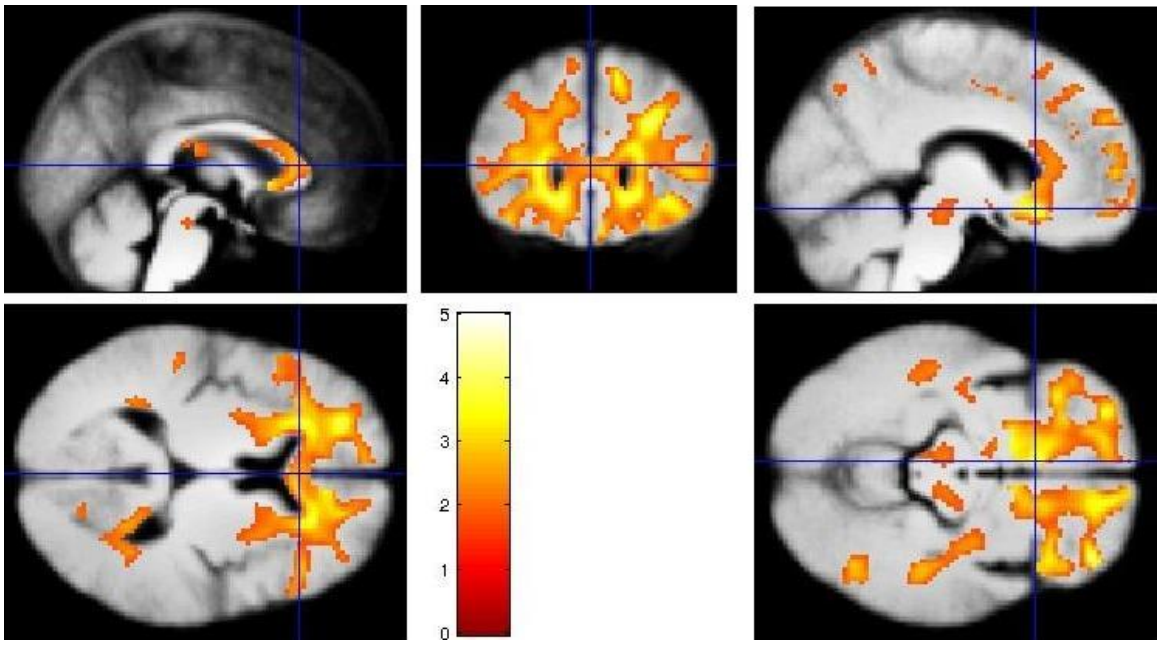

Figure 52: *Group* level *white* matter volume *decreases* in patients with MDD with increased BDI scores.

7.4 Discussion

The results show that when distinguishing MDD subjects and controls, grey matter may be more informative than white matter as it achieved the higher classification accuracy of 85% (compared with 71%). However, white matter was found to be more predictive of current symptom severity, both with and without feature selection, as assessed through the BDI, HAM-D and MADRS scores.

Achieving a high classification accuracy using grey matter images is unsurprising as it has been reported in the literature at a comparable level (Costafreda *et al.,* 2009a; Kipli *et al.,* 2013; Mwangi *et al.,* 2012a; Termenon *et al.,* 2013). Although the white matter prediction achieved a lower significance than the grey matter, the result was still significant and it is the first study which has demonstrated that there are consistent differences between the white matter in MDD subjects and healthy controls which can distinguish subjects on an individual level with significant accuracy.

The grey matter classification was driven by brain regions identified using feature selection. These brain regions included the caudate, insula and periventricular grey matter – all of which have been reported to be decreased in previous group level meta-analyses (Bora *et al.,* 2012; Fu *et al.,* 2003; Kempton *et al.,* 2011; Koolschijn *et al.,* 2009).

All DTI comparisons between MDD and controls found no FA increases MDD subjects (Cole *et al.,* 2012; Kieseppä *et al.,* 2010; Korgaonkar *et al.,* 2011; Li *et al.,* 2007; Ma *et al.,* 2007; Steele *et al.,* 2005; Zhu *et al.,* 2011; Zou *et al.,* 2008) but the largest region identified in both the classification and VBM analyses was an increase of white matter in the cingulate gyrus. The cause of this increase is unclear; it is unlikely to be due to artefact as it would have to be a consistent artefact in one group compared to the other for the region to be included in the classification. The unexpected increase in white matter is unlikely to be due to outliers for the same reason. A limitation of this study is that a number of the subjects were taking anti-depressant medication, however it is unclear what effect this would have on white matter as there have been no studies investigating this. Although psychiatric neuroimaging studies typically report reductions in volume, increases in white matter volume have been reported in other disorders such as ASD (Herbert *et al.*, 2003; Herbert *et al.,* 2004), treatment-naïve obsessive compulsive disorder (Atmaca *et al.,*

2007), body dysmorphic disorder (Rauch *et al.,* 2003) and schizophrenia (Suzuki *et al.,* 2002).

There are two different potential neurobiological explanations for this increased cingulate gyrus region. The simplest explanation is that there is a genuine increase of white matter in patients compared with controls in the cingulate gyrus. Although this might explain why a lesion in this area during a cingulotomy could alleviate symptoms, it is still unclear why this would not be reported in previous DTI studies. An alternative explanation is that modulation, which takes place during pre-processing, increases the intensity of tissue in brain regions when the volume is decreased (and vice-versa) when normalising towards a template as modulation attempts to preserve the total amount of white matter across the whole brain. Therefore if the white matter volume around the cingulate gyrus was consistently larger in patients (in native space), then the intensity would be consistently increased in the same region after modulation. Increased white matter volume in the cingulate gyrus is a far more likely neurobiological explanation as it could be due to a number of things such as insufficient synaptic pruning during adolescence (Paus *et al.,* 2008) or, more likely, a decrease in white matter integrity in the cingulate gyrus (Bennett *et al.,* 2010). If white matter integrity was reduced in the cingulate gyrus then MD would be increased and/or FA would be reduced. In the DTI studies of MDD FA decreases have been reported in the posterior cingulate cortex and corpus callosum (Cole *et al.,* 2012; Zhu *et al.,* 2011) and the only study which reported MD differences found an increase in the corpus callosum (Cole *et al.,* 2012) suggesting that white matter integrity was reduced around the cingulate gyrus. A reduction in white matter integrity may be due to cellular differences (such as the density of axons, the level of myelination, axonal diameter and inflammation) or larger, voxel-sized differences (such as a decrease in the directional organisation of fibres within a bundle), it is possible that the increased white matter volume finding in the cingulate gyrus was identified due to a combination of these factors (Bennett *et al.,* 2010).

As this is the first comparison of the white matter (T_1-weighted) images from MDD patients and controls, this finding requires replication. However, as all subjects in this study have DTI data, further investigation can take place to see if the cingulate gyrus differences are observable using DTI analyses and if further support for either of the arguments for a neurobiological difference between the groups can be acquired.

Mwangi *et al.* (2012b) found that they were able to predict the BDI score using whole brain grey matter images but not the HAM-D score. However, when using a similar approach, this study was able to predict HAM-D (and MADRS which was not discussed by Mwangi *et al.*) but not BDI. This discrepancy may be due to the fact that the present study had a wider range of symptom severity within the MDD group and/or because it is a single centre study. In any event, this work requires study of a larger dataset to investigate this issue further. The HAM-D and MADRS scores were both able to be significantly predicted based on both grey and white matter whole brain images, with the white matter images providing the best predictions of severity scores, whereas the BDI score could not be significantly predicted.

Adding feature selection to the process did not dramatically improve the results. Using univariate feature selection, the grey matter images were able to predict the BDI scores but could no longer predict the HAM-D and MADRS scores significantly and the white matter results gave a poorer fit compared with the whole brain white matter results. Multivariate feature selection (RFE) improved the significance level of the grey matter MADRS and HAM-D predictions (compared with the whole brain results) and all three variables were significantly predicted using white matter images and RFE. However, there is a concern that, although the predicted values correlate with the true scores, the mean absolute error and root mean squared error values are higher than the corresponding whole brain, meaning that the prediction contains more noise. One particularly interesting result was the addition of non-linear kernels to RFE predictions. Although non-linear kernels did not improve the results using whole brain images or univariate feature selection-based predictions, both the Gaussian and RBF kernels dramatically improved the prediction of the HAM-D and, to a lesser extent, MADRS scores when prediction was based on multivariate feature selection (RFE) and white matter images.

There are a number of limitations of this study. First, although the number of subjects in this study is comparable with similar studies, the results require replication in a larger study. Second, patients were taking a range of antidepressant medications at the time of scanning; however it is unclear to what extent this had an effect on these results. Sapolsky (2001) has argued that it is unlikely that grey matter volume reductions in MDD are as a result of antidepressant medication as there is evidence for antidepressant-induced neurogenesis with no arguments for reductions,

however, the small amount of increased grey matter volume may be a medication effect which needs further investigation. As there are no studies discussing the effects of antidepressant medication on white matter, it is unclear what effect it may have. Current medication status is also a potential confound when predicting symptom severity scores, however, Table 9 shows that there is no obvious link between current medication and symptom severity. Finally, as the MDD group were recruited with a past or present diagnosis of MDD the differences between the two groups may not have been as distinct as other studies, however, this range of symptom severity may have been an advantage in the prediction of symptom severity.

To summarise, it was possible to replicate the accurate classification of grey matter images to distinguish MDD subjects and healthy controls (Costafreda *et al.*, 2009a; Kipli *et al.*, 2013; Mwangi *et al.*, 2012a; Termenon *et al.*, 2013) and it was possible to extend this to accurately predicting diagnosis using the white matter component of structural MR images. It was possible to replicate the prediction of symptom severity using whole brain structural MRI (Mwangi *et al.*, 2012b), however the severity scores which were able to be significantly predicted using the whole brain images were the HAM-D and MADRS scores but not the BDI score, contrary to the findings by Mwangi *et al.* (2012b). Furthermore, white matter tends to have an improved accuracy of prediction of symptom severity compared with grey matter. Although these results require replication in a larger population, these results provide encouragement that machine learning methods can increase the understanding of the neurobiology of MDD.

Table 9: Current Medication and State Illness Severity. No patients had psychotic symptoms and quetiapine was prescribed as an augmentation agent for antidepressants (Dorée *et al.*, 2007), similar to the long established use of lithium, L-tryptophan and triiodothyronine in treatment resistant depression. No obvious relationship between current medication and state illness severity was present. 'mg' indicates total dose per day, 'mcg' total micrograms per day.

	Primary AD	Secondary AD	Augmentation	Anti-psychotics	HAM-D
1	fluoxetine (100 mg)	trazodone (150 mg)			21
2	venalfaxine (525 mg)	trazodone (150 mg)			18
3	isocarboxazid (40 mg)			quetiapine (75 mg)	19
4	venlafaxine (300 mg)		lithium (200 mg)		8
5	sertraline (100 mg)	trazodone (200 mg)		quetiapine (300 mg)	11
6	sertraline (300 mg)	trazodone (300 mg)	triiodothyronine (20 mcg)	quetiapine (800 mg)	24
7	isocarboxazid (70 mg)				18
8	venalfaxine (225 mg)				18
9	sertraline (100 mg)			quetiapine (100 mg)	13
10	fluoxetine (60 mg)	mirtazapine (45 mg)	lithium (900 mg)		21
11				chlorpromazine (150 mg)	29
12	sertraline (200 mg)			quetiapine (300 mg)	16

13	tranylcyprom ine (70 mg)				14
14	venlafaxine (300 mg)		L-Tryptophan (6000 mg)		19
15	venlafaxine (525 mg)	mirtazapine (45 mg)			4
16	citalopram (60 mg)				14
17	venlafaxine (75 mg)				12
18	citalopram (10 mg)				13
19	phenelzine (60 mg)		L-Tryptophan (3000 mg), lithium (1000 mg)	quetiapine (75 mg)	16
20	venalfaxine (300 mg)			quetiapine (200 mg)	14

Chapter 8: High Accuracy Individual Diagnostic Classification in MDD using fMRI

8.1 Introduction

As anhedonia is one of the main symptoms of MDD, a blunted response to rewarding events is anticipated in patients and has been reported in the literature (Zhang *et al.,* 2013). A recent meta-analysis, investigating reward processing in MDD, found that decreased activity in the caudate, cerebellum, thalamus, anterior cingulate, putamen and insula and increased activity in the cuneus, middle frontal gyrus, superior frontal and fusiform gyrus, frontal lobe and lingual gyrus, were present in all types of rewarding stimuli investigated (Zhang *et al.,* 2013). When investigating reward processing specifically associated with monetary rewards, the caudate, thalamus, insula and precuneus were found to have decreased activity and the inferior, middle and superior frontal gyrus, inferior parietal lobule, cuneus and anterior cingulate showed increased activity in MDD (Zhang *et al.,* 2013). As the current study investigates brain activity at the outcome time, the most relevant finding by Zhang and colleagues (2013) was that they only identified decreased activation in the caudate in MDD during the analysis of monetary reward activations at the outcome time.

The monetary incentive delay task consists of rewarding (win) and aversive (loss) events. At the feedback time, Knutson *et al.* (2008) found that rewarding events activated the medial prefrontal cortex, posterior cingulate cortex, caudate and hippocampus in both the unmedicated MDD group and the healthy control group. Also the putamen and sublenticular extended amygdala were activated in the controls, but not the MDD group (Knutson *et al.,* 2008). When the avoiding a loss vs. receiving a loss contrast was investigated, the middle frontal gyri, parietal cortex, sublenticular extended amygdala and putamen were activated in the control group, whereas the MDD group only showed activation in the head of the caudate (Knutson *et al.,* 2008). In a similar study, Pizzagalli *et al.* (2009) found that MDD subjects had significantly weaker activations in the nucleus accumbens and dorsal caudate during rewarding feedback.

In another reward-related task, MDD patients were found to have greater activation in the inferior frontal gyrus and thalamus when receiving a reward

(Smoski *et al.,* 2009). Also, failure to win resulted in greater activation in the caudate, auditory cortex, BA 41, occipital regions and frontal medial cortex in controls and in the MDD group greater activations were found in the middle, inferior, and orbitofrontal cortex (Smoski *et al.,* 2009).

In general, the most consistent finding in reward-based tasks at the outcome time is that MDD patients tend to have a decreased activation in the caudate and this was the only region identified in the meta-analysis by Zhang *et al.* (2013). The response to aversive events in MDD, however, is unclear as no studies investigating losing compared with avoiding a loss (rather than avoiding a loss compared with losing) could be identified.

There are a number of studies that have attempted to predictively classify MDD patients vs. controls, however, only one study has applied machine learning to a reward or loss based paradigm. Hahn *et al.* (2011) proposed that combining contrasts from three different fMRI paradigms, an emotional processing paradigm and two modified monetary incentive delay tasks, involving an attempt to win money and avoid losing as much as possible, would improve classification accuracy more than could be achieved for an individual contrast. When attempting to classify fMRI contrasts individually, Hahn *et al.* found that the highest accuracy (72%) was obtained when subjects anticipated avoiding a loss (Hahn *et al.,* 2011). However, when 3 of the 15 conditions were combined, the accuracy increased to 83% (sensitivity – 80%, specificity – 87%). The three conditions which led to the highest accuracy were neutral facial expressions, actual large reward and anticipation of no loss (Hahn *et al.,* 2011). As two of the three fMRI contrasts found to be relevant to distinguishing depression from healthy controls by Hahn *et al.* (2011) were related to their reward and loss task, the application of the reward and loss paradigm in this chapter, may add to the literature.

The study presented in this chapter used the reward and loss fMRI paradigm outlined in Chapter 6 to investigate neural responses in MDD and healthy controls. It is an instrumental task which requires the acquisition of learning to win points and to avoid losing points by choosing one of two pairs of stimuli and receiving feedback on the outcome of the choice. Control 'baseline' stimuli were also present. The rewarding and aversive events were analysed separately. Each contrast was investigated to determine whether brain activity patterns could accurately classify MDD and healthy controls on an individual level, using a similar approach to the one

that was used to classify structural MRI. In addition, possible correlations between symptom severity and brain activity were tested.

8.2 Method

8.2.1 Subjects

Event-related fMRI scans were obtained from subjects participating in neuroimaging studies at the Clinical Research Centre, Ninewells Hospital and Medical School in Dundee, UK. Informed consent was obtained from all volunteers. The study was approved by the local Ethics Committee.

Twenty adults with a past or present diagnosis of MDD were recruited from the Advanced Interventions Service in Dundee. One scan was excluded from the analysis due to a failure to adequately perform the fMRI task.

Diagnosis was made according to MINI PLUS V5.0 criteria (Sheehan and Lecrubier, 1992). Exclusion criteria included potentially confounding diagnoses – any other primary psychiatric disorder, substance misuse or significant head injury. 18 MDD participants were being treated with one or more anti-depressant medication (venalfaxine (6), sertraline (3), trazodone (3), citalopram (2), fluoxetine (2), mirtazapine (2), isocarboxazid (1), ltryptophan (1), phenelzine (1), and tranylcypromine (1)). In addition, 7 MDD participants were being treated with anti-psychotic medications (quetiapine (6) and chlorpromazine (1)) and 3 MDD participants were being given lithium augmentation.

Twenty-one healthy controls with no lifetime history of MDD were recruited mostly from partners, relatives and friends of patients and underwent psychiatric screening using the same semi-structured interview schedule as the patients. None of the controls had a history of current or past psychiatric or neurological disorder and none were taking medication. To balance the two groups in terms of total subjects, two randomly selected controls were removed from the dataset during machine learning.

All MDD and control volunteers had a predicted pre-morbid Full Scale Intelligence Quotient above 106 as assessed by the NART. Handedness was assessed using the EHI (Oldfield, 1971). Apart from 2 left-handed subjects in the control group and 1 and 3 ambidextrous subjects in the control and patient groups

respectively, all subjects were right-handed (with one patient having failed to complete the EHI test).

8.2.2 Image Acquisition

For each participant functional whole-brain images were acquired using a 3T Siemens Magnetom TrioTim syngo scanner using an EPI (echo-planar imaging) sequence with the following parameters: TR = 2500 ms, TE = 30 ms, flip angle = 90°, FOV = 224 mm, matrix = 64 x 64, 37 slices, voxel size 3.5x3.5x3.5 mm, slice thickness 3.5 mm, inter-slice gap = 0.5 mm. The first four BOLD (Blood-oxygen-level dependent) volumes were discarded as standard.

The fMRI paradigm was a modified version of the Pessiglione task (Pessiglione *et al.,* 2006), incorporating rewarding, neutral and aversive events into one task. It has been described fully in Chapter 6.

8.2.3 Image Pre-processing

All scans were visually inspected for artefacts and particular care was taken to identify gross artefacts (McRobbie *et al.,* 2010). A small number of individual volumes showed gross artefacts, all of which were due to head movement. The affected volumes were replaced by the average of the two neighbouring volumes.

Pre-processing was done using SPM8 (http://www.fil.ion.ucl.ac.uk/spm). Images were first realigned towards the first image in each time series and co-registered to the SPM8 MNI EPI template. The average realigned, co-registered images for each subject were used as a template to normalise each realigned and co-registered volume to the SPM8 EPI template image. The resultant images were smoothed with an 8 mm FWHM Gaussian kernel.

8.2.4 fMRI Analyses

An event related random effects design was used for analysis. The times of each type of feedback (reward pair: win/nothing, loss pair: nothing/lose, and neutral pair: no change/nothing) were modelled as truncated delta functions and convolved with the BOLD function. These six vectors were entered into standard first level analyses for

each subject. The six motion realignment parameters, as output by SPM, were entered into the design matrix as covariates to eliminate motion driven artefacts. Four contrasts were defined during the first level analysis. The first two were winning vs. failing to win (where 'nothing' was presented during the reward pair) and losing vs. avoiding a loss (where 'nothing' was presented during the loss pair). These two contrasts will be shortened for future reference to the basic win and loss contrasts. The other two contrasts involved including the neutral condition to eliminate any non-task-based activations/deactivations. These involved contrasting the basic win and loss contrasts with the basic neutral contrast ('no change' vs. 'nothing'), and will be referred to as the controlled win and loss contrasts hereafter.

Each contrast image was entered into second level analyses to test for within-group (one-sample t-test) activations and between-group differences (MDD vs. control two-sample t-test). Also, the patient group was investigated further by performing a regression with various symptom severity scores, such as the MADRS, the HAM-D, the Hamilton Anxiety Rating Scale (HAM-A) the BDI and the Beck Hopelessness Scale (BHS). Higher scores in each of these scores indicate more severe symptoms.

For the conventional *group* level VBM analysis, the null hypothesis of no difference in brain structure between patients and controls was tested using an unpaired t-test as implemented in SPM8. Significance was defined as $p<0.05$ at a whole brain corrected cluster level (Slotnick *et al.*, 2003).

8.2.5 Individual Scan Classification

Machine learning to generate individual subject predictions was implemented in Matlab (The Mathworks Inc.) using an SVM toolbox (Schwaighofer, 2001) and custom Matlab scripts. SVM analysis consisted of two stages: training the classifier, then testing the accuracy using data not used for training. Standard LOOCV was used for training with the SVM parameters being selected on the basis of training stage accuracy. As in Chapter 7, the feature selection method applied during the training stage was the thresholded t-test, as implemented in the SPM toolbox.

After the first level analyses, four different contrasts were extracted for each subject. These contrasts were classified independently in each case, in the same way as grey and white matter images were independently classified in the previous section.

8.3 Results

8.3.1 Participant Characteristics

Age and IQ did not differ significantly (t-test, $p>0.1$) and gender was not significantly different (as assessed by a chi-square calculation) between groups. The MDD group mean age was 50.8 years (standard deviation 10.6) mean IQ was 122.8 (standard deviation 4.8). The control group mean age was 45.5 years (standard deviation 13.0) and the mean IQ 122.5 (standard deviation 5.8). There were no significant differences in task performance between groups.

The average HAM-D, MADRS and BDI illness severity rating scores in the MDD group were 16.1, 22.5 and 32.2, indicating depression severity in the moderate range. The mean HAMA, BHS and MGH-S scores in the MDD group were 15.8, 14.1 and 13.2, respectively.

These results are outlined in Table 10.

Table 10: Clinical and task peformance descriptors for the MDD and healthy control groups in the fMRI analysis. Variables are shown as mean (standard deviation). *chi-square test with other tests being t-tests.

	MDD	Controls	
Age	50.79 (10.6)	46.14 (13.97)	n.s.
IQ	122.58 (4.78)	116.95 (27.38)	n.s.
Female/Total*	15/19	15/21	n.s.
HAM-D	16.00 (5.72)	0.48 (0.93)	<0.001
MADRS	22.05 (7.93)	0.48 (1.03)	<0.001
BDI	32.42 (11.65)	0.43 (0.87)	<0.001
HAMA	15.84 (5.66)	0.43 (0.98)	<0.001
BHS	14.05 (5.36)	1.43 (1.47)	<0.001
MGH-S	13.24 (10.78)	N/A	N/A
Number of vouchers won	33.58 (4.96)	34.90 (4.35)	n.s.
Number of vouchers lost	29.05 (4.97)	29.10 (3.71)	n.s.
Total task score	4.53 (6.45)	5.81 (5.78)	n.s.

8.3.2 Within-Group and Between-Group Analyses

The basic and controlled contrasts, defined in section 8.2.4, provided similar results; however, the latter gave a clearer output as activations due to visual stimuli in the occipital lobe were removed. Therefore, the results from the basic contrasts are omitted at this stage.

On reward trials, healthy controls were found to activate primarily in the nucleus accumbens, caudate, medial orbitofrontal cortex and the posterior cingulate (Figure 53 (left)). MDD patients exhibited activation in the insula and a weaker signal in the medial orbitofrontal cortex (Figure 53 (right)).

When this contrast was entered into a between-group analysis (patients vs. controls), a significant failure to activate the nucleus accumbens, medial orbitofrontal cortex, posterior cingulate and, to a lesser extent, caudate was identified in the patient group (Figure 54 (left)). Furthermore, the insula activation identified in the within-group patient analysis shows up as *increased* in patients in the between-group analysis (Figure 54 (right)).

On loss trials, healthy controls were found to deactivate in the nucleus accumbens and hippocampus whereas patients only deactivated the nucleus accumbens (Figure 55 (left and right correspondingly)). Whilst the control group did not show any relevant activation during aversive events, the patient group showed *increased* activation in the midbrain and insula (Figure 56).

When comparing the controlled loss contrast between groups, the most significant brain region was the hippocampus (Figure 57), which was as a result of patients' failure to deactivate this region. The insula which patients activated, and the control group did not, was also significantly different between groups (Figure 57).

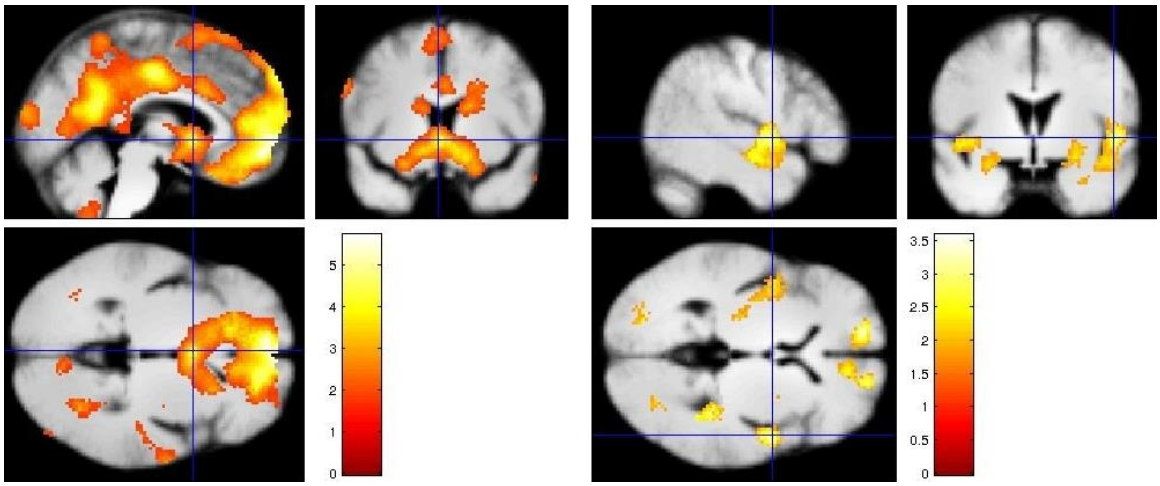

Figure 53: Within-group analyses of the controlled win contrast, displaying activations in controls (left) and patients (right).

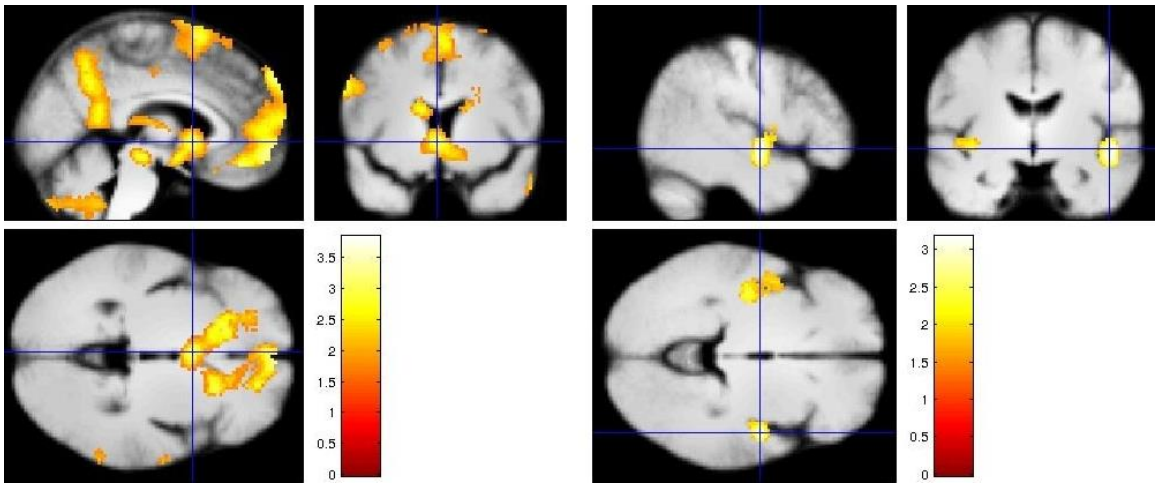

Figure 54: Between-group analysis of the controlled win contrast, displaying regions of increased activation in controls compared with patients (left) and increased activation in patients compared with controls (right).

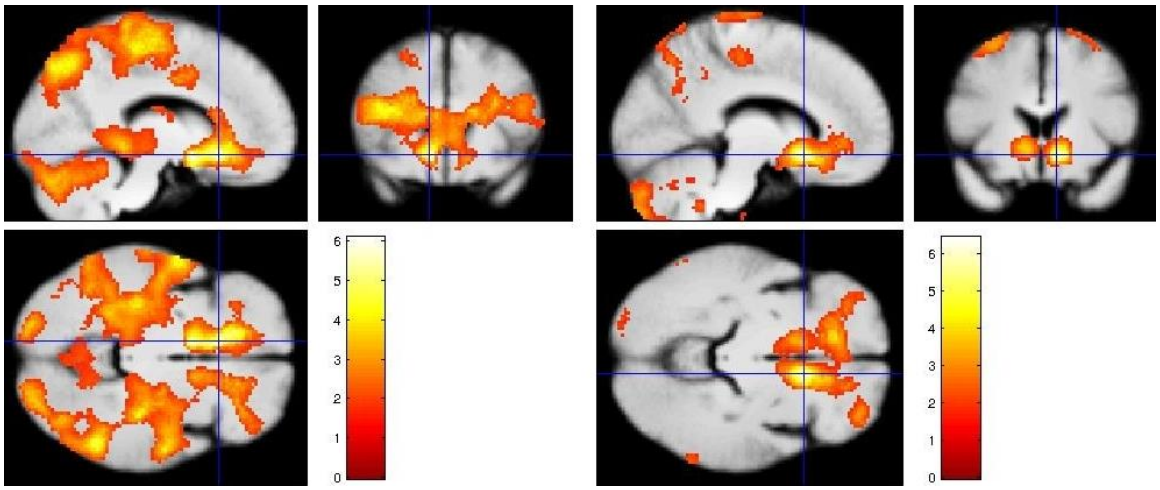

Figure 55: Within-group analyses of the controlled loss contrast, displaying deactivations in controls (left) and patients (right).

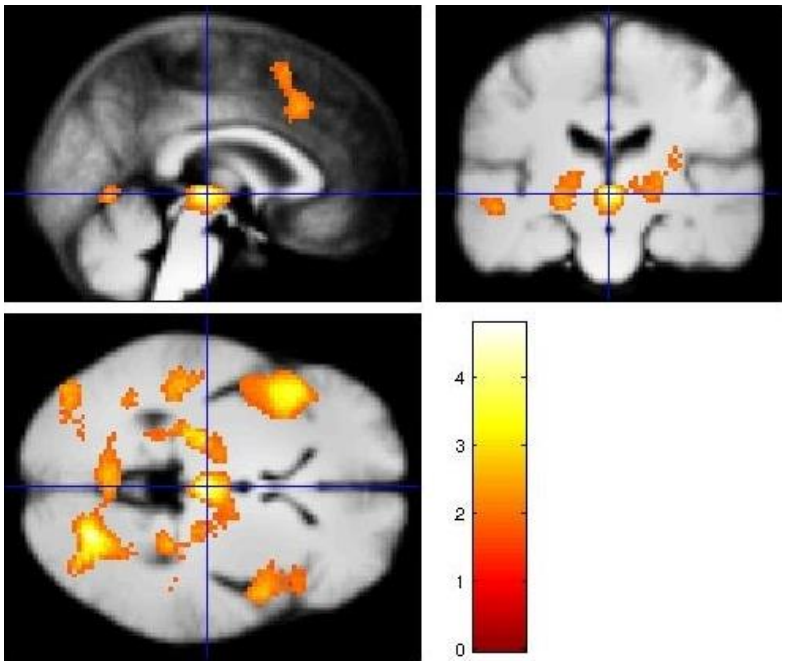

Figure 56: Within-group analyses of the controlled loss contrast, displaying activations in the patient group.

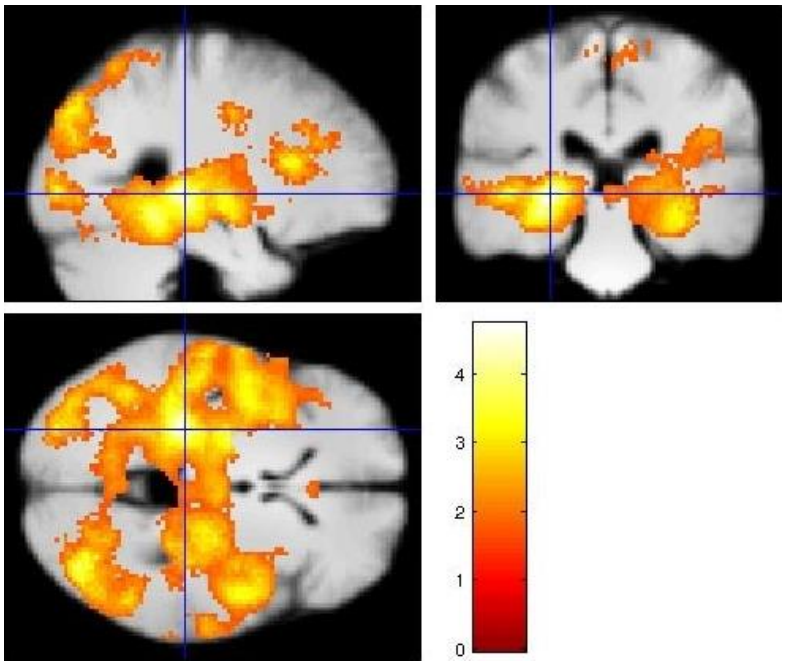

Figure 57: Between-group analysis of the controlled loss contrast, displaying regions of increased activation in patients compared with controls.

8.3.3 Individual Subject SVM Predictions

A Gaussian SVM was used to analyse 19 fMRI images of adults with a past or present diagnosis of MDD and 19 fMRI images of control subjects matched for age, gender and IQ. Feature selection was implemented using t-tests with a variable threshold that was optimised during cross-validation. The analysis was done using the four contrasts separately.

The analysis using the basic win contrast images resulted in an individual subject predictive accuracy of 79% (sensitivity 0.79, specificity 0.79, $\chi^2 = 10.5$, $p = 0.0012$). The basic loss contrast images obtained an accuracy of 84% (sensitivity 0.89, specificity 0.79, $\chi^2 = 15.3$, $p < 0.0001$).

Including the neutral, control condition, increased both accuracies to 84% (sensitivity 0.79, specificity 0.89, $\chi^2 = 15.3$, $p < 0.0001$) for the controlled win contrasts and 97% (sensitivity 0.95, specificity 1.0, $\chi^2 = 30.5$, $p << 0.0001$) for the controlled loss contrast.

8.3.4 Brain Regions identified using Feature Selection

The brain regions identified in the classification of the controlled contrasts were very small. The nucleus accumbens and medial orbitofrontal cortex were identified in the controlled win contrast classification and a single region in the hippocampus was identified in the controlled loss contrast classification. These are shown in Figure 58, with the brain regions identified during feature selection highlighted with respect to the corresponding between-group VBM results for each contrast.

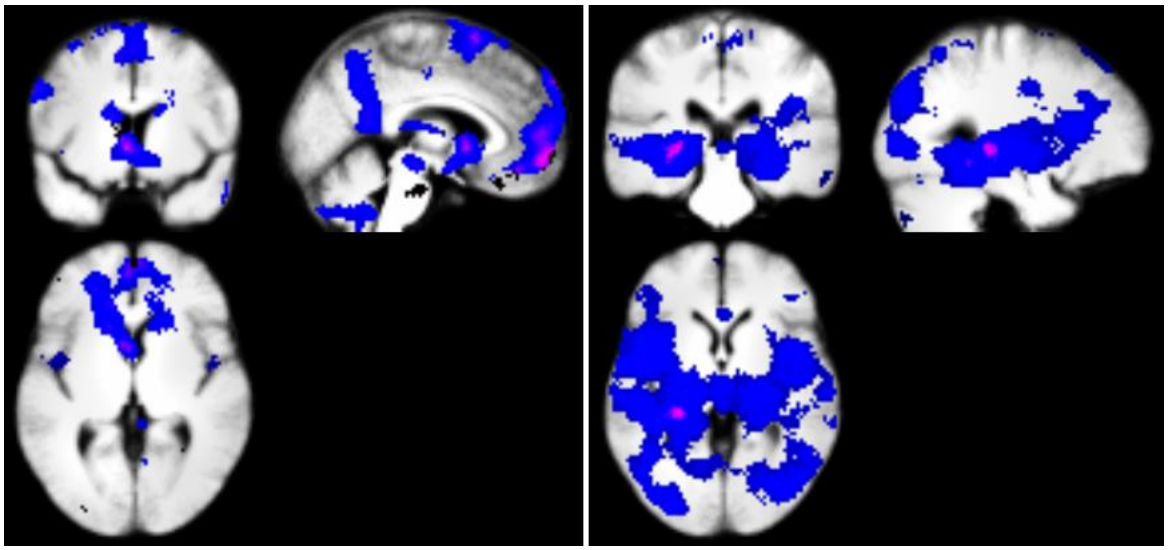

Figure 58: The overlap between the regions identified during VBM analysis (blue and pink) and the brain regions identified during the classification of MDD patients and controls (pink) using a controlled win contrast (left) and a controlled lose contrast (right).

8.3.5 Correlations with Severity Scores

Linear regressions were performed using patients' fMRI contrast images to test which brain activity correlated with symptom severity scores. Whilst the basic and controlled reward contrasts did not reveal any notable regions with respect to severity scores, the basic and controlled loss contrasts correlated more strongly with the severity scores investigated.

The nucleus accumbens activity was found to negatively correlate with the MADRS and HAM-D scores (Figure 59) using the basic loss contrast (but not the controlled loss contrast). This suggests that the more severely depressed patients deactivated the nucleus accumbens more strongly when a loss was experienced as the within-group analysis in the patient group showed that the nucleus accumbens was deactivated during loss trials. However, as the control group also deactivated this region and there were no significant between-group differences in this region, this suggests patients might deactivate the nucleus accumbens more than controls, but not to a significant degree. This could be investigated in future analyses.

The BDI correlated with an increase in activation in the controlled loss contrasts in the insula (Figure 60). This was an expected result as the patient group showed an increased activation in the insula in the within-group analysis which the controls did not, leading to a between-group difference in this region.

A region of the anterior cingulate significantly correlated with the HAD-A score when the controlled loss contrast was investigated (Figure 61). This means that the more anxious patients had more activity in the anterior cingulate while experiencing a loss.

Finally, the BHS positively correlated with a large hippocampal region during loss trials (Figure 62). As reported in section 8.3.2, the hippocampus significantly deactivated in controls but the patient group failed to significantly deactivate the hippocampus. The correlation with the BHS shows that increased feelings of hopelessness in patients correlate with less deactivation in the hippocampus. The extent of the subject's negative attitudes, or pessimism, about the future correlated strongly with the same brain area which distinguished patients and controls with 97% accuracy, demonstrating potential as an imaging biomarker of MDD in the hippocampus.

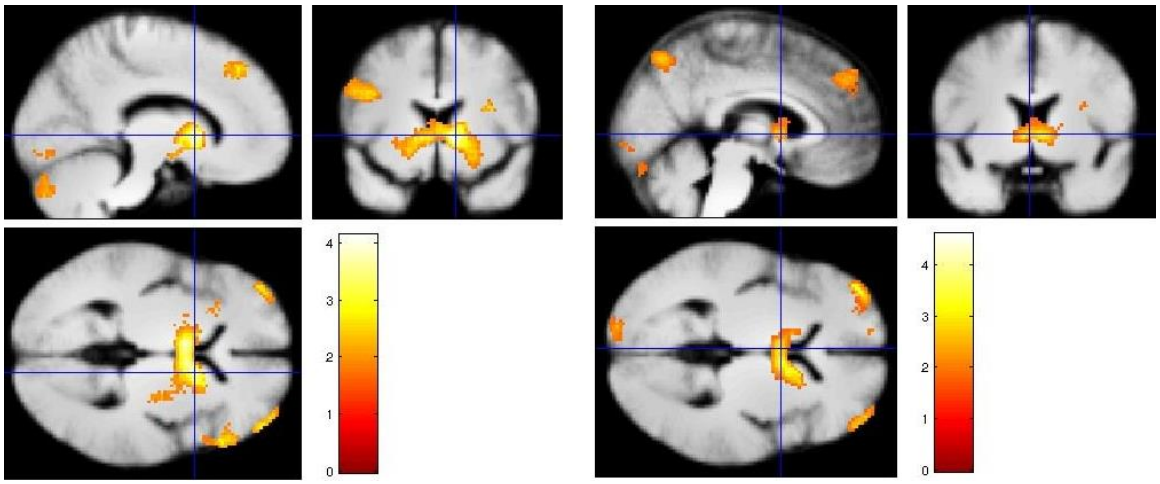

Figure 59: Group-level negative correlations between the basic loss contrast and total scores on the MADRS (left) and HAM-D (right).

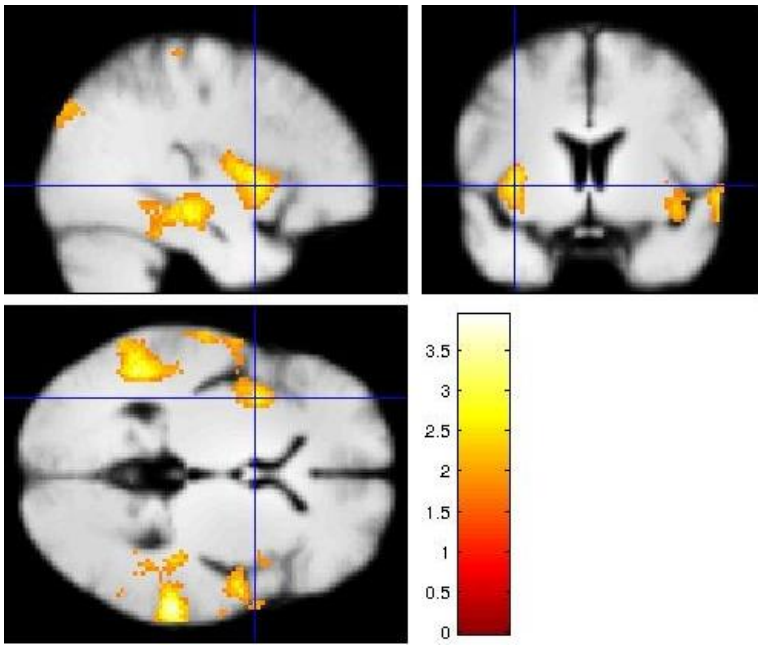

Figure 60: Group-level positive correlations between the controlled loss contrast and total score on the BDI.

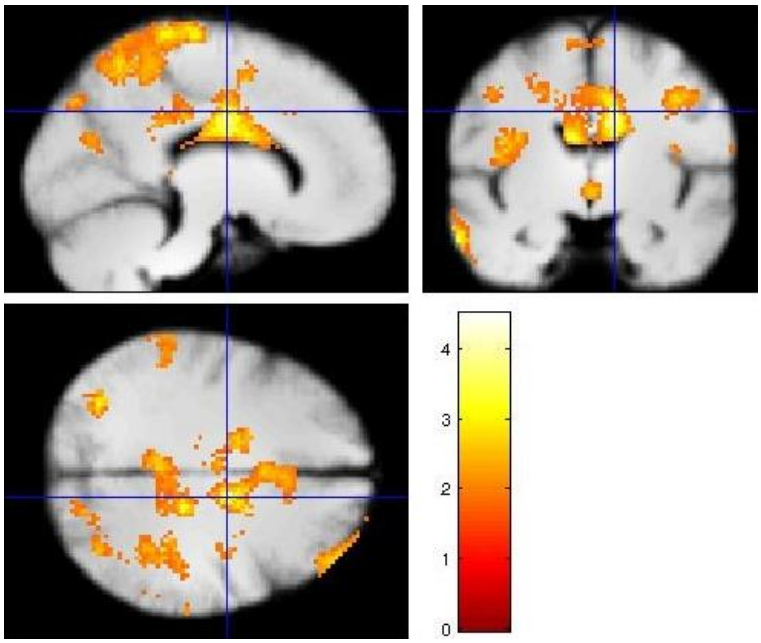

Figure 61: Group-level positive correlations between the controlled loss contrast and HAM-A.

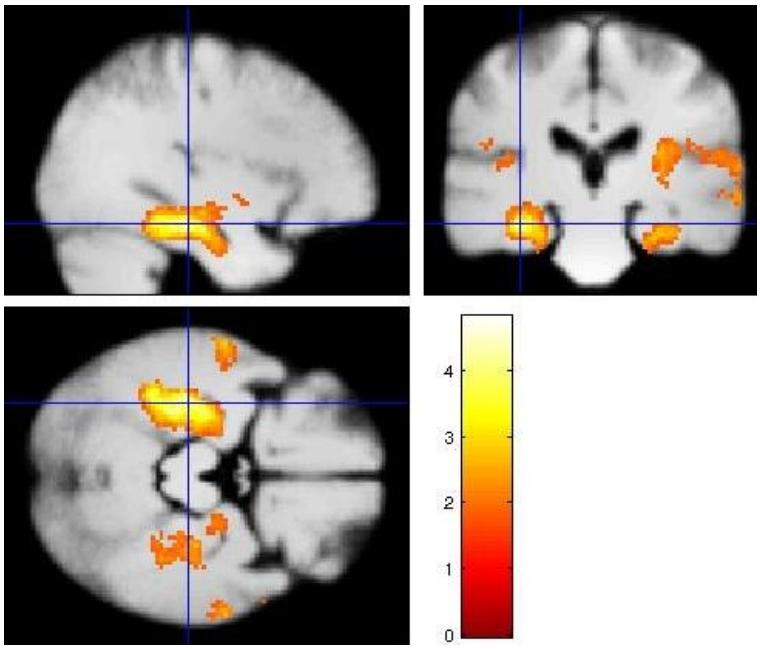

Figure 62: Group-level positive correlations between the controlled loss contrast and total score on the BHS.

8.4 Discussion

When predicting diagnosis using the controlled loss contrast, it was possible to almost perfectly distinguish the MDD group from controls on an *individual* basis. The classification was based on the finding that MDD patients failed to deactivate the hippocampus whilst controls successfully deactivated this region. The correlation identified between activity in the hippocampus and the BHS supports the assertion that this is a syndrome-based abnormality (and not a confound) as increased feelings of hopelessness in patients correlated with less deactivation in the hippocampus. The hippocampus is strongly involved in tasks involving learning and memory and so it is possible that the overactivity found in the hippocampus in MDD patients during loss trials could be due to the patients being hypersensitive to previous aversive events.

Abnormalities in hippocampal volume have been previously reported in MDD (Sapolsky, 2001) and these have been found to remain in remission (Sheline *et al.*, 1996). Whilst the patients in this analysis did not appear to have any structural differences in the hippocampus, grey matter in the hippocampus was found to negatively correlate with both the MADRS and HAM-D total scores, indicating there was a correlation between increased symptom severity and decreased hippocampal volume. An inverse relationship between glucocorticoid levels and hippocampus volume has been reported in animal and human studies (Sapolsky, 2001; Tessner *et al.*, 2007). The glucocorticoid, cortisol, is released in response to stress. As a large proportion of patients with MDD have been found to hypersecrete cortisol (Sapolsky, 2001), increased stress levels in MDD patients may explain why the hippocampus volume has been reported to be decreased.

Patients' abnormal brain activity in the hippocampus when experiencing a loss may also be due to stress. The hippocampus has a large amount of glucocorticoid receptors (Sapolsky, 2001) so if stress levels increase when a loss is received, cortisol is released which could increase hippocampal activity. It would be interesting to investigate the pattern of activity throughout the brain with structural differences taken into account as, if the hippocampal volume is reduced in patients compared to controls, the over-activation to aversive stimuli in patients would become even stronger when the smaller volume is taken into account.

The insula is another region that was found to have significantly different activity between patients and controls. This region was also identified as having a

positive correlation with patients' BDI scores, in that the more severely depressed patients tended to activate the insula more when experiencing a loss. Similar to the hippocampus, the insula is another region that was found to have reduced grey matter in MDD patients – making the increased activation more remarkable.

The nucleus accumbens was deactivated in both groups during the loss contrasts; however the MADRS and HAM-D scores were found to negatively correlate with patients in this region using the basic loss contrast. It is possible that the patients may deactivate the nucleus accumbens more than controls.

Anxiety, as assessed by the HAM-A questionnaire, was found to be correlated with the anterior cingulate activity in patients. Overactivity in this region when experiencing a loss is an interesting finding because it is approximately the same region where a lesion is created during a cingulotomy – a surgical procedure which is used in treatment resistant depression.

The lower classification accuracy of 85% achieved using the controlled reward contrast is also an interesting result and is equal to the highest accuracy achieved when classifying the same groups using structural MRI. The brain regions that were selected using the feature selection method included the nucleus accumbens and the medial orbitofrontal cortex – both of which failed to significantly activate in MDD patients. The meta-analysis by Zhang *et al.* (2013) suggested that the only region which would demonstrate abnormal (reduced) activity in MDD patients at the reward outcome time was the caudate. This study found that although the caudate had reduced activity in the MDD group compared to controls, there were a number of other regions that were more significantly reduced, such as the nucleus accumbens, the medial orbitofrontal cortex and the posterior cingulate. Although reduced activation in the nucleus accumbens was not found in the meta-analysis, a study by Pizzagalli *et al.* (2009), one of the studies in the meta-analysis that had most similarities with the paradigm used in this study, reported the same result. Similar to the loss contrast, MDD patients also showed an increased activation in the insula, however, unlike the loss contrast, it was not found to be significantly correlated with BDI scores.

The limitations of this study are similar to those outlined in Chapter 7: the results require replication in a larger study, the patient group were recruited with a past or present diagnosis of MDD and patients were taking a range of antidepressant medications at the time of scanning. It seems unlikely that the effects of medication

were a major confound in this study as a number of brain regions identified in the classifier and fMRI analyses correlated with various symptom severity scores and previous results in the literature, suggesting that the results are based on diagnosis rather than a medication effect.

To summarise, it was possible to accurately classify MDD subjects and healthy controls on an individual level using their biological responses to rewarding and aversive events. The brain regions which showed abnormal activity in the loss contrast were also found to be correlated with various severity scores, increasing the confidence in the results. The especially striking result is the high classification accuracy when machine learning methods were applied to the controlled loss contrast. This result requires replication in a larger dataset, but it is clear that the response to aversive stimuli is an under-researched and important research area that may be able to increase the understanding of MDD.

Chapter 9: Conclusion

The application of machine learning to neuroimaging data has the potential to improve the clinical treatment and understanding of psychiatric disorders. This thesis has shown the potential of pattern recognition algorithms to complement clinical diagnosis and the potential to identify biomarkers of psychiatric disorders, using feature selection in conjunction with machine learning techniques.

Machine learning-based methods have been successfully applied to predict MPH response in children and adolescents with ADHD using only demographic and clinical variables and neuropsychological test scores. Using only a small number of variables this method achieved an accuracy of 77% when predicting on novel data. However, if a wider range of uncorrelated variables were included (potentially including genetic or neuroimaging data) it is possible that accuracy could be increased further.

In addition, it has been shown that machine learning algorithms can be combined with structural MRI images to predictively diagnose both ADHD and MDD, achieving accuracies of 93% and 85% respectively. Notably, white matter was found to be more predictive of ADHD diagnosis than grey matter, with a large brainstem volume reduction identified in children and adolescents with ADHD. The accuracy achieved and brain regions identified when predicting MDD diagnosis using grey matter images was comparable with similar studies (Costafreda *et al.*, 2009a; Kipli *et al.*, 2013; Mwangi *et al.*, 2012a; Termenon *et al.*, 2013). Interestingly, although the white matter MDD prediction accuracy was poorer than the grey matter prediction accuracy, MDD subjects were found to have increased white matter volume in the cingulate gyrus and posterior cingulate which was predictive of MDD. Increases in white matter volume have been reported in other disorders such as ASD (Herbert *et al.*, 2003; Herbert *et al.*, 2004), treatment-naïve obsessive compulsive disorder (Atmaca *et al.*, 2007), body dysmorphic disorder (Rauch *et al.*, 2003) and schizophrenia (Suzuki *et al.*, 2002). The white matter component of structural MRI is rarely investigated which may explain why it has not been reported previously, however, these findings require further examination using DTI. All subjects in the MDD study have DTI data which will be analysed in the future. A potential next step of these diagnostic classification studies is to test for multiple psychiatric disorders or multiple comorbid disorders. For example, a

method to diagnose individual subjects with unipolar or bipolar depression would be beneficial to clinicians.

The MDD patients' structural MR images were also used to predict symptom severity scores with white matter differences providing a more accurate prediction than grey matter. An interesting extension to this work could be to investigate if it is possible to predict the change of symptom severity over time. A longitudinal study of this type could identify brain differences which aid recovery, potentially providing a target for treatment.

Finally, the machine learning methods were also shown to be able to successfully distinguish MDD patients and healthy controls by their brain activity in response to rewarding and aversive stimuli. When predicting on the basis of response to rewarding events, 84% accuracy was achieved with patients' failure to activate the nucleus accumbens and medial orbitofrontal cortex driving the classification. The highest accuracy reported in this thesis was obtained when prediction was based on the responses to aversive events. Patients' failure to deactivate the hippocampus when receiving a loss contributed to the classifier distinguishing patients and controls with 97% accuracy. Although fMRI resolution is poorer than structural MRI, the results obtained in this study suggest that fMRI may be able to provide more reliable predictions with machine learning algorithms than brain structure in MDD. Therefore future studies should investigate a wide range fMRI tasks to investigate whether they can be reliably used in machine learning-based predictions and increase the understanding of psychiatric disorders.

Psychiatric disorders have previously been thought to originate from "functional" rather than structural differences in the brain. However, the work within this thesis shows that differences exist with regard to both brain structure and function, when comparing patients with MDD and healthy controls. It would be interesting to investigate how these two sets of brain abnormalities are linked and whether they can be combined to help increase understanding of MDD, plus identify the mechanisms which lead to this debilitating disorder.

Future work may investigate whether a classifier which used images from a number of different modalities (e.g. structural and functional MRI, DTI, demographic and clinical variables, neuropsychological test scores and genetics data) could provide accurate and robust predictions to clinical problems.

There are a number of potential limitations to the methods implemented in this thesis which require further discussion. The machine learning-based methods implemented in this thesis require spatial normalisation when applied to neuroimaging data so that each voxel corresponds to the same brain region. However, by definition, spatial normalisation loses subject-specific data by distorting all subjects towards a template meaning potentially important individual subject structural abnormalities and irregularities are removed. Future work could involve creating a procedure to extract a number of characteristics (e.g. size, surface area and volume of each brain region) from the native space images to enter into a classifier.

Many of the subjects in the ADHD and MDD groups analysed in this thesis had received treatment at the time of scanning. As discussed in Chapter 6, medication-naïve subjects with ADHD are currently being recruited to a study aimed at addressing this limitation in the ADHD prediction in Chapter 5. The MDD subjects analysed in this thesis were treatment resistant and so had an extensive medication history. It is unclear to what extent medication history could affect these analyses but future work could compare previously medicated and medication naïve subjects to investigate this further.

Another potential limitation of the MDD studies in this thesis is that treatment-resistant MDD patients may have more pronounced structural and functional abnormalities when compared with subjects with first-episode MDD. It is unclear if a classifier developed using treatment-resistant subjects would obtain a higher, lower or similar accuracy when diagnosing first-episode MDD subjects. Intuitively, the pattern of brain regions which are abnormal in MDD may be easier to identify in the most severely depressed subjects which would increase the likelihood of training on the neurobiological abnormalities related to MDD (Fu and Costafreda, 2013). However, if the brain regions affected by MDD vary with length of symptoms or treatment then a first-episode cohort may be required for a clinically useful prediction of diagnosis. This potential limitation merits further investigation to study the structural and functional brain alterations of subjects with MDD in a longitudinal study from the first episode onwards.

More work is required to replicate all the results obtained in this thesis on much larger datasets but the initial results are encouraging. Furthermore, these methods also need to be applied to more interesting clinical questions such as the iBOCA study (described in Chapter 6), which aims to predict treatment response in

ADHD. That study acquired both structural and functional MR images and there are a number of different analyses planned once the recruitment is complete, as described in Chapter 6.

There are many exciting applications of machine learning algorithms in psychiatric neuroimaging; however, it is important to note that machine learning is an active research field in itself. These techniques are continually being improved, developed and are becoming more robust. MR imaging is another independent research area with MR technology still improving and the accuracy and sophistication of normalisation algorithms progressing rapidly. Higher resolution images with fewer artefacts and more accurate normalisation techniques could increase the reliability and accuracy of machine learning-based predictions.

These advances require investigation in the future and the author intends to continue to investigate various machine learning algorithms to determine which methods generally perform best for different pattern recognition problems. For example, Gaussian Processes are able to provide probabilistic output which may improve on results obtained using the RVR algorithm. They also have the advantage of being able to make multi-class predictions, a limitation of both SVM and RVM.

Another approach to investigate various machine learning algorithms could include using Monte Carlo simulations to test various properties of classifiers prior to application in psychiatric neuroimaging. These could be applied to identify how robust each classifier is to outliers in the data or how sensitive it is to the class imbalance problem. This would provide a clear rationale when deciding which machine learning algorithm to implement in each study.

A major assumption in all supervised machine learning studies is that the labels provided to train and test a classifier are correct. However, in psychiatric diagnosis there is no gold standard so some subjects may be diagnosed incorrectly, increasing the difficulty in identifying the underlying pattern of a psychiatric disorder (Fu and Costafreda, 2013). Furthermore, studies that contain misdiagnosed subjects cannot be expected to achieve perfect classification accuracy as these misdiagnosed subjects would influence both the training stage and the testing stage during cross-validation. Therefore, it is possible that a machine learning-based study in psychiatry could identify the underlying pattern of a disorder without achieving 100% accuracy during cross-validation due to misdiagnosed subjects. The use of unsupervised machine learning algorithms eliminates the issue of misdiagnosis during the training

stage as these methods attempt to identify the underlying patterns in unlabelled data. This thesis only investigated supervised machine learning but studies using unsupervised machine learning algorithms have recently emerged with positive results (Zeng *et al.,* 2013) – removing potential confounds such as diagnosis in the training subjects. Clearly, the potential issue of misdiagnosis remains during the testing stage but the influence of the misdiagnosed subjects is reduced when unsupervised machine learning algorithms are used.

In summary, the results presented in this thesis demonstrate the potential of machine learning algorithms in psychiatry, demonstrating the success of these algorithms to predict diagnosis in two different psychiatric disorders using different imaging modalities. The studies contained within this thesis are a proof of concept and require replication in larger samples. These methods have started to be applied to clinically relevant questions which are less understood such as prognosis (Costafreda *et al.,* 2009a), symptom severity (Mwangi *et al.,* 2012b), identification of patients at risk of developing disorder (Koutsouleris *et al.,* 2011; Koutsouleris *et al.,* 2009), and an estimation of the likelihood of response to treatment (Gong *et al.,* 2011). In addition, with the use of feature selection, machine learning studies have the potential to augment the knowledge of the neurobiology of various psychiatric disorders.

References

Adler CM, Delbello MP, Mills NP, Schmithorst V, Holland S, Strakowski SM. Comorbid ADHD is associated with altered patterns of neuronal activation in adolescents with bipolar disorder performing a simple attention task. Bipolar Disorders 2005; 7: 577-588.

Afshar F, Watkins ES, Yap JC. Stereotaxic Atlas of the Human Brainstem and Cerebellar Nuclei: A Variability Study. New York: Raven Press, 1978.

Agarwal N, Port JD, Bazzocchi M, Renshaw PF. Update on the use of MR for assessment and diagnosis of psychiatric diseases. Radiology 2010; 255: 23.

Alexander GE, DeLong MR, Strick PL. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annual Review Of Neuroscience 1986; 9: 357-381.

Alpaydin E. Introduction to machine learning: MIT Press, 2004.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV-TR: American Psychiatric Publishing, Inc., 2000.

Anderson CM, Polcari A, Lowen SB, Renshaw PF, Teicher MH. Effects of methylphenidate on functional magnetic resonance relaxometry of the cerebellar vermis in boys with ADHD. The American Journal Of Psychiatry 2002; 159: 1322-1328.

Arnsten AF. Catecholamine modulation of prefrontal cortical cognitive function. Trends In Cognitive Sciences 1998; 2: 436-447.

Arnsten AF, Steere JC, Hunt RD. The contribution of alpha 2-noradrenergic mechanisms of prefrontal cortical cognitive function. Potential significance for attention-deficit hyperactivity disorder. Archives Of General Psychiatry 1996; 53: 448-455.

Arnsten AFT, Li B-M. Neurobiology of Executive Functions: Catecholamine Influences on Prefrontal Cortical Functions. Biological Psychiatry 2005; 57: 1377-1384.

Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage 2007; 38: 95-113.

Ashburner J, Barnes G, Chen C-C, Daunizeau J, Flandin G, Friston K, *et al.* SPM8 Manual: Functional Imaging Laboratory, University College London, 2012: 475.

Atmaca M, Yildirim H, Ozdemir H, Tezcan E, Kursad Poyraz A. Volumetric MRI study of key brain regions implicated in obsessive-compulsive disorder. Progress in Neuro-Psychopharmacology and Biological Psychiatry 2007; 31: 46-52.

Banaschewski T, Coghill D, Danckaerts M, Dopfner M, Rohde L, Sergeant JA, *et al.* ADHD and Hyperkinetic Disorder. New York: Oxford University Press, 2010.

Bennett IJ, Madden DJ, Vaidya CJ, Howard DV, Howard JH. Age-related differences in multiple measures of white matter integrity: A diffusion tensor imaging study of healthy aging. Human brain mapping 2010; 31: 378-390.

Berquin PC, Giedd JN, Jacobsen LK, Hamburger SD, Krain AL, Rapoport JL, *et al.* Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. Neurology 1998; 50: 1087-1093.

Bishop CM. Pattern recognition and machine learning: Springer, 2006.

Bohland JW, Saperstein S, Pereira F, Rapin J, Grady L. Network, anatomical, and non-imaging measures for the prediction of ADHD diagnosis in individual subjects. Frontiers in systems neuroscience 2012; 6.

Booth JR, Burman DD, Meyer JR, Lei Z, Trommer BL, Davenport ND, *et al.* Larger deficits in brain networks for response inhibition than for visual selective attention in attention deficit hyperactivity disorder (ADHD). Journal Of Child Psychology And Psychiatry, And Allied Disciplines 2005; 46: 94-111.

Bora E, Fornito A, Pantelis C, Yücel M. Gray matter abnormalities in major depressive disorder: a meta-analysis of voxel based morphometry studies. Journal of affective disorders 2012; 138: 9-18.

Bray S, Chang C, Hoeft F. Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. Frontiers In Human Neuroscience 2009; 3: 32-32.

Brotman MA, Rich BA, Guyer AE, Lunsford JR, Horsey SE, Reising MM, *et al.* Amygdala activation during emotion processing of neutral faces in children with severe mood dysregulation versus ADHD or bipolar disorder. The American Journal Of Psychiatry 2010; 167: 61-69.

Brown MRG, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, *et al.* ADHD-200 Global Competition: diagnosing ADHD using personal

characteristic data can outperform resting state fMRI measurements. Frontiers in systems neuroscience 2012; 6.

Bussing R, Grudnik J, Mason D, Wasiak M, Leonard C. ADHD and conduct disorder: an MRI study in a community sample. The World Journal Of Biological Psychiatry: The Official Journal Of The World Federation Of Societies Of Biological Psychiatry 2002; 3: 216-220.

Cao Q, Zang Y, Zhu C, Cao X, Sun L, Zhou X, *et al.* Alerting deficits in children with attention deficit/hyperactivity disorder: event-related fMRI evidence. Brain Research 2008; 1219: 159-168.

Cao X, Cao Q, Long X, Sun L, Sui M, Zhu C, *et al.* Abnormal resting-state functional connectivity patterns of the putamen in medication-naive children with attention deficit hyperactivity disorder. Brain Research 2009; 1303: 195-206.

Carmona S, Vilarroya O, Bielsa A, Trèmols V, Soliva JC, Rovira M, *et al.* Global and regional gray matter reductions in ADHD: A voxel-based morphometric study. Neuroscience Letters 2005; 389: 88-93.

Castellanos FX, Giedd JN, Marsh WL, Hamburger SD, Vaituzis AC, Dickstein DP, *et al.* Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. Archives Of General Psychiatry 1996; 53: 607-616.

Castellanos FX, Lee PP, Sharp W, Jeffries NO, Greenstein DK, Clasen LS, *et al.* Developmental trajectories of brain volume abnormalities in children and adolescents with attention-deficit/hyperactivity disorder. JAMA: The Journal Of The American Medical Association 2002; 288: 1740-1748.

Castellanos FX, Sharp WS, Gottesman RF, Greenstein DK, Giedd JN, Rapoport JL. Anatomic brain abnormalities in monozygotic twins discordant for attention deficit hyperactivity disorder. American Journal of Psychiatry 2003; 160: 1693-1696.

Chamberlain SR, Muller U, Blackwell AD, Clark L, Robbins TW, Sahakian BJ. Neurochemical Modulation of Response Inhibition and Probabilistic Learning in Humans. Science 2006; 311: 861-863.

Chang C-W, Ho C-C, Chen J-H. ADHD classification by a texture analysis of anatomical brain MRI data. Frontiers in systems neuroscience 2012; 6.

Chaves R, Ramirez J, Gorriz JM, Lopez M, Salas-Gonzalez D, Alvarez I, *et al.* SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test

NMSE feature selection with feature correlation weighting. Neuroscience Letters 2009; 461: 293-297.

Chen C-H, Ridler K, Suckling J, Williams S, Fu CHY, Merlo-Pich E, *et al.* Brain imaging correlates of depressive symptom severity and predictors of symptom improvement after antidepressant treatment. Biological psychiatry 2007; 62: 407-414.

Cheng W, Ji X, Zhang J, Feng J. Individual classification of ADHD patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. Frontiers In Systems Neuroscience 2012; 6: 58-58.

Coghill D, Banaschewski T. The genetics of attention-deficit/hyperactivity disorder. Expert Review of Neurotherapeutics 2009; 9: 1547-1565.

Coghill D, Nigg J, Rothenberger A, Sonuga-Barke E, Tannock R. Whither causal models in the neuroscience of ADHD? Developmental Science 2005; 8: 105-114.

Coghill D, Rohde LA, Banaschewski T. Attention Deficit Hyperactivity Disorder. Biological Child Psychiatry 2008; 24.

Coghill DR, Rhodes SM, Matthews K. The neuropsychological effects of chronic methylphenidate on drug-naive boys with attention-deficit/hyperactivity disorder. Biological Psychiatry 2007; 62: 954-962.

Cohen J. Statistical power analysis for the behavioral sciences: Academic Press, 1977.

Colby JB, Rudie JD, Brown JA, Douglas PK, Cohen MS, Shehzad Z. Insights into multimodal imaging classification of ADHD. Frontiers In Systems Neuroscience 2012; 6: 59-59.

Cole J, Chaddock CA, Farmer AE, Aitchison KJ, Simmons A, McGuffin P, *et al.* White matter abnormalities and illness severity in major depressive disorder. The British Journal of Psychiatry 2012; 201: 33-39.

Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20: 273-297.

Costafreda SG, Chu C, Ashburner J, Fu CHY. Prognostic and Diagnostic Potential of the Structural Neuroanatomy of Depression. PLoS ONE 2009a; 4: e6353.

Costafreda SG, Khanna A, Mourão-Miranda J, Fu CHY. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. Neuroreport 2009b; 20: 637-641.

Craddock RC, Holtzheimer III PE, Hu XP, Mayberg HS. Disease state prediction from resting state functional connectivity. Magnetic Resonance In Medicine: Official Journal Of The Society Of Magnetic Resonance In Medicine / Society Of Magnetic Resonance In Medicine 2009; 62: 1619-1628.

Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods: Cambridge University Press, 2000.

D'Agata F, Caroppo P, Boghi A, Coriasco M, Caglio M, Baudino B, *et al.* Linking coordinative and executive dysfunctions to atrophy in spinocerebellar ataxia 2 patients. Brain Structure & Function 2011; 216: 275-288.

Dai D, Wang J, Hua J, He H. Classification of ADHD children through multimodal magnetic resonance imaging. Frontiers in systems neuroscience 2012; 6.

Dash M, Liu H. Feature Selection for Classification. Intelligent Data Analysis 1997; 1: 131-156.

De Martino F, Valente G, Staeren Nl, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage 2008; 43: 44-58.

Del Campo N, Chamberlain SR, Sahakian BJ, Robbins TW. The roles of dopamine and noradrenaline in the pathophysiology and treatment of attention-deficit/hyperactivity disorder. Biological Psychiatry 2011; 69: e145-e157.

Denney CB, Rapport MD. Predicting methylphenidate response in children with ADHD: theoretical, empirical, and conceptual models. Journal Of The American Academy Of Child And Adolescent Psychiatry 1999; 38: 393-401.

Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. Journal Of Neuroscience Methods 2002; 118: 115-128.

Diedrichsen J. A spatially unbiased atlas template of the human cerebellum. Neuroimage 2006; 33: 127-138.

Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N. A probabilistic MR atlas of the human cerebellum. Neuroimage 2009; 46: 39-46.

Dollar P. Piotr's Image and Video Matlab Toolbox, 2011.

Döpfner M, Lehmkuhl G. Diagnostik-System für psychische Störungen im Kindes- und Jugendalter nach ICD-10 und DSM-IV. Bern: Huber, 1998.

Dorée J-P, Rosiers JD, Lew V, Gendron A, Elie R, Stip E, *et al.* Quetiapine augmentation of treatment-resistant depression: a comparison with lithium. Current Medical Research and Opinion 2007; 23: 333-341.

Duchesnay E, Cachia A, Boddaert N, Chabane N, Mangin J-F, Martinot J-L, *et al.* Feature selection and classification of imbalanced datasets: application to PET images of children with autistic spectrum disorders. Neuroimage 2011; 57: 1003-1014.

Durston S, Davidson MC, Mulder MJ, Spicer JA, Galvan A, Tottenham N, *et al.* Neural and behavioral correlates of expectancy violations in attention-deficit hyperactivity disorder. Journal Of Child Psychology And Psychiatry, And Allied Disciplines 2007; 48: 881-889.

Durston S, Mulder M, Casey BJ, Ziermans T, van Engeland H. Activation in ventral prefrontal cortex is sensitive to genetic vulnerability for attention-deficit hyperactivity disorder. Biological Psychiatry 2006; 60: 1062-1070.

Durston S, Pol HEH, Schnack HG, Buitelaar JK, Steenhuis MP, Minderaa RB, *et al.* Magnetic resonance imaging of boys with attention-deficit/hyperactivity disorder and their unaffected siblings. Journal of the American Academy of Child & Adolescent Psychiatry 2004; 43: 332-340.

Durston S, Tottenham NT, Thomas KM, Davidson MC, Eigsti I-M, Yang Y, *et al.* Differential patterns of striatal activation in young children with and without ADHD. Biological Psychiatry 2003; 53: 871-878.

Ecker C, Marquand A, Mourão-Miranda J, Johnston P, Daly EM, Brammer MJ, *et al.* Describing the brain in autism in five dimensions--magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. The Journal Of Neuroscience: The Official Journal Of The Society For Neuroscience 2010a; 30: 10612-10623.

Ecker C, Rocha-Rego V, Johnston P, Mourão-Miranda J, Marquand A, Daly EM, *et al.* Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. NeuroImage 2010b; 49: 44-56.

Elkis H, Friedman L, Wise A, Meltzer HY. Meta-analyses of studies of ventricular enlargement and cortical sulcal prominence in mood disorders: comparisons with controls or patients with schizophrenia. Archives of General Psychiatry 1995; 52: 735.

Ellison-Wright I, Ellison-Wright Z, Bullmore E. Structural brain change in Attention Deficit Hyperactivity Disorder identified by meta-analysis. BMC Psychiatry 2008; 8: 51-51.

Eloyan A, Muschelli J, Nebel MB, Liu H, Han F, Zhao T, *et al.* Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. Frontiers In Systems Neuroscience 2012; 6: 61-61.

Epstein JN, Casey BJ, Tonev ST, Davidson M, Reiss AL, Garrett A, *et al.* Assessment and prevention of head motion during imaging of patients with attention deficit hyperactivity disorder. Psychiatry Research: Neuroimaging 2007a; 155: 75-82.

Epstein JN, Casey BJ, Tonev ST, Davidson MC, Reiss AL, Garrett A, *et al.* ADHD- and medication-related brain activation effects in concordantly affected parent-child dyads with ADHD. Journal Of Child Psychology And Psychiatry, And Allied Disciplines 2007b; 48: 899-913.

Epstein JN, Delbello MP, Adler CM, Altaye M, Kramer M, Mills NP, *et al.* Differential patterns of brain activation over time in adolescents with and without attention deficit hyperactivity disorder (ADHD) during performance of a sustained attention task. Neuropediatrics 2009; 40: 1-5.

Ernst M, Rumsey J, Munson S. Update of Functional Neuroimaging in Child Psychiatry. In: Fu CHY, Senior C, Russell T, Weinberger DR and Murray R, editors. Neuroimaging in psychiatry: Martin Dunitz, 2003.

Fair DA, Bathula D, Nikolas MA, Nigg JT. Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. Proceedings Of The National Academy Of Sciences Of The United States Of America 2012a; 109: 6769-6774.

Fair DA, Nigg JT, Iyer S, Bathula D, Mills KL, Dosenbach NUF, *et al.* Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. Frontiers in systems neuroscience 2012b; 6.

Fang P, Zeng L-L, Shen H, Wang L, Li B, Liu L, *et al.* Increased cortical-limbic anatomical network connectivity in major depression revealed by diffusion tensor imaging. PloS one 2012; 7: e45972.

Fassbender C, Zhang H, Buzy WM, Cortes CR, Mizuiri D, Beckett L, *et al.* A lack of default network suppression is linked to increased distractibility in ADHD. Brain Research 2009; 1273: 114-128.

Fava M. Diagnosis and definition of treatment-resistant depression. Biological psychiatry 2003; 53: 649-659.

Filipek PA, Semrud-Clikeman M, Steingard RJ, Renshaw PF, Kennedy DN, Biederman J. Volumetric MRI analysis comparing subjects having attention-deficit hyperactivity disorder with normal controls. Neurology 1997; 48: 589-601.

Fletcher T. Support Vector Machines Explained: University College London (UCL), 2009.

Floares A, Jakary A, Bornstein A, Deicken R. Neural Networks and Classification & Regression Trees Are Able to Distinguish Female with Major Depression from Healhy Controls Using Neuroimaging Data. Neural Networks, 2006. IJCNN'06. International Joint Conference on: IEEE, 2006: 4605-4611.

Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD. Statistical Parametric Mapping: The Analysis of Functional Brain Images. London: Academic Press, 2007.

Frodl T, Skokauskas N. Meta-analysis of structural MRI studies in children and adults with attention deficit hyperactivity disorder indicates treatment effects. Acta Psychiatrica Scandinavica 2012; 125: 114-126.

Fu CHY, Costafreda SG. In Review Neuroimaging-Based Biomarkers in Psychiatry: Clinical Opportunities of a Paradigm Shift. Canadian Journal of Psychiatry 2013; 58.

Fu CHY, Mourão-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SCR, *et al.* Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biological psychiatry 2008; 63: 656-662.

Fu CHY, Russell T, Murray R, Weinberger DR. Neuroimaging in psychiatry: Martin Dunitz, 2003.

Gao W, Zeng L-L, Shen H, Hu D. Estimating medication status via resting-state functional connectivity in major depression. Intelligent Science and Intelligent Data Engineering: Springer, 2012: 153-159.

Gilbert SJ, Spengler S, Simons JS, Steele JD, Lawrie SM, Frith CD, *et al.* Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. Journal Of Cognitive Neuroscience 2006; 18: 932-948.

Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, *et al.* Prognostic prediction of therapeutic response in depression using high-field MR imaging. Neuroimage 2011; 55: 1497-1503.

Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 2001; 14: 21-36.

Gradin VB, Baldacchino A, Balfour D, Matthews K, Steele JD. Abnormal Brain Activity during a Reward and Loss Task in Opiate Dependent Patients receiving Methadone Maintenance Therapy. Neuropsychopharmacology 2013.

Grotegerd D, Suslow T, Bauer J, Ohrmann P, Arolt V, Stuhrmann A, *et al.* Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study. European archives of psychiatry and clinical neuroscience 2012: 1-13.

Guitart-Masip M, Chowdhury R, Sharot T, Dayan P, Duzel E, Dolan RJ. Action controls dopaminergic enhancement of reward representations. Proceedings Of The National Academy Of Sciences Of The United States Of America 2012; 109: 7511-7516.

Guyon I, Elisseeff A. An introduction to variable and feature selection. J. Mach. Learn. Res. 2003; 3: 1157-1182.

Habes I, Krall SC, Johnston SJ, Yuen KSL, Healy D, Goebel R, *et al.* Pattern classification of valence in depression. NeuroImage: Clinical 2013.

Hahn T, Marquand AF, Ehlis A-C, Dresler T, Kittel-Schneider S, Jarczok TA, *et al.* Integrating neurobiological markers of depression. Archives of general psychiatry 2011; 68: 361.

Haubold A, Peterson BS, Bansal R. Annual Research Review: Progress in using brain morphometry as a clinical tool for diagnosing psychiatric disorders. Journal of Child Psychology and Psychiatry 2012; 53: 519-535.

Hazari H, Christmas D, Matthews K. The clinical utility of different quantitative methods for measuring treatment resistance in major depression. Journal of affective disorders 2013; 150: 231-236.

Herbert MR, Ziegler DA, Deutsch CK, O'Brien LM, Lange N, Bakardjiev A, *et al.* Dissociations of cerebral cortex, subcortical and cerebral white matter volumes in autistic boys. Brain 2003; 126: 1182-1192.

Herbert MR, Ziegler DA, Makris N, Filipek PA, Kemper TL, Normandin JJ, *et al.* Localization of white matter volume increase in autism and developmental language disorder. Annals of Neurology 2004; 55: 530-540.

Hermann B, Jones J, Dabbs K, Allen CA, Sheth R, Fine J, *et al.* The frequency, complications and aetiology of ADHD in new onset paediatric epilepsy. Brain 2007; 130: 3135-3148.

Hill DE, Yeo RA, Campbell RA, Hart B, Vigil J, Brooks W. Magnetic resonance imaging correlates of attention-deficit/hyperactivity disorder in children. Neuropsychology 2003; 17: 496-506.

Hoekzema E, Carmona S, Tremols V, Gispert JD, Guitart M, Fauquet J, *et al.* Enhanced neural activity in frontal and cerebellar circuits after cognitive training in children with attention-deficit/hyperactivity disorder. Human Brain Mapping 2010; 31: 1942-1950.

Hoexter MQ, Miguel EC, Diniz JB, Shavitt RG, Busatto GF, Sato JR. Predicting obsessive-compulsive disorder severity combining neuroimaging and machine learning methods. Journal of Affective Disorders 2013.

Humphrey N. A History of the Mind. New York: Simon and Schuster, 1992.

Hynd GW, Semrud-Clikeman M, Lorys AR, Novey ES, Eliopulos D. Brain morphology in developmental dyslexia and attention deficit disorder/hyperactivity. Archives Of Neurology 1990; 47: 919-926.

Ingalhalikar M, Kanterakis S, Gur R, Roberts T, Verma R, Jiang T, *et al.* DTI Based Diagnostic Prediction of a Disease via Pattern Classification. In: Jiang T, Navab N, Pluim J and Viergever M, editors. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2010. Vol 6361. Berlin / Heidelberg: Springer-Verlag Berlin / Heidelberg, 2010: 558-565.

Jacobson NS, Truax P. Clinical significance: A statistical approach to denning meaningful change in psychotherapy research. Journal of consulting and clinical psychology 1991; 59: 12-19.

Johnston BA, Mwangi B, Matthews K, Coghill D, Steele JD. Predictive classification of individual magnetic resonance imaging scans from children and adolescents. European Child & Adolescent Psychiatry 2012: 1-12.

Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, *et al.* Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. Journal Of The American Academy Of Child And Adolescent Psychiatry 1997; 36: 980-988.

Kelly AMC, Margulies DS, Castellanos FX. Recent advances in structural and functional brain imaging studies of attention-deficit/hyperactivity disorder. *Current Psychiatry Reports* 2007; 9: 401-407.

Kempton MJ, Salvador Z, Munafò MR, Geddes JR, Simmons A, Frangou S, *et al.* Structural neuroimaging studies in major depressive disorder: meta-analysis and comparison with bipolar disorder. Archives of general psychiatry 2011; 68: 675.

Keren NI, Lozar CT, Harris KC, Morgan PS, Eckert MA. In vivo mapping of the human locus coeruleus. Neuroimage 2009; 47: 1261-1267.

Kieseppä T, Eerola M, Mäntylä R, Neuvonen T, Poutanen V-P, Luoma K, *et al.* Major depressive disorder and white matter abnormalities: a diffusion tensor imaging study with tract-based spatial statistics. Journal of affective disorders 2010; 120: 240-244.

Kim MJ, Hamilton JP, Gotlib IH. Reduced caudate gray matter volume in women with major depressive disorder. Psychiatry Research: Neuroimaging 2008; 164: 114-122.

Kipli K, Kouzani AZ, Joordens M. Evaluation of Feature Selection Algorithms for Detection of Depression from Brain sMRI Scans. Proceedings of 2013 ICME International Conference on Complex Medical Engineering. Beijing, 2013.

Klöppel S, Abdulkadir A, Jack Jr CR, Koutsouleris N, Mourão-Miranda J, Vemuri P. Diagnostic neuroimaging across diseases. Neuroimage 2012; 61: 457-463.

Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, *et al.* Automatic classification of MR scans in Alzheimer's disease. Brain 2008; 131: 681-9.

Knutson B, Bhanji JP, Cooney RE, Atlas LY, Gotlib IH. Neural responses to monetary incentives in major depression. Biological psychiatry 2008; 63: 686-692.

Kobel M, Bechtel N, Weber P, Specht K, Klarhofer M, Scheffler K, *et al.* Effects of methylphenidate on working memory functioning in children with attention

deficit/hyperactivity disorder. European Journal Of Paediatric Neurology: EJPN: Official Journal Of The European Paediatric Neurology Society 2009; 13: 516-523.

Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine 2001; 23: 89-109.

Konrad K, Neufang S, Hanisch C, Fink GR, Herpertz-Dahlmann B. Dysfunctional attentional networks in children with attention deficit/hyperactivity disorder: evidence from an event-related functional magnetic resonance imaging study. Biological Psychiatry 2006; 59: 643-651.

Koolschijn P, van Haren NEM, Lensvelt-Mulders GJLM, Pol H, Hilleke E, Kahn RS. Brain volume abnormalities in major depressive disorder: A meta-analysis of magnetic resonance imaging studies. Human brain mapping 2009; 30: 3719-3735.

Korgaonkar MS, Grieve SM, Koslow SH, Gabrieli JDE, Gordon E, Williams LM. Loss of white matter integrity in major depressive disorder: Evidence using tract-based spatial statistical analysis of diffusion tensor imaging. Human brain mapping 2011; 32: 2161-2171.

Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Moller H-J, Riecher-Rössler A. Disease Prediction in the At-Risk Mental State for Psychosis Using Neuroanatomical Biomarkers: Results From the FePsy Study. Schizophrenia Bulletin 2011.

Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, *et al.* Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch Gen Psychiatry 2009; 66: 700-12.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience 2009; 12: 535-540.

Levy F, Swanson JM. Timing, space and ADHD: the dopamine theory revisited. The Australian And New Zealand Journal Of Psychiatry 2001; 35: 504-511.

Li C-T, Lin C-P, Chou K-H, Chen I, Hsieh J-C, Wu C-L, *et al.* Structural and cognitive deficits in remitting and non-remitting recurrent depression: a voxel-based morphometric study. Neuroimage 2010; 50: 347-356.

Li L, Ma N, Li Z, Tan L, Liu J, Gong G, *et al.* Prefrontal white matter abnormalities in young adult with major depressive disorder: a diffusion tensor imaging study. Brain research 2007; 1168: 124-128.

Liang S-F, Hsieh T-H, Chen P-T, Wu M-L, Kung C-C, Lin C-Y, *et al.* Differentiation between resting-state fMRI data from ADHD and normal subjects: Based on functional connectivity and machine learning. Fuzzy Theory and it's Applications (iFUZZY), 2012 International Conference on: IEEE, 2012: 294-298.

Lim L, Marquand A, Cubillo AA, Smith AB, Chantiluke K, Simmons A, *et al.* Disorder-Specific Predictive Classification of Adolescents with Attention Deficit Hyperactivity Disorder (ADHD) Relative to Autism Using Structural Magnetic Resonance Imaging. PloS one 2013; 8: e63660.

Linden David EJ. The Challenges and Promise of Neuroimaging in Psychiatry. Neuron 2012; 73: 8-22.

Liu F, Guo W, Yu D, Gao Q, Gao K, Xue Z, *et al.* Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PloS one 2012; 7: e40968.

Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. The Lancet 2006; 367: 1747-1757.

Lord A, Horn D, Breakspear M, Walter M. Changes in community structure of resting state functional connectivity in unipolar depression. PloS one 2012; 7: e41282.

Ma N, Li L, Shu N, Liu J, Gong G, He Z, *et al.* White matter abnormalities in first-episode, treatment-naive young adults with major depressive disorder. American Journal of Psychiatry 2007; 164: 823-826.

Ma Q, Zeng L-L, Shen H, Liu L, Hu D. Altered cerebellar-cerebral resting-state functional connectivity reliably identifies major depressive disorder. Brain research 2012.

Mai JK, Assheuer J, Paxinos G. Atlas of the human brain. Dusseldorf: Academic Press, 1997.

Marquand AF, Mourão-Miranda J, Brammer MJ, Cleare AJ, Fu CHY. Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. Neuroreport 2008; 19: 1507-1511.

Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS medicine 2006; 3: e442.

Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J. A Probabilistic Atlas of the Human Brain: Theory and Rationale for Its Development: The International Consortium for Brain Mapping (ICBM). NeuroImage 1995; 2: 89-101.

McRobbie DW, Moore EA, Graves MJ, Prince MR. MRI from Picture to Proton. New York: Cambridge University Press, 2010.

Meyfroidt G, Güiza F, Ramon J, Bruynooghe M. Machine learning techniques to examine large patient databases. Best Practice & Research Clinical Anaesthesiology 2009; 23: 127-143.

Milham MP, Fair D, Mennes M, Mostofsky SH. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in Systems Neuroscience 2012; 6.

Moorhead TWJ, Gountouna V-E, Job DE, McIntosh AM, Romaniuk L, Lymer GKS, *et al.* Prospective multi-centre Voxel Based Morphometry study employing scanner specific segmentations: Procedure development using CaliBrain structural MRI data. BMC medical imaging 2009; 9: 8.

Mostofsky SH, Reiss AL, Lockhart P, Denckla MB. Evaluation of cerebellar size in attention-deficit hyperactivity disorder. Journal Of Child Neurology 1998; 13: 434-439.

Mostofsky SH, Rimrodt SL, Schafer JGB, Boyce A, Goldberg MC, Pekar JJ, *et al.* Atypical motor and sensory cortex activation in attention-deficit/hyperactivity disorder: a functional magnetic resonance imaging study of simple sequential finger tapping. Biological Psychiatry 2006; 59: 48-56.

Mourão-Miranda J, Almeida JRC, Hassel S, de Oliveira L, Versace A, Marquand AF, *et al.* Pattern recognition analyses of brain activation elicited by happy and neutral faces in unipolar and bipolar depression. Bipolar Disorders 2012a; 14: 451-460.

Mourão-Miranda J, Hardoon DR, Hahn T, Marquand AF, Williams SCR, Shawe-Taylor J, *et al.* Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. NeuroImage 2011; 58: 793-804.

Mourão-Miranda J, Oliveira L, Ladouceur CD, Marquand A, Brammer M, Birmaher B, *et al.* Pattern recognition and functional neuroimaging help to discriminate

healthy adolescents at risk for mood disorders from low risk adolescents. PloS one 2012b; 7: e29482.

Mueller A, Candrian G, Grane VA, Kropotov JD, Ponomarev VA, Baschera G-M. Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: a validation study. Nonlinear Biomedical Physics 2011; 5: 5-5.

Mwangi B, Ebmeier KP, Matthews K, Douglas Steele J. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain: A Journal Of Neurology 2012a; 135: 1508-1521.

Mwangi B, Matthews K, Steele JD. Prediction of illness severity in patients with major depression using structural MR brain scans. Journal Of Magnetic Resonance Imaging: JMRI 2012b; 35: 64-71.

Mwangi B, Tian TS, Soares JC. A Review of Feature Reduction Techniques in Neuroimaging. Neuroinformatics 2013: 1-16.

Nakao T, Radua J, Rubia K, Mataix-Cols D. Gray matter volume abnormalities in ADHD: voxel-based meta-analysis exploring the effects of age and stimulant medication. The American Journal Of Psychiatry 2011; 168: 1154-1163.

Nouretdinov I, Costafreda SG, Gammerman A, Chervonenkis A, Vovk V, Vapnik V, *et al.* Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage 2011; 56: 809-813.

Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 1971; 9: 97-113.

Oliveira L, Ladouceur CD, Phillips ML, Brammer M, Mourão-Miranda J. What Does Brain Response to Neutral Faces Tell Us about Major Depression? Evidence from Machine Learning and fMRI. PloS one 2013; 8: e60121.

Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience & Biobehavioral Reviews 2012; 36: 1140-1152.

Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007; 69: 659-677.

Passarotti AM, Sweeney JA, Pavuluri MN. Neural correlates of response inhibition in pediatric bipolar disorder and attention deficit hyperactivity disorder. Psychiatry Research 2010; 181: 36-43.

Paus T, Keshavan M, Giedd JN. Why do many psychiatric disorders emerge during adolescence? Nat Rev Neurosci 2008; 9: 947-957.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 2006; 442: 1042-1045.

Peterson BS, Potenza MN, Wang Z, Zhu H, Martin A, Marsh R, *et al.* An FMRI study of the effects of psychostimulants on default-mode processing during Stroop task performance in youths with ADHD. The American Journal Of Psychiatry 2009; 166: 1286-1294.

Pizzagalli DA, Holmes AJ, Dillon DG, Goetz EL, Birk JL, Bogdan R, *et al.* Reduced caudate and nucleus accumbens response to rewards in unmedicated subjects with major depressive disorder. The American journal of psychiatry 2009; 166: 702.

Plant C, Teipel SJ, Oswald A, Böhm C, Meindl T, Mourão-Miranda J, *et al.* Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. Neuroimage 2010; 50: 162-174.

Plessen KJ, Bansal R, Zhu H, Whiteman R, Amat J, Quackenbush GA, *et al.* Hippocampus and amygdala morphology in attention-deficit/hyperactivity disorder. Archives Of General Psychiatry 2006; 63: 795-807.

Plichta MM, Scheres A. Ventral-striatal responsiveness during reward anticipation in ADHD and its relation to trait impulsivity in the healthy population: A meta-analytic review of the fMRI literature. Neuroscience & Biobehavioral Reviews 2013.

Pliszka SR, Glahn DC, Semrud-Clikeman M, Franklin C, Perez III R, Xiong J, *et al.* Neuroimaging of inhibitory control areas in children with attention deficit hyperactivity disorder who were treatment naive or in long-term treatment. The American Journal Of Psychiatry 2006; 163: 1052-1060.

Polanczyk G, de Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. The American Journal Of Psychiatry 2007; 164: 942-948.

Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning: MIT Press, 2006.

Rasmussen PM, Madsen KH, Lund TE, Hansen LK. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. Neuroimage 2011; 55: 1120-1131.

Rauch SL, Phillips KA, Segal E, Makris N, Shin LM, Whalen PJ, *et al.* A preliminary morphometric magnetic resonance imaging study of regional brain volumes in body dysmorphic disorder. Psychiatry Research: Neuroimaging 2003; 122: 13-19.

Riccio CA, Reynolds CR, Lowe P, Moore JJ. The continuous performance test: a window on the neural substrates for attention? Archives of Clinical Neuropsychology 2002; 17: 235-272.

Rubia K, Cubillo A, Smith AB, Woolley J, Heyman I, Brammer MJ. Disorder-specific dysfunction in right inferior prefrontal cortex during two inhibition tasks in boys with attention-deficit hyperactivity disorder compared to boys with obsessive-compulsive disorder. Human Brain Mapping 2010a; 31: 287-299.

Rubia K, Halari R, Cubillo A, Mohammad A-M, Brammer M, Taylor E. Methylphenidate normalises activation and functional connectivity deficits in attention and motivation networks in medication-naive children with ADHD during a rewarded continuous performance task. Neuropharmacology 2009a; 57: 640-652.

Rubia K, Halari R, Cubillo A, Mohammad A-M, Scott S, Brammer M. Disorder-specific inferior prefrontal hypofunction in boys with pure attention-deficit/hyperactivity disorder compared to boys with pure conduct disorder during cognitive flexibility. Human Brain Mapping 2010b; 31: 1823-1833.

Rubia K, Halari R, Smith AB, Mohammad M, Scott S, Brammer MJ. Shared and disorder-specific prefrontal abnormalities in boys with pure attention-deficit/hyperactivity disorder compared to boys with pure CD during interference inhibition and attention allocation. Journal Of Child Psychology And Psychiatry, And Allied Disciplines 2009b; 50: 669-678.

Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics (Oxford, England) 2007; 23: 2507-2517.

Sapolsky RM. Depression, antidepressants, and the shrinking hippocampus. Proceedings of the National Academy of Sciences 2001; 98: 12320-12322.

Sartorius A, Henn FA. Deep brain stimulation of the lateral habenula in treatment resistant major depression. Medical hypotheses 2007; 69: 1305-1308.

Sato JR, Hoexter MQ, Fujita A, Rohde LA. Evaluation of pattern recognition and feature extraction methods in ADHD prediction. Frontiers in systems neuroscience 2012; 6.

Schneider B, Prvulovic D. Novel biomarkers in major depression. Current opinion in psychiatry 2013; 26: 47-53.

Schneider S, Unnewehr S, Margraf J. Diagnostisches interview psychischer Störungen im Kindes- und Jugendalter (Kinder-DIPS). Heidelberg: Springer, 2009.

Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the Support of a High-Dimensional Distribution. Neural Computation 2001; 13: 1443-1471.

Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, *et al.* PRoNTo: Pattern Recognition for Neuroimaging Toolbox. Neuroinformatics 2013.

Schwaighofer A. SVM toolbox, 2001.

Seidman LJ, Valera EM, Makris N. Structural brain imaging of attention-deficit/hyperactivity disorder. Biological Psychiatry 2005; 57: 1263-1272.

Shafritz KM, Marchione KE, Gore JC, Shaywitz SE, Shaywitz BA. The effects of methylphenidate on neural systems of attention in attention deficit hyperactivity disorder. The American Journal Of Psychiatry 2004; 161: 1990-1997.

Shah PJ, Ebmeier KP, Glabus MF, Goodwin GM. Cortical grey matter reductions associated with treatment-resistant chronic unipolar depression. Controlled magnetic resonance imaging study. The British journal of psychiatry 1998; 172: 527-532.

Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis: Cambridge University Press, 2004.

Sheehan DV, Lecrubier Y. Mini-International Psychiatric Interview. MINI PLUS. English Version 1992; 5.

Sheline YI, Wang PW, Gado MH, Csernansky JG, Vannier MW. Hippocampal atrophy in recurrent major depression. Proceedings of the National Academy of Sciences 1996; 93: 3908-3913.

Shen L, Kim S, Qi Y, Inlow M, Swaminathan S, Nho K, *et al.* Identifying Neuroimaging and Proteomic Biomarkers for MCI and AD via the Elastic Net. In: Liu T, Shen D, Ibanez L and Tao X, editors. Multimodal Brain Image Analysis. Vol 7012. Berlin/Heidelberg: Springer-Verlag, 2011: 27-34.

Sladky R, Friston KJ, Trostl J, Cunnington R, Moser E, Windischberger C. Slice-timing effects and their correction in functional MRI. Neuroimage 2011; 58: 588-594.

Slifer KJ, Koontz KL, Cataldo MF. OPERANT-CONTINGENCY-BASED PREPARATION OF CHILDREN FOR FUNCTIONAL MAGNETIC RESONANCE IMAGING. Journal of applied behavior analysis 2002; 35: 191-194.

Slotnick SD, Moo LR, Segal JB, Hart J, Jr. Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. Brain Research. Cognitive Brain Research 2003; 17: 75-82.

Slotnick SD, Schacter DL. The nature of memory related activity in early visual areas. Neuropsychologia 2006; 44: 2874-2886.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004; 14: 199-222.

Smoski MJ, Felder J, Bizzell J, Green SR, Ernst M, Lynch TR, *et al.* fMRI of alterations in reward selection, anticipation, and feedback in major depressive disorder. Journal of affective disorders 2009; 118: 69-78.

Solanto MV, Schulz KP, Fan J, Tang CY, Newcorn JH. Event-related FMRI of inhibitory control in the predominantly inattentive and combined subtypes of ADHD. Journal Of Neuroimaging: Official Journal Of The American Society Of Neuroimaging 2009; 19: 205-212.

Solmaz B, Dey S, Ravishankar Rao A, Shah M. ADHD classification using bag of words approach on network features Proc. SPIE 8314. San Diego, California, USA, 2012.

Somol P, Pudil P, Novovičová J, Paclík P. Adaptive floating search methods in feature selection. Pattern Recognition Letters 1999; 20: 1157-1163.

Song S, Zhan Z, Long Z, Zhang J, Yao L. Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. Plos One 2011; 6: e17191-e17191.

Steele JD, Bastin ME, Wardlaw JM, Ebmeier KP. Possible structural abnormality of the brainstem in unipolar depressive illness: a transcranial ultrasound and diffusion tensor magnetic resonance imaging study. Journal of Neurology, Neurosurgery & Psychiatry 2005; 76: 1510-1515.

Stephan KE, Friston KJ, Frith CD. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. Schizophrenia Bulletin 2009; 35: 509-527.

Stonnington CM, Chu C, Klöppel S, Jack Jr. CR, Ashburner J, Frackowiak RSJ. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 2010; 51: 1405-1413.

Suskauer SJ, Simmonds DJ, Fotedar S, Blankner JG, Pekar JJ, Denckla MB, *et al.* Functional magnetic resonance imaging evidence for abnormalities in response selection in attention deficit hyperactivity disorder: differences in activation associated with response inhibition but not habitual motor response. Journal Of Cognitive Neuroscience 2008; 20: 478-493.

Suzuki M, Nohara S, Hagino H, Kurokawa K, Yotsutsuji T, Kawasaki Y, *et al.* Regional changes in brain gray and white matter in patients with schizophrenia demonstrated with voxel-based analysis of MRI. Schizophrenia Research 2002; 55: 41-54.

Takahashi T, Yücel M, Lorenzetti V, Walterfang M, Kawasaki Y, Whittle S, *et al.* An MRI study of the superior temporal subregions in patients with current and past major depression. Progress in Neuro-Psychopharmacology and Biological Psychiatry 2010; 34: 98-103.

Talairach J, Tournoux P. Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging. New York: Thieme Medical Publishers, 1988.

Termenon M, Graña M, Besga A, Echeveste J, Pérez JM, Gonzalez-Pinto A. Diagnosis of Bipolar Disorder Based on Principal Component Analysis and SVM. Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013: Springer, 2013: 569-578.

Tessner KD, Walker EF, Dhruv SH, Hochman K, Hamann S. The relation of cortisol levels with hippocampus volumes under baseline and challenge conditions. Brain research 2007; 1179: 70-78.

Theodoridis S, Koutroumbas K. Pattern recognition: Elsevier/Academic Press, 2006.

Tipping ME. Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 2001; 1: 211-244.

Tripp G, Wickens JR. Research review: dopamine transfer deficit: a neurobiological theory of altered reinforcement mechanisms in ADHD. Journal Of Child Psychology And Psychiatry, And Allied Disciplines 2008; 49: 691-704.

Tsujimoto S, Genovesio A, Wise SP. Frontal pole cortex: encoding ends at the end of the endbrain. Trends in Cognitive Sciences 2011; 15: 169-176.

Tukey JW. Exploratory data analysis. Reading, MA 1977; 231.

Vaidya CJ, Bunge SA, Dudukovic NM, Zalecki CA, Elliott GR, Gabrieli JDE. Altered neural substrates of cognitive control in childhood ADHD: evidence from functional magnetic resonance imaging. The American Journal Of Psychiatry 2005; 162: 1605-1613.

Vakili K, Pillay SS, Lafer B, Fava M, Renshaw PF, Bonello-Cintron CM, *et al.* Hippocampal volume in primary unipolar major depression: a magnetic resonance imaging study. Biological Psychiatry 2000; 47: 1087-1090.

van 't Ent D, van Beijsterveldt CEM, Derks EM, Hudziak JJ, Veltman DJ, Todd RD, *et al.* Neuroimaging of response interference in twins concordant or discordant for inattention and hyperactivity symptoms. Neuroscience 2009; 164: 16-29.

Vapnik V, Golowich SE, Smola A. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems 1997; 9: 281-287.

Vapnik VN. The nature of statistical learning theory: Springer, 1995.

Vapnik VN. Statistical Learning Theory: Wiley, 1998.

Voeller KK. Toward a neurobiologic nosology of attention deficit hyperactivity disorder. Journal Of Child Neurology 1991; 6 Suppl: S2-S8.

Volkow ND, Wang G-J, Newcorn J, Fowler JS, Telang F, Solanto MV, *et al.* Brain dopamine transporter levels in treatment and drug naive adults with ADHD. Neuroimage 2007a; 34: 1182-1190.

Volkow ND, Wang G-J, Newcorn J, Telang F, Solanto MV, Fowler JS, *et al.* Depressed dopamine activity in caudate and preliminary evidence of limbic involvement in adults with attention-deficit/hyperactivity disorder. Archives Of General Psychiatry 2007b; 64: 932-940.

Wang L, Zhu C, He Y, Zang Y, Cao Q, Zhang H, *et al.* Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. Human Brain Mapping 2009; 30: 638-649.

Wang Y, Fan Y, Bhatt P, Davatzikos C. High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. Neuroimage 2010; 50: 1519-1535.

Watkins AE, Scheaffer RL, Cobb GW. Statistics: From Data to Decision: John Wiley & Sons, 2009.

Wechsler D. WISC-III: Wechsler intelligence scale for children. San Antonio, TX: Psychological Corporatio, Harcourt Brace Jovanovich, 1991.

Wechsler D. The Wechsler intelligence scale for children. London: Pearson Assessment, 2004.

Weiß RH. Grundintelligenztest Skala 2 (CFT 20) mit Wortschatztest (WS) und Zahlenfolgentest (ZF). Braunschweig: Westermann, 1998.

Yerys BE, Jankowski KF, Shook D, Rosenberger LR, Barnes KA, Berl MM, *et al.* The fMRI success rate of children and adolescents: typical development, epilepsy, attention deficit/hyperactivity disorder, and autism spectrum disorders. Human Brain Mapping 2009; 30: 3426-3435.

Yu Y, Shen H, Zeng L-L, Ma Q, Hu D. Convergent and Divergent Functional Connectivity Patterns in Schizophrenia and Depression. PLOS ONE 2013; 8: e68250.

Zang Y-F, He Y, Zhu C-Z, Cao Q-J, Sui M-Q, Liang M, *et al.* Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. Brain & Development 2007; 29: 83-91.

Zeng L-L, Shen H, Liu L, Wang L, Li B, Fang P, *et al.* Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 2012; 135: 1498-1507.

Zeng LL, Shen H, Liu L, Hu D. Unsupervised classification of major depression using functional connectivity MRI. Human brain mapping 2013.

Zhang W-N, Chang S-H, Guo L-Y, Zhang K-L, Wang J. The neural correlates of reward-related processing in major depressive disorder: A meta-analysis of functional magnetic resonance imaging studies. Journal of affective disorders 2013.

Zhu C-Z, Zang Y-F, Cao Q-J, Yan C-G, He Y, Jiang T-Z, *et al.* Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. Neuroimage 2008; 40: 110-120.

Zhu CZ, Zang YF, Liang M, Tian LX, He Y, Li XB, *et al.* Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder. Medical Image Computing And Computer-Assisted Intervention: MICCAI 2005; 8: 468-475.

Zhu X, Wang X, Xiao J, Zhong M, Liao J, Yao S. Altered white matter integrity in first-episode, treatment-naive young adults with major depressive disorder: a tract-based spatial statistics study. Brain research 2011; 1369: 223-229.

Zou K, Huang X, Li T, Gong Q, Li Z, Ou-yang L, *et al.* Alterations of white matter integrity in adults with major depressive disorder: a magnetic resonance imaging study. Journal of psychiatry & neuroscience: JPN 2008; 33: 525.