Challenges in Resting State Functional Connectivity Analysis: Removal of Head Motion Artifacts and Machine Learning-based Disease Classification

by

Pradyumna Lanka

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 06, 2017

Keywords: Functional MRI, Functional connectivity, Motion artifacts, Prospective motion correction, Disease classification, Machine Learning

Approved by

Gopikrishna Deshpande, Chair, Associate Professor, Electrical & Computer Engineering
Thomas S. Denney, Director, Auburn University MRI Research Center
Stanley J. Reeves, Professor, Electrical & Computer Engineering

Abstract

In recent years, the use of resting-state functional Magnetic Resonance Imaging (rs-fMRI) for examining the brain function in healthy and clinical populations has increased drastically. Simultaneous modulations in neural activity between remote brain regions when the subject does not perform an explicit task, also called resting state functional connectivity (RSFC) are associated with the presence or absence of neurological disorders. However, some challenges remain to be addressed for widespread use of RSFC as a tool for disease classification. In this thesis, we address two crucial issues associated with RSFC. In the first part of this thesis, we examine how in-scanner head motion can cause artifactual changes in RSFC and evaluate the utility of an image based prospective motion correction in reducing the head motion artifacts in RSFC derived metrics. Our results indicate that the use of prospective motion correction combined with commonly used retrospective motion correction methods was able to visibly reduce the artifactual changes in RSFC. In the second part of this thesis, we examine the issues associated with the use of RSFC for disease diagnosis. Specifically, we evaluate how variations in age ranges and the data acquisition site of the sample can affect the performance of machine learning classifiers especially in heterogeneous disease populations with small sample sizes. We observe that the use of small, homogenous subject samples might give inflated measures of accuracy possibly due to overfitting. Finally, we recommend the use of a hold-out test data or a replication dataset to reproduce the classification performances to ensure good generalization across the disease population.

Acknowledgments

I would like to thank several people who contributed their time and effort and played a role in the successful completion of my thesis. First, I would also like to express my sincere gratitude to my graduate adviser Dr. Gopikrishna Deshpande for his guidance and support throughout my master's program. He has been an inspiration to me and helped me become a better researcher. I am certainly indebted to him for his efforts in reviewing my thesis. I would certainly thank my committee members, Dr. Thomas Denney and Dr. Stanley Reeves for graciously agreeing to serve on my committee. The courses they offered helped me immensely to understand the intricacies of image/signal processing and MRI which served as a foundation for my research. I would also thank Dr. Bogdan Wilamowski for inculcating my interest in machine learning and neural networks.

I would also like to thank all my friends at the Auburn University MRI Center, for their help and support during my research. I always enjoyed having intellectually stimulating conversations with them. My thesis would not have been possible without large-scale neuroimaging data sharing initiative like Autism Brain Imaging Data Exchange (ABIDE), Alzheimer's Disease Neuroimaging Initiative (ADNI), Attention Deficit Hyperactivity Disorder (ADHD) Consortium, whose data has been invaluable for my research. Finally, I would like to thank my family for their unconditional love and support and for believing in my abilities.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Magnetic Resonance Imaging (MRI)

The human brain is one of the most complex structures known to us, with many of its mysteries yet to be revealed. Probably one of the most significant breakthroughs in the examination the human body in general and the human brain, in particular, came from Magnetic Resonance Imaging. The idea of Nuclear magnetic resonance (NMR) in solids was first discovered by two nuclear physicists Edward Purcell [1] and Felix Bloch [2]. The principles of NMR later were later extended to imaging by Paul Lauterbur and Peter Mansfield who subsequently won the 2003 Nobel Prize in medicine [3]. It has since been used extensively in clinical applications for disease diagnosis.

There are three steps in creating a Magnetic Resonance Image. The first step is producing the signal by spin excitation of the Hydrogen H^1 nucleus (protons) in the human body. The second step is to localize the signal to its source. Finally, the third step involves generating image contrast to differentiate the various tissues in the body [4].

The protons in the human body have a zero net magnetic moment as they are directed arbitrarily in the absence of a magnetic field. A large static magnetic field (B0) formed by a huge solenoid shaped electromagnet in the MRI scanner forces all protons to align in a direction parallel or antiparallel to the field's direction. This strength of the static magnetic field usually 1.5T, 3T or 7T, ultimately determines the SNR of the resulting image. Most protons align in the parallel direction than antiparallel, about the static magnetic field and precess around their axis like little spinning tops at a rate determined by Larmor frequency, which in turn is dependent on the

nucleus of the atoms and the static magnetic field strength (B0). This random alignment of protons creates a net non-zero magnetic moment in the direction parallel to B0 field.

A Radio Frequency (RF) magnetic pulse (denoted by B1), is applied in a direction perpendicular to the B0 field at the Larmor Frequency of a nucleus, usually hydrogen, due to their abundance in the human body. The protons absorb the energy and rotation axis of the protons is flipped to an angle called the Flip Angle (FA). Once the RF pulse is turned off, the proton loses its energy to its environment and tries to get back to its original state at the rate known as longitudinal (T1) relaxation rate R1 (1/T1). Once the proton is tipped over, there is a net magnetic moment in the direction perpendicular to the static magnetic field. The proton then interacts with neighboring protons (spin-spin interaction) and goes out of phase with respect to its neighbors, at the rate given by the transverse (T2) relaxation rate R2 (1/T2) of the nucleus, thus reducing the magnetic moment to zero in the transverse direction. The net magnetic moment of the nucleus in the transverse direction is detected as a signal by the receiver coil. However, the spin-spin interaction is not the only way in which the electron loses its transverse component of the magnetic moment. Local magnetic inhomogeneity can cause the spin of the nucleus to decay faster, with a decay rate given by $R2^*$ ($1/T2^*$).

The three gradient coils achieve the localization of the signals to specific regions in the human body, in the three directions. These gradient coils create small differences in the net magnetic field strength in the body across their respective directions so that each location in the 3D space experience a unique field strength. This gradation in the magnetic field causes the protons at each location to precess at slightly varying speed. By matching the band of frequencies of the RF pulse to the frequency of the region, we can obtain the signal from only that particular region in the tissue, thereby localizing the source of the signal to a specific region in the body.

Continuously change the RF pulse of the transmit coil to the whole gamut of frequencies across all the magnetic field gradients in a magnetic field, we can obtain a three-dimensional volume of the entire tissue.

The domain the data is collected in MRI is called k-space. Since the data is obtained as a net collection of frequencies across the spectrum, a Fourier Transform can be used to separate the different frequency components, each of which represents a signal from every region in the tissue. The intensity of the signal obtained depends on Transverse Relaxation rate, longitudinal relaxation rate, and the proton density. Since different tissues in the human body have different relaxation times and proton densities, this factor can be exploited to create the appropriate contrast between various tissues in the human body. Image contrasts help in separating tissues from their surroundings, such as detecting cancerous cells from non-cancerous cells, identifying gray matter from white matter in the brain, etc.

1.2 fMRI & BOLD contrast

FMRI is an indirect measure of neural activity and can be used to infer the function of the brain over time [5]. In fMRI, we rapidly collect data from human brain at smaller time intervals using a pulse sequence called echo-planar imaging (EPI). EPI requires rapid changes in the gradient coils to collect all the data from the brain in a few seconds. The neural activity of the brain is inferred from Blood Oxygenation Level Dependent (BOLD) contrast which was developed by Seiji Ogawa et al. in 1990 [6]. The BOLD contrast is dependent on the differences in the magnetic properties of oxygenated and de-oxygenated blood. Oxygenated blood is diamagnetic, so it does not interact with the surrounding magnetic field, whereas de-oxygenated blood is paramagnetic which means it distorts the surrounding magnetic field (susceptibility) which reduces the T2* relaxation time, thereby reducing the MR signal. Neural activation of a region in

the brain increases its metabolic demand, causing a slight dip in the oxygenated blood relative to the deoxygenated blood, resulting in a drop the MR signal. This reduction in the oxygenated blood prompts a substantial increase in the flow of the blood, rich with oxygen to the region. This rush of oxygenated blood causes an increase in the signal which usually peaks at around 5 seconds following the neural response. The signal then falls, until the relative concentrations of the oxygenated and the deoxygenated blood gets back to their previous level. The changes in MR signal in response to the neural activity is called Hemodynamic response (HR) and can by modeled by a hemodynamic response function (HRF). In fact, the BOLD response can be modeled as a convolution of the underlying neuronal activity with the HRF. Due to the slow hemodynamic response (HR), sometimes the fMRI time series is deconvolved to identify the underlying neural activity [7].

FMRI captures a snapshot of the brain across time repetition time (TR). Functional MRI provides excellent spatial resolution compared to EEG but poor temporal resolution due to the hemodynamic response and a long TR [8]. However, larger static field strengths (B0) provide a decent compromise between the SNR, spatial resolution, and the temporal resolution.

1.3 Resting-state fMRI

Resting-state fMRI (Rs-fMRI) measures the spontaneous fluctuations in the MR signals of the brain when the person is not performing an explicit task. Regions of the brain do not work independently, they work in coordination with each other and organize to form of networks. Hence the field of neuroimaging has moved neuropsychological localization in the early 1990s to more connectionist approaches lately [9]. These networks can be identified by connectivity models obtained from Resting-state fMRI. The extent to which two brain regions are co-activated can be determined using the Pearson's correlation coefficient between their fMRI time-

series. Several brain networks in low-frequency have been identified so far since the discovery of the Sensory-motor Network [10]. The default-mode network, dorsal attention network, sensory-motor network, visual processing network, auditory-phonological network and self-referential network, are some of the Resting state Networks (RSN) [11]. Resting state functional connectivity can be used a proxy for identifying the regions in the brain that co-activate with a seed region for several tasks tend to be positively correlated with each other at rest. The resting state fMRI connectivity have been shown to be sensitive to age and neurological disorders [12, 13, 14]. Hence the search for imaging-based biomarkers remains the holy grail of resting state fMRI connectivity due to its ease of acquiring data and its insensitivity to task performance.

1.4 Machine Learning

There are several definitions for machine learning. According to Tom Mitchell, "A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E [15]." The field of machine learning draws inspiration from Artificial Intelligence, Statistics and Pattern Recognition. Machine learning can be broadly be divided into three categories Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Supervised machine learning can expressed as a set of inputs and outputs, where the performance can be evaluated as correctly predicting the output for unknown sets of inputs. Supervised learning algorithms are further divided into classification when the outputs are discrete and Regression when the outputs are continuous. In this thesis, our focus is on supervised classification methods.

Each observation is described by a set of attributes/predictors/features. Supervised learning can be defined as learning or optimizing the model function $\boldsymbol{y = f(w, x)}$ that maps the relationship between the inputs *x* and outputs *y* by estimating the parameters *w*. We can then use the model to

predict the outputs for unseen inputs. The algorithm which is used to learn the model form the training data is called a learning algorithm. A machine learning model, can be obtained from different learning algorithms which make different assumptions of the nature of the data. The data which is used to build the model is called training data. Unfortunately, since in a typical learning procedure we try to minimize the cost function on the training data, the function might actually learn the noise in the training data as well, hence the performance of model on the training obtained might not generalize well to the unseen data. Hence a separate dataset, also called a hold-out test data is generally used for obtaining an unbiased estimate of the performance of the model.

Hyperparameters are a set of knobs that determine the model of the algorithm. They are set before the learning commences, unlike model parameters whose values will be estimated by the learning algorithm. Grid search is the easiest way to estimate the optimal hyperparameters for the model. We used cross-validation to measure the generalization accuracy and chose the model which has the best performance to determine the optimum set of hyperparameters.

1.5 Thesis organization

Two of the most important issues in resting state fMRI connectivity are the artifacts associated with head motion and disease diagnosis using resting state fMRI connectivity metrics. Because of the growing interests in the applications of resting state fMRI connectivity in recent years and the issues limiting its widespread appeal, we decide to tackle these two issues in this thesis. The first chapter gives a general introduction of MRI and fMRI, along with defining a few concepts in machine learning. We introduce some basic principles of signal generation in MRI and fMRI. We also discuss the meaning of a few terms commonly encountered in machine learning. In the second chapter, we discuss the effectiveness of prospective motion correction in correcting

motion artifacts. We first define the nature of the motion artifacts associated with in-scanner head motion in 47 subjects scanned with Prospective Acquisition CorrEction (PACE) sequence. Using metrics derived from resting state fMRI connectivity we evaluate in detail the effectiveness of the PACE sequence in reducing motion artifacts in the rs-fMRI data. In the third chapter, we discuss how age and site variability might affect the classification performance, especially in small datasets. Using neuroimaging data from 4 datasets and applying 18 different machine learning algorithms, we show how the classifiers might overfit homogeneous data to give inflated measures of accuracy. These performance measures, unfortunately, do not generalize well to the general population. We summarize our findings in the Conclusion in Chapter 5. Finally, we explain the machine learning classifiers we used in the appendix A.

1.6 Bibliography

[1] E. M. Purcell, H. C. Torrey and R. V. Pound, "Resonance Absorption by Nuclear Magnetic Moments in a Solid," *Physical Review,* vol. 69, pp. 37-38, 1946.

[2] F. Bloch, W. W. Hansen and M. Packard, "Nuclear Induction," *Physical Review,* vol. 69, p. 127, 1946.

[3] A. Filler, The History, Development and Impact of Computed Imaging in Neurological Diagnosis and Neurosurgery: CT, MRI, and DTI., Avialable form Nature Preceedings <http://hdl.handle.net/10101/npre.2009.3267.1>, 2009.

[4] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, . M. R. Thompson and R. Venkatesan, Magnetic Resonance Imaging: Physical Principles and Sequence Design, 2nd Edition, Wiley-Blackwell, 2014.

[5] S. A. Huettel, A. W. Song and G. McCarthy, Functional Magnetic Resonance Imaging, Third Edition, Sunderland, MA, USA: Sinauer Associates, Inc., 2014.

[6] S. Ogawa, T. M. Lee, A. R. Kay and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *PNAS,* vol. 87, pp. 9868-9872, 1990.

[7] G.-R. Wu, W. Liao, S. Stramaglia, J.-R. Ding, H. Chen and D. Marinazzo, "A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data," *Medical Image Analysis,* vol. 17, no. 3, pp. 365-374, 2013.

[8] S. Chen and X. Li, "Functional Magnetic Resonance Imaging for Imaging Neural Activity in the Human Brain: The Annual Progress," *Computational and Mathematical Methods in Medicine,* vol. 2012, no. 613465, p. 9, 2012.

[9] S. L. Small, "Connectionist networks and language disorders," *Journal of Communication Disorders,* vol. 27, no. 4, pp. 305-323, 1994.

[10] B. Biswal, Zerrin Yetkin, F., Haughton, Victor M. and Hyde, James S., "Functional connectivity in the motor cortex of resting human brain using echo-planar mri," *Magn. Reson. Med.,* vol. 34, no. 4, pp. 537-541, 1995.

[11] M. P. van den Heuvel and H. E. Hulshoff Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity, European Neuropsychopharmacology," *European Neuropsychopharmacology,* vol. 20, no. 8, pp. 519-534, 2010.

[12] D. Pinter, C. Beckmann, M. Koini, E. Pirker, N. Filippini, A. Pichler, S. Fuchs, F. Fazekas and C. Enzinger, "Reproducibility of Resting State Connectivity in Patients with Stable Multiple Sclerosis," *PLOS ONE,* vol. 11, no. 3, p. e0152158, 2016.

[13] J. Damoiseaux, C. Beckmann, E. Sanz Arigita, F. Barkhof, P. Scheltens, C. Stam, S. Smith and S. Rombouts, "Reduced resting-state brain activity in the “default network” in normal aging," *Cerebral Cortex,* vol. 18, no. 8, pp. 1856-1864, 2008.

[14] J. S. Damoiseaux, "Resting-state fMRI as a biomarker for Alzheimer's disease?," *Alzheimer's Research & Therapy,* vol. 4, no. 2, p. 8, 2012.

[15] T. M. Mitchell, Machine Learning, New York, USA: McGraw-Hill, Inc, 1997.

Chapter 2

An evaluation of the effectiveness of Prospective Acquisition CorrEction (PACE) for reducing motion-related artifacts in resting state fMRI data

Abstract

Resting state functional connectivity (RSFC) derived from blood oxygenation level dependent (BOLD) functional magnetic resonance imaging (fMRI) has been extensively used due to its sensitivity to brain function and its alterations in clinical populations. Head movement in the scanner causes spurious signal changes in the BOLD signal, confounding RSFC estimates. We examined the effectiveness of Prospective Acquisition Correction (PACE) in reducing motion artifacts in BOLD data. Using PACE-corrected RS-fMRI data obtained from 44 subjects and subdividing them into low and high motion cohorts, we investigated voxel-wise motion-BOLD relationships, the distance-dependent functional connectivity artifact and the correlation between head motion and connectivity metrics such as posterior cingulate seed based connectivity and degree centrality. Our results indicate that, when PACE is used in combination with standard retrospective motion correction strategies, it provides two principal advantages over conventional echo-planar imaging (EPI) RS-fMRI data: (i) PACE was effective in eliminating significant negative motion-BOLD relationships, shown to be associated with signal dropouts caused by head motion, and (ii) Censoring with a lower threshold (frame-wise displacement > 0.5mm) and a smaller window around the motion corrupted time-point provided qualitatively equivalent reductions in the motion artifact with PACE when compared to a more conservative threshold of 0.2 mm required with conventional EPI data. This will likely provide substantial savings in data which would otherwise be lost to censoring. Given that PACE is available as an option in the EPI product sequence provided by Siemens, it has negligible overhead in terms of scan time, sequence modifications or additional setup (which are typical of other prospective

motion correction methods), and hence presents an attractive option for head motion correction in high throughput resting state BOLD imaging.

2.1 Introduction

Head motion is one of the major sources of artifacts in functional Magnetic Resonance Imaging (fMRI). Head motion is said to cause large spatially varying signal changes across the brain. Realignment corrects the changes in brain position, but it does not take into consideration the changes in the image intensity associated with motion. Head motion, particularly in the direction perpendicular to the slice selection is susceptible to artifacts due to magnetic field inhomogeneity and spin-excitation history effects [1].

Resting state functional connectivity measures the synchronicity of the brain activity in different regions of the brain and has become quite popular in the last decade due to its sensitivity to development, aging and pathology. However, motion can severely affect the validity of resting state fMRI (RS-fMRI) studies, particularly in hyperkinetic populations as the motion induced variance changes could potentially drive resting state functional connectivity metrics in the same direction as one would expect due to disease or aging.

Most motion correction approaches are typically classified into prospective motion correction and retrospective motion correction. In prospective motion correction, the motion is corrected for before or during the acquisition of the volumes, whereas retrospective motion correction methods correct for motion after the acquisition of the volumes. Rigid-body realignment, nuisance signal regression, modeling the effects of the head motion on the blood oxygenation level dependent (BOLD) signal using motion parameters and removing the fitted response, temporal band pass filtering, motion censoring or spike regression, group level correction or

some combinations of the above approaches are routinely used with varying degrees of success in retrospective motion correction [2, 3, 4]

Realignment is the first step in retrospective motion correction. Though it aims to make each voxel correspond to the same region in the brain in the fMRI time series by selecting a suitable rigid-body transformation, it does not eliminate the changes in the image intensity associated with head motion. In nuisance signal regression, we regress out the mean signal corresponding to areas of Cerebrospinal fluid (CSF), White Matter (WM) and the GS (Global Signal) signals from the BOLD signal, with the hope that we would remove the variance of non-neural origin attributable to head motion, heartbeat, respiration and other sources of scanner noise. WM and CSF signals, though commonly used, fail to remove a significant amount of variance in the BOLD signal due to head motion [2, 4]. Global signal regression (GSR) has been reported in previous studies to remove significant amount of variance associated with head motion from the BOLD signal [2, 4, 3]. However, there have been some concerns that GSR introduces an artificial negative bias in the correlation coefficients, driving them downward everywhere in the brain to the point of introducing spurious anti-correlations [5] and is also said to create artifactual group differences in functional connectivity [6].

Spin History effects result in the signal intensity of the current acquisition to become a complex non-linear function of the current position, as well as previous positions [1]. In order to address this, one could model the effects of head movement on the BOLD signal by using a second order polynomial containing the current motion parameters and few previous (in time) motion parameters and remove the fitted response from the BOLD signal. Either volume based parameters or voxel-Specific motion parameters have been used, though no significant benefit in using voxel-specific parameters have been reported [3, 2]. 24 parameters (R_t, R_t^2, R_{t-1}, R_{t-1}^2) and

36 parameters (R_t, R_t^2, R_{t-1}, R_{t-1}^2, R_{t-2}, R_{t-2}^2) models, where *R* indicates the 6 realignment parameters and *t* indicates the number of volumes back in time used to correct for motion. The justification for using complex models is that it increases the fit compared to models using a lower number of parameters [3]. However, going further back in time results in loss of degrees of freedom and may sometimes not justify the increase in fit. Also, fewer number of parameters might be appropriate for low-motion subjects as increasing the number of parameters causes the model to over-fit the data and thereby reduce the sensitivity to the underlying neural activity [3].

Regressions of nuisance signals derived from WM/CSF and the motion parameters cannot effectively model motion induced variance in fMRI time-series [2, 4, 3]. Alternatively, in motion censoring or scrubbing, the motion corrupted time points and the adjacent time points that exceed a threshold defined by a QC (quality control) metric derived from the data such as DVARS (Derivative of root mean squared variance over voxels) or Frame-wise Displacement (FD) [7, 4, 8], are marked. Either the data for those time points are removed, or they are interpolated from the adjacent time points and removed after preprocessing. Similar to scrubbing, spike regression models the motion induced spikes in fMRI data and removes the fitted response, effectively eliminating the influence of the corrupted time points in the fMRI time series [3, 9]. Scrubbing/Censoring/Spike Regression, especially in high-motion datasets could potentially lead to loss of a large quantity of data [10] and in turn result in noisy estimates of the functional connectivity [11]. Further, scrubbing could introduce discontinuities in the data which may invalidate many analysis methods used thereafter. Even though interpolation has been suggested as a way to avoid these discontinuities, interpolation is at best a guess and still amounts to loss of original data. Filtering the signal in the frequency band of 0.008/0.01 - 0.08/0.1 Hz eliminates

frequencies corrupted by respiration and other artifacts and might improve sensitivity to the underlying neuronal fluctuations.

A more recent approach is retrospective correction at the group level where in the mean subject head motion is regressed out by adding it as a nuisance variable in the second-level (group level) general linear model (GLM) by removing within-group/between-group variance which can be attributed to the differences in head motion of the subjects in the groups. So within-group regression can be used to eliminate the assumed linear effects of head motion on functional connectivity. However, group level correction by using mean subject motion has its drawbacks. The subject's head motion may co-vary with factors of interest such as age and disease, thereby underestimating the relationship between functional connectivity and the factors of interest. A summary of currently available retrospective motion correction methods and their effectiveness can be found in [12, 8].

Given the difficulty in properly modeling the motion effects on the BOLD signal, prospective motion correction methods have gained increasing prominence. Although prospective motion correction has been in vogue for more than a decade, recent research (the flurry of articles that have appeared since Power et al. [7]) demonstrating the inadequacy of retrospective methods (most glaringly the rigid body realignment approach) suggests that it is imperative to evaluate prospective motion correction approaches in the context of motion effects on resting state functional connectivity. Most prospective methods estimate the position of the head during scanning by using external tracking devices [13, 14, 15, 16, 17]. A review of prospective motion correction methods in fMRI can be found in Zaitev et al. [18]. Since these methods use an external device to independently record head motion and correct the gradients in near real-time, they require elaborate setups, the subjects to wear a "marker" and sequence modification.

Consequently, they are unsuitable for high throughput routine scanning. Alternatively, Prospective Acquisition CorrEction (PACE) [19] is an image based online motion detection and correction sequence which tracks the subject's head location to keep the position of the head fixed relative to the scanner' coordinate frame thereby reducing spin history effects associated with head motion [19]. Using an image-based motion detection algorithm, the head motion parameters are estimated and fed back into the scanner so that the slice positioning and orientation are adjusted before the acquisition of the next volume. PACE accounts for motion based on the current volume realignment parameters and adjusts the position for the next volume acquisition of to the calculated head position by adjusting the magnetic field gradients in the gradient coils. Since the position of the previous volume is used to acquire the current volume, there is a residual motion that cannot be accounted by PACE. That said, it requires no additional setup in terms of external devices, does not require subjects to wear any "targets" and is a functionality that is in-built in FDA-approved echo-planar imaging (EPI) sequences on Siemens scanners (and hence does not require a sequence modification). For all these reasons, PACE is suitable for high throughput routine imaging and therefore is worthy of being evaluated in the context of motion-BOLD relationship.

We had three main goals for this paper. First, we were particularly interested in understanding the following effects obtained from RS-fMRI data acquired using an EPI-PACE sequence: (i) the spurious motion-BOLD relationships [2], (ii) the motion induced distance dependent functional connectivity artifact [20, 3, 7, 4] and (iii) the effect of motion on RS-fMRI connectivity based metrics such as Degree Centrality and PCC (posterior cingulate cortex) seed based FC. Second, we examined if a combination of prospective and retrospective motion correction methods could do a better job of reducing motion artifacts in BOLD data compared to using PACE alone.

Finally, we wished to determine neural correlates of head motion by performing hemodynamic deconvolution of the BOLD signal (which was corrected for motion using the best possible combination of PACE and retrospective correction) to uncover the underlying latent neural signals and then correlating it with head motion. Hemodynamic deconvolution was performed to remove the delay between head motion and the BOLD response so that their correlation would be meaningful.

2.2 Methods

2.2.1 Subjects

A total of 47 healthy adult subjects (20 males/27 females, age 25.1 ± 5 years) with no history of any neurological disorders were selected for this study. The subjects were instructed to relax, keep their eyes open, not think about anything in specific and keep their head as still as possible for the duration of the scans. Appropriate padding was provided to keep the head as still as possible in the scanner. All subjects gave informed consent, and the scanning procedure was performed in accordance with the guidelines and the approval of the Institutional Review Board at Auburn University.

2.2.2 Data Acquisition

All subjects were scanned with a 3T MAGNETOM Verio scanner (Siemens Healthcare, Erlangen, Germany) using an EPI – PACE sequence with a 32 channel head coil and the following acquisition parameters: TR of 1000 ms, TE of 29 ms, Flip Angle of 90° with 16 slices, matrix= 64×64, voxel size = 3.5×3.5×5 mm^3. The number of time points acquired for each subject ranged from 250-1000. A T1 weighted MPRAGE anatomical image (TE=2 ms,

TR=1900 ms, 176 slices with 1x1x1 mm^3 voxel size) was also acquired for all the subjects to aid in spatial normalization.

2.2.3 Preprocessing of the RS-fMRI data

The pre-processing of the RS- fMRI data was performed using Data Processing Assistant for Resting-State fMRI (DARPSF) toolbox [21]. The first five time points were removed from the time series to allow for T1 equilibration. Slice timing correction was applied to each slice in every volume to account for the different acquisition times of the slices. The volumes were then realigned using a six-parameter (three translations, three rotations) rigid body transformation to account for the head motion by optimizing the minimum squared difference (MSD) cost function by a two-pass procedure. After realignment, the T1 weighted anatomical image from each subject was registered to the mean functional image. Linear and quadratic detrending were performed to remove low-frequency drift. Mean WM and CSF signals were regressed from the time series to remove non-BOLD related signal variance. Also, the 24 parameter motion regression proposed by Friston (Friston-24) consisting of the six realignment parameters, their temporal derivatives and the squares of them, were regressed from the resting state fMRI BOLD time series.

2.2.4 Calculation of DVARS and head motion metrics

DVARS (Derivative of root mean squared variance over voxels) is the square root of mean square value of the temporal derivative of the intensities of the BOLD signal, calculated backward from the current time point to the previous time point over a voxel, ROI or the entire brain [7, 4]. Traditionally, motion metrics are calculated from the realignment parameters, and their accuracy is limited by the accuracy of the estimates of the realignment parameters.

Common metrics which capture subject head motion are the Total Displacement (TD) and Framewise Displacement (FD). TD is measure of the change in the position of the head from its initial postion, while FD is a measure of the change in the position of the head from the previous time point to the current time point and is calculated using realignment parameters of both the time points. It measures relative displacement rather than the absolute displacement of the head. In the case of motion correction by PACE, since the slice positioning is adjusted on the fly for every volume, these realignment parameters, and the FD metrics are a measure of the residual motion, relative to the scanner that is uncorrected by PACE rather than the actual motion of the head. Since all voxels in the brain do not move in a similar direction, it is essential to capture the individual movements of the voxel to understand the localized changes in signal intensities. So, along with volumetric metrics of the head's framewise displacement (FD_{vol}) which assigns a single value of head motion to the entire brain, we also calculated the voxel-specific framewise displacement (FD_{vox}) which uses the 6 realignment parameters to estimate the relative displacement of every voxel at each time-point. This enabled the computation of the displacement of each voxel with respect to the previous time point. More details on this approach are available in Satterthwaite, et al. [3], Yan, et al. [2]. The following motion metrics were calculated for each subject for every time point: FD_{FSL} [22], FD_{Power} [7], and $FD_{VanDijk}$ [23], all of which are voulume specific metrics. We also estimated the voxel-specific framewise displacement (FD_{vox}) [3]. The relationship between the different FD metrics was ascertained by plotting the subject mean of $mean_{sp}FD_{vox}$ (spatial mean of the framewise displacement across all voxels in brain for each volume) on the X-axis and the different volumetric FD metrics (FD_{FSL}, FD_{Power}, $FD_{VanDijk}$) on the Y-axis [2]. We also calculated the $mean_{sp}TD_{vox}$, which is the voxel-wise mean of the total displacement of the brain in the scanner.

2.2.5 Examination of motion -BOLD relationships in PACE data

As voxel displacement is not spatially constant across the brain due to the combination of head rotations and head translations, voxel-wise analysis of the motion-BOLD relationships would be more appropriate to study the localized effect of head motion on BOLD signal intensity. Therefore, to understand the spatially varying relationships between head motion and the BOLD signal, the BOLD signal was pre-processed with several combinations of nuisance signal regressors: (i) CSF + WM regression, (ii) CSF + WM + GS regression, (iii) CSF + WM + Friston-24 motion regression, (iv) CSF + WM + GS + Friston-24 motion regression, (v) CSF + WM + Friston-24 motion regression + motion censoring (FD_{power} threshold >0.5 mm and 1 back and 2 forward volumes regressed from the model), and (vi) CSF + WM + GS + Friston-24 motion regression + motion censoring. These pipelines were evaluated for each of the 44 subjects. Three of the subjects were eliminated because they did not have the necessary 3 minutes of data required for stable estimation of rs-fcMRI (Resting State-functional Connectivity MRI) metrics [2] after censoring. The Pearson's correlation coefficient was calculated between the voxel-specific framewise displacement (FD_{vox}) and the BOLD signal for every voxel (pre-processed using the six pipelines mentioned above), and for all the volumes in the time series as described by Yan et al. [2]. With motion censoring, the same volumes which were removed from the BOLD signal were also removed from the FD_{vox} to calculate voxel-wise correlation between motion and BOLD. Fischer's z transformation was performed on the resultant correlation maps to improve the normality of the data distribution. The resultant z -maps were then normalized to the standard MNI template (3 mm^3 cubic voxels) and the resulting volumes were smoothed with a 4.5 mm^3 Gaussian kernel. A one-sample t-test was performed on the normalized correlation maps with a significance level of $p<0.05$ (FDR corrected) to investigate consistent patterns of

motion-BOLD relationships within the group. In order to further investigate the nature of the motion-BOLD relationships for different levels of head-motion, we divided our dataset into two subsets, a higher motion group (FD_{FSL} = 0.152±0.062 mm) and a lower motion group (FD_{FSL} = 0.077±0.014 mm) containing 22 subjects each based on their mean [FD_{FSL}] with similar sex and age profiles in both the subgroups. We then proceeded to repeat the procedure described above for both the data sets separately to compare the motion-BOLD signal relationships in the high-motion dataset with the low-motion dataset.

2.2.6 Examining the motion- induced distance dependent artifact in functional connectivity

Head motion tends to distort functional connectivity metrics by inflating connectivity estimates between closer regions and reducing the connectivity between farther regions, as the voxels which are far from each other are less likely to experience similar movements, thus giving rise to the decaying effect of functional connectivity as a function of the head motion and the distance between them [7, 23, 20]. This is called as motion-induced distance dependent artifact in functional connectivity throughout the paper. Power et al. reported that online motion correction by PACE did not ameliorate the distance dependent changes in functional connectivity induced by head motion [7]. However, they did not show the corresponding results and did not elaborate it further. Therefore, to understand the effect of online motion correction on functional connectivity and to reveal the distance dependence artifact, we followed a procedure used previously [3, 20] to characterize the effects of head motion artifacts in PACE data. After preprocessing the data, the volumes were normalized to the standard MNI template (3 mm^3 cubic voxels) and were smoothed with a 4.5 mm^3 Gaussian kernel, following which, the time-series were filtered with a band pass filter with bandwidth of 0.01-0.1 HZ.

160 ROIs, as defined by the Dosenbach 160 atlas [24], were extracted from the brain. Each ROI was modeled as a sphere with 10 mm diameter and the mean resting state pre-processed BOLD signal was obtained for each ROI. Functional connectivity was calculated as the correlation between the time series of every pair of ROIs, resulting in a functional connectivity matrix consisting of 12,720 elements ((160 x 160)-160/2) for each of the 44 subjects. These connectivity values were then correlated with the mean head motion obtained from each subject, i.e. mean (FD_{FSL}). These correlations were plotted on the y-axis of a scatter plot with the Euclidean distance between ROIs on the x-axis. The estimated correlation between FD and RSFC was then used to compare the success of PACE with a combination of retrospective motion correction methods to correct for spurious changes caused in functional connectivity due to head motion. This procedure was repeated for all the combinations of nuisance, motion, and spike repressors discussed previously, and the results compared for the high-motion and the low-motion subgroups.

2.2.7 Impact of head motion censoring threshold on the removal of motion induced artifacts

Motion censoring is a trade-off between the quality and quantity of data. If PACE does correct for the lingering effects of spin history in the BOLD time series after the motion has ended, then a modest threshold with a small censoring window around the motion corrupted time-points would provide us with a good compromise. Therefore, to understand the impact of motion censoring on the reduction of motion artifacts, we considered four cases of motion censoring using a milder threshold of 0.5 mm and a stricter threshold of 0.2 mm.: (i) Censoring of volumes whose FD_{power}>0.5 mm and one volume after the motion corrupted volume (denoted as FD>0.5mm, 0B+1F, i.e. zero backward and one forward volumes are removed), (ii) Censoring

of volumes whose FD_{power}>0.5 mm as well as one volume before and two volumes after the motion corrupted volume (FD>0.5mm, 1B+2F), (iii) Censoring of volumes whose FD_{power}>0.2mm as well as one volume after the motion corrupted volume (FD>0.2mm, 0B+1F), (iv) Censoring of volumes whose FD_{power}>0.2 mm as well as one volume before and two volumes after the motion corrupted volume (FD>0.5mm, 1B+2F). We examined both the motion-BOLD relationships and the distance dependent connectivity artifact for each of the four cases to make a sound judgment on the appropriate motion threshold that prevents excessive loss of data.

2.2.8 Calculation of RS-fMRI based connectivity metrics

We calculated two RS-fMRI based metrics: 1) Network Degree Centrality (DC) and 2) PCC Seed Based Functional Connectivity (PCC-FC). These were chosen in order to (i) Evaluate the spatial relationship between functional connectivity metrics and motion and (ii) Compare the effectiveness of the motion correction strategies in High-motion and Low-motion Subgroups. Network Degree Centrality (DC) was calculated as the weighted sum of significant positive connections for every voxel in the brain [25, 26, 2]. A connection was deemed significant if the correlation coefficient exceeded a threshold of 0.25 ($p< 0.0001$). A subject level z-score was calculated by subtracting the mean for all voxels and dividing by the standard deviation. These subject level z-maps were registered to the MNI template and smoothed with a 4.5 mm^3 Gaussian kernel. PCC seed based functional connectivity (PCC-FC) was estimated by extracting the mean time series from the posterior cingulate cortex (PCC: 0, -53, 26; diameter=10 mm) and then calculating the Pearson's correlation coefficient with other voxels in the brain as was done in other studies [3, 23, 2]. This was done in the standardized space after preprocessing, filtering, normalization and smoothing. These correlation values were transformed to z values using

Fischer's r to z transformation. We calculated the correlation between the DC/PCC-FC maps and FD_{vox} for each subject across every voxel to obtain maps indicating the relationships between motion and FC metrics. To compare the effectiveness of the motion correction strategy across preprocessing pipelines with various nuisance regressors, we performed a t-test for high-motion and low-motion subgroups separately.

2.2.9 Deconvolution of BOLD data to examine the neural correlates of head motion

Deconvolution has been previously employed in inferring the underlying latent (unmeasured) neuronal activity in the resting state BOLD signal [27, 28], especially given that the HRF has different properties (such as time-to-peak, FWHM and response height) across different brain regions as well as between different subjects [29, 30, 31, 32]. HRF variability can have an influence on the values of functional connectivity by artificially elevating the correlations of the underlying neural activity or suppressing them. This point is illustrated in Figure 2.1. Since some of the low-frequency fluctuations in the BOLD signal could potentially be associated with the neural components of head-motion, we investigated if these regions could be identified in PACE data. A few studies have previously reported that the positive FD -BOLD relationships or FD-fcMRI relationships in the motor cortex might have a neurological origin [2, 33]. It is apparent that any neural activity reflects in the BOLD signal after a delay due to the hemodynamic response function. Therefore, in order to investigate the correspondence between neural activity and head motion, we need to take this delay into consideration. This can be achieved by deconvolving the HRF from BOLD data to recover the underlying latent neural signals and then finding the correlation between head motion and the estimated latent neural signals. HRF deconvolution removes the delay between neural activity and head motion and makes temporal correlation between them meaningful. According to our knowledge, no study has specifically

used the deconvolution procedure to investigate the presence of neural components of head motion in motion–BOLD relationships. So we performed a voxel-wise resting-state hemodynamic deconvolution as described in Wu et al. [28] using PACE data processed with the following retrospective correction strategies: CSF + WM + Friston-24 motion regression and CSF + WM + GS + Friston-24 motion regression. Subsequently, the de-convolved time-series, i.e. latent neural signals were correlated with voxel-specific framewise displacement (FD_{vox}). The BOLD signals were not band pass filtered before deconvolution; otherwise, a similar preprocessing strategy was used.

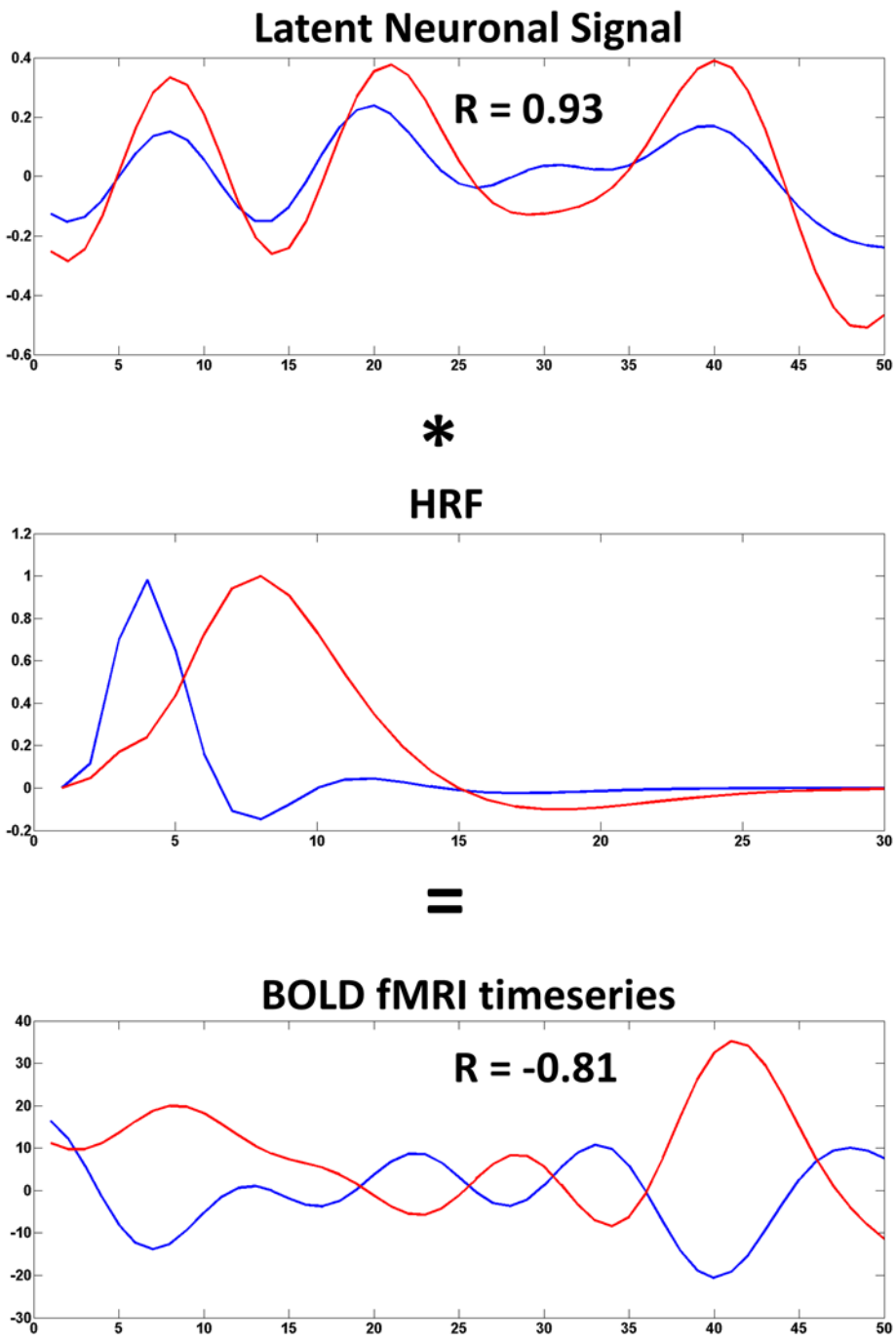


Figure 2.1. This figure illustrates how deconvolution of the BOLD timeseries (only 50 time points are shown though there were 250 time points in total) could change the functional connectivity between two fMRI timeseries. Note that the correlation coefficient for the BOLD fMRI timeseries for two voxels is -0.81 whereas for the correlation coefficient between latent neuronal signals in those same two voxels is 0.93.

2.3 Results

2.3.1 Examination of BOLD time-series data

Figure 2.2 shows PACE corrected resting state BOLD time-series extracted from the PCC in a representative high-motion subject as well as a low-motion subject and provides an insight into the effect of head motion on prospectively motion-corrected BOLD signal. The effect of each of the preprocessing steps on the time courses of the BOLD signal (PCC), DVARS (PCC) and DVARS (Whole Brain), each of which are obtained from PACE-corrected data, can be discerned. Figure 2.2 also displays motion metrics such as mean [TD_{vox} (PCC)] (total displacement of PCC), $mean_{sp}TD_{vox}$ (total displacement of the brain) , mean [FD_{vox} (PCC)], $mean_{sp}FD_{vox}$, FD_{FSL}, FD_{Power}, and the 6 realignment parameters. There is a linear relationship between residual motion metrics, and FD_{Power} seems to have the highest of all the FD_{vol} measures and that FD_{FSL} and $mean_{sp}FD_{vox}$, closely align with each other. This fact is confirmed when we plot subject wise summary motion metrics as shown in Figure 2.3. FD_{FSL} and mean [$mean_{sp}$ FD_{vox}] were highly correlated at 0.998 and so was FD_{Power} with $mean_{sp}$ FD_{vox} at 0.988. This is in tune with the previous result published by Yan, et al. [2], comparing motion metrics in non-PACE BOLD data using larger number of subjects.

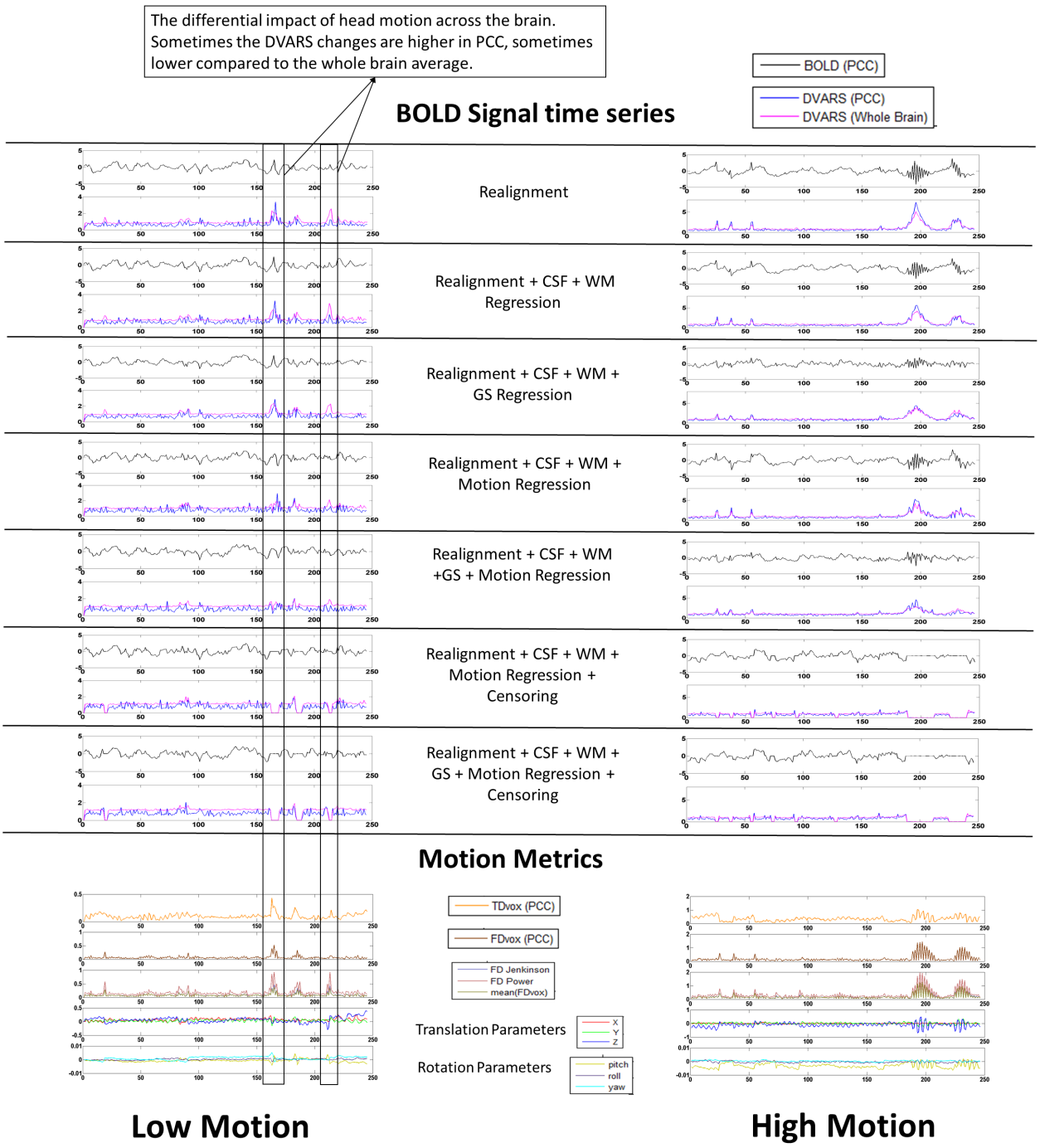


Figure 2.2. The PACE-corrected time-series extracted from the posterior cingulate cortex (PCC: 0,-53,26; 10 mm diameter sphere) at every step in the preprocessing pipeline for a representative subject in the high-motion (right) and low-motion (left) subgroups. Please note that the range of the y-axis for both the groups are the same for BOLD time series and range from -5 to 5. However for the motion metrics plots the range on the y-axes are different in the left and right panels in order to better visualize the type of motion in low-motion subjects. Large changes in the head position are associated with large changes in head motion. Regression of nuisance

variables was not successful in eliminating large spikes in head motion in the high-motion subject, but they were relatively successful in the low motion subject. The differential impact of head motion on the signal changes across the brain is illustrated as well.

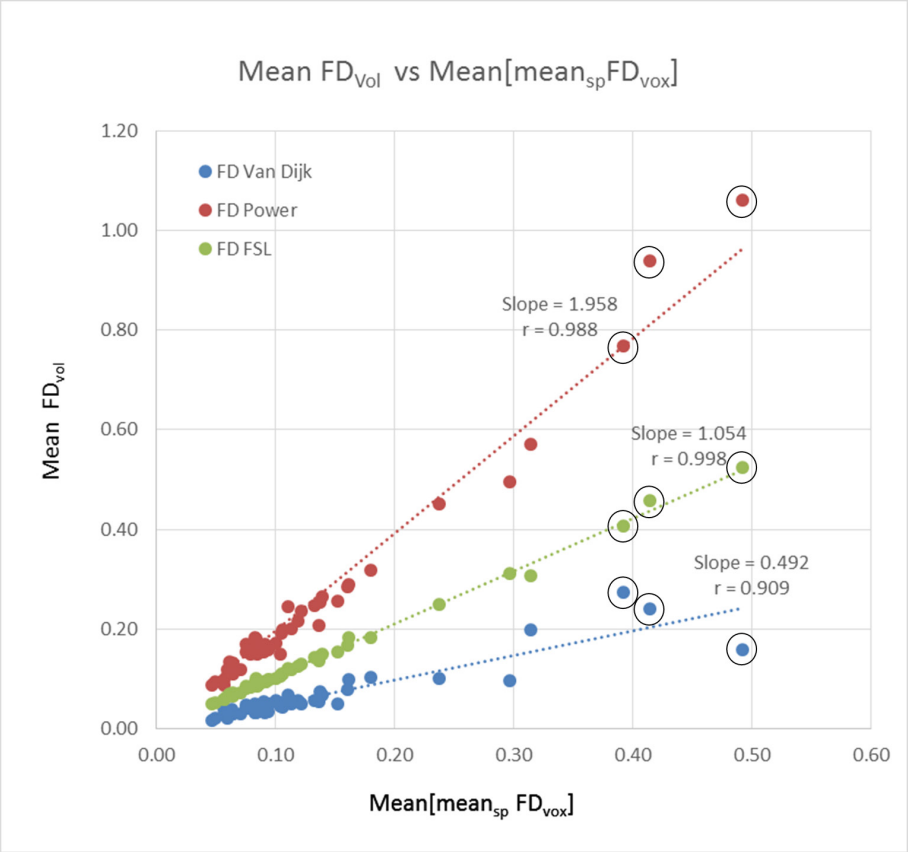


Figure 2.3. The correlation between the different volumetric Framewise Displacement (FD_{vol}) metrics with the voxel-wise derived Framewise Displacement (FD_{vox}). As reported in Yan et al, (2013), all three volumetric measures of FD_{vol} were highly correlated with mean[mean$_{sp}$(FD_{vox})] and the slope of FD_{FSL} was twice that of $FD_{Van\ Dijk}$ but half of FD_{Power}. The circled subjects in the figure were removed from the study due to excessive head motion and the lack of sufficient time-points required for stable estimation of RS-fMRI metrics after motion censoring.

Looking at the impact of the motion on the PACE-corrected BOLD time series in Figure 2.2, we observe that large changes in the head position roughly correspond to the significant changes in the PACE-corrected BOLD time series. While this has been shown to be true for non-PACE data

[4, 2, 7, 3], we show here that the same is true for PACE-corrected data as well although the magnitude of such changes may be different in PACE data. DVARS, which measures the change in the BOLD signal, approximately follows the sharp rise and fall in head motion (as captured in the framewise displacement). It is also worth noting that some head movements were associated with larger signal changes in the PCC compared to the the whole brain signal and in some cases, it was other way around. This points to the differential impact of head motion on the PACE-corrected BOLD signal in different regions of the brain. WM and CSF regression did not seem to alleviate the motion artifact, however, GS regression and to some extent motion regression appears to have reduced the artifact. In the high motion subject who was relatively still except for two large head movements, a ringing effect (rapid changes) in the PACE-corrected BOLD signal and associated FD metrics can be observed. One possible reason for this could be that, due to the prospective correction by PACE, the scanner seems to be adjusting to the motion, thus causing a distinct effect on the BOLD time courses. Of course, these patterns appear to have reduced after WM, CSF, and GSR, and motion parameter regression but is still distinctly present after nuisance covariate regression, giving merit to the argument that censoring the motion corrupted volumes is the best way to eliminate the artifactual effects of residual head motion on the PACE-corrected BOLD Signal. Although the data is shown for just two subjects, it effectively demonstrates the limitations of nuisance signal regression in preprocessing Rs-fMRI data obtained with PACE. To understand how much variance is explained by the Friston-24 motion parameters and the six realignment parameters, we estimated the average PACE-corrected BOLD signal variance explained by the 24 motion parameters and six realignment parameters for each subject and averaged the results over all subjects as shown in Figure 2.4A and Figure 2.4B, respectively. This result is pretty similar to the one reported by Satterthwaite, et

al. [3] (Figure 2.4C), who used adjusted R^2 maps to illustrate the signal varaince explained by using six realignment parameters. The BOLD signal from regions which are farthest from the centre of the brain are affected the most by head motion and consequntly more variance in the PACE-corrected BOLD signal is explained by the 24 motion parameters. Also, compared to six realignment parameters (Figure 2.4B, Figure 2.4C) use of 24 motion parameters (Figure 2.4A) explained a lot more variance (as observed by their adjusted R^2 values) across the brain.

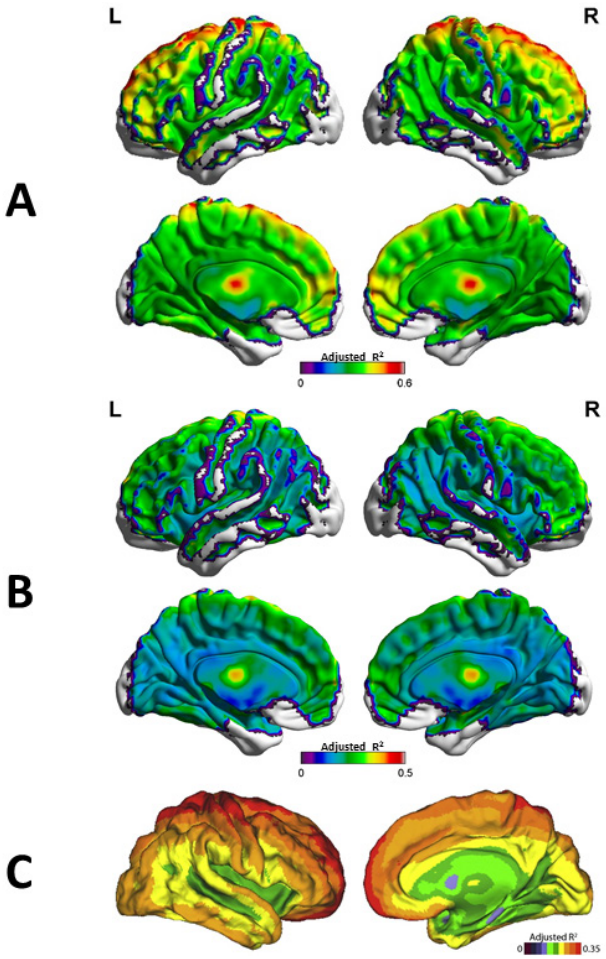


Figure 2.4. The average BOLD signal variance (adjusted R^2) explained by the 24 regressors used in the Friston-24 motion regression model (A) and the six realignment parameters (B). (C) Adjusted R^2 maps showing amount of signal variance explained by 6 standard from Satterthwaite et al., 2013. Figures A and B are similar, except for that fact that 24 motion regressors (A) explain far more variance across the brain compared to using just 6 motion parameters (B). These motion regressors explain a modest amount of variance in the brain, with more variance

explained in the frontal regions and less variance explained in other (especially posterior) regions. This is to be expected given that frontal regions experience more displacement than other regions of the brain (Yan et al., 2013).

2.3.2 Voxel-wise relationships between framewise displacement and PACE-corrected BOLD signal

The sensitivity of the BOLD time series to head motion artifacts is spatially varying and this can be characterized by the FD-BOLD relationship. Consistent linear relationships (or correlations) between FD_{vox} and the BOLD signal across the brain is indicative of the motion artifact and can affect the estimation of functional connectivity between brain regions. Figure 2.5 shows the raw FD_{vox}-BOLD correlation maps as well as maps thresholded at ($T > 4.95$, $p<0.05$, FDR corrected). As was observed with the PACE-corrected time-series data, WM, CSF and motion regression did not remove significant motion-BOLD relationships. An interesting observation is that negative motion-BOLD relationships which are associated with large head movements [2] were absent in the thresholded maps obtained from PACE-corrected data. It can be seen from Figure 2.5 that none of the voxels exceed the negative threshold, implying no significant negative motion-BOLD relationships were present. It is noteworthy that results from non PACE-corrected data reported before show significant negative motion-BOLD relationships [2]. With the addition of GSR, the large positive relationships were reduced across the brain, but negative correlations were introduced. However, it should be noted that none of the negative motion-BOLD correlations were significant. With relatively modest motion censoring (FD>0.5mm, 1B+2F) and without GSR, motion BOLD relationships were not significant ($p>0.05$), and very few positive relationships survived the thresholds. In order to obtain an equivalent result with non-PACE data, Yan et al., had to use a much stricter censoring threshold of $FD_{Power}>0.2$mm coupled with GSR [2]. If we used GSR or increased the censoring threshold to those used by Yan et al., all

motion-BOLD relationships were eliminated. This indicates that one could use liberal censoring (thereby retaining more data) and avoid the confounding effects of GSR and yet eliminate all negative, and most positive motion-BOLD relationships using PACE data.

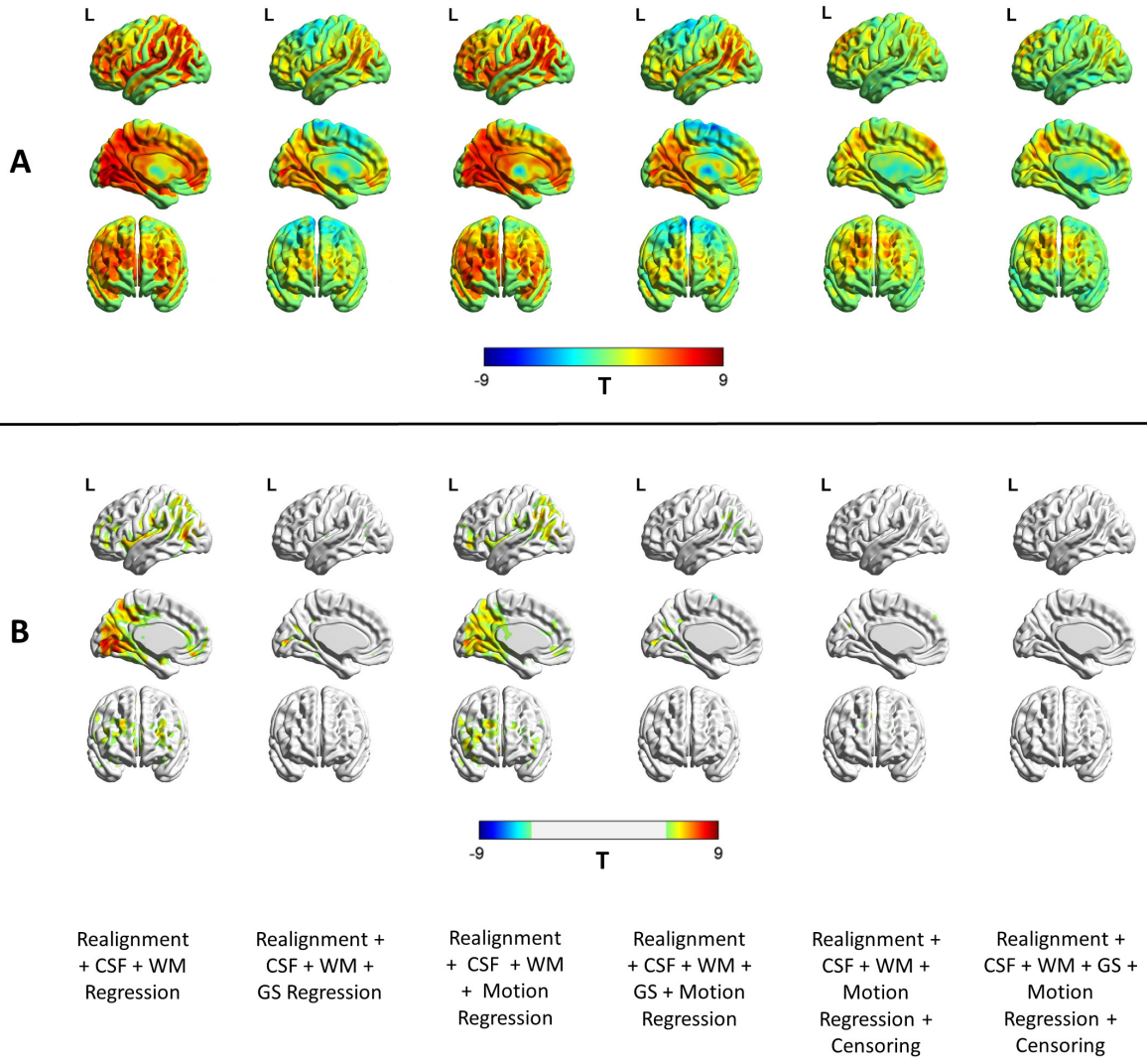


Figure 2.5. Illustration of the reduction in the relationship between motion and PACE-corrected BOLD data for different nuisance variable regressors. The un-thresholded T maps are shown in (A) and the thresholded ($p<0.05$, FDR corrected) maps are shown in (B). The results indicate that motion regression did not remove motion-BOLD relationships visibly. However, GS regression did seem to reduce these relationships, with some regions now showing a negative correlation. (B) After the nuisance variance regressions, some regions did exhibit significant positive relationships with the BOLD signal, though no negative relationships remained. With censoring, both positive and negative relationships are almost absent.

To better understand the patterns of these relationships between high-motion and low-motion subjects, we repeated the analysis separately for high-motion and low-motion subgroups. The result is shown in Figure 2.6 and the corresponding thresholded T-maps for high and low motion subgroups shown in Figure 2.7. The results show small motion-BOLD correlations for low-motion subgroup as expected, with significant correlations ($p<0.05$, FDR corrected) only restricted to the visual areas after WM and CSF regression. Further steps of preprocessing eliminated even those correlations to below significance. However, the relationships for high-motion subgroup relationships reduced to below chance levels in most areas only after motion censoring. Qualitatively, motion-BOLD correlations obtained from PACE data appear to be smaller in magnitude and spatial extent when compared to those obtained from non-PACE data (in both low and high motion subjects) reported previously [2].

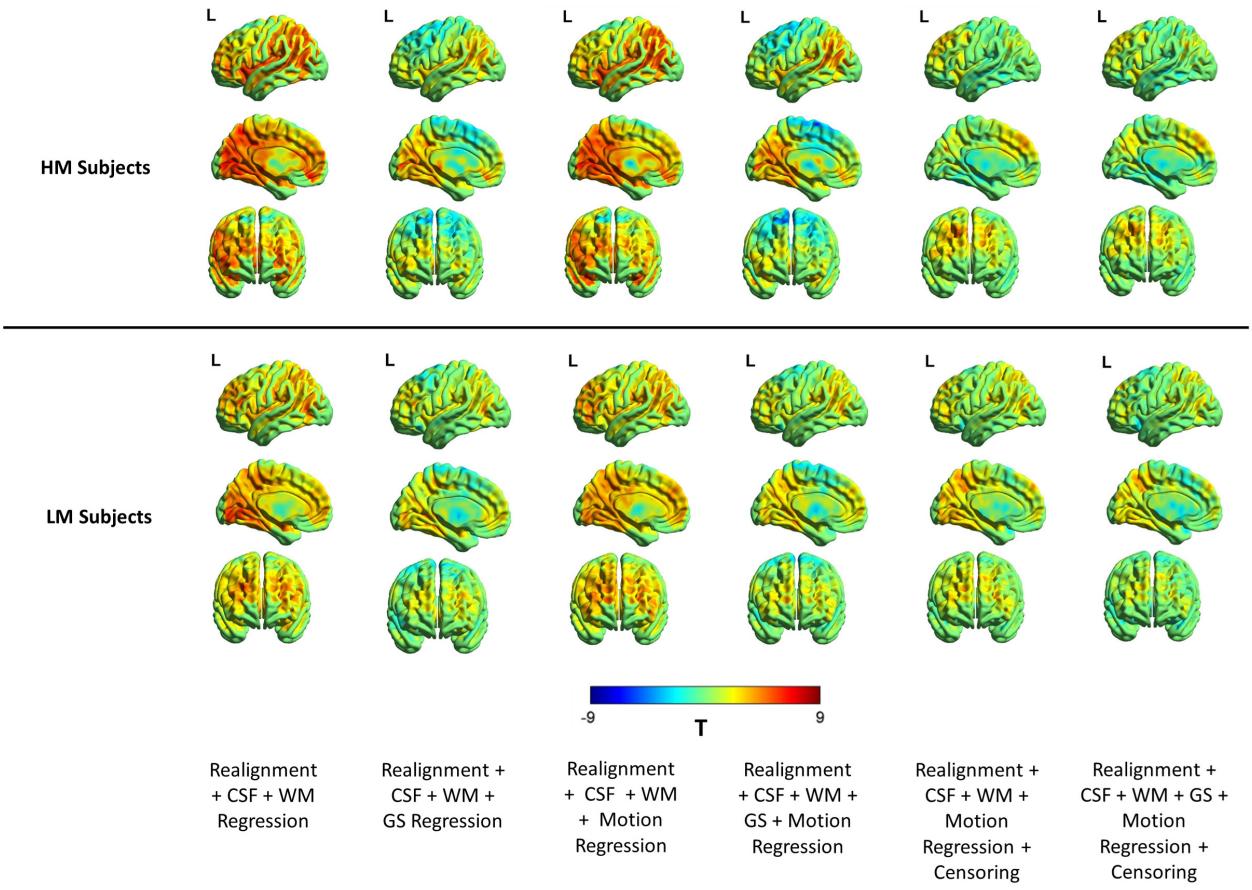


Figure 2.6**.** The un-thresholded T maps illustrating the relationship between the PACE-corrected BOLD signal and voxel-specific framewise displacement for the high motion and the low motion Subgroups. CSF, WM, and motion regression are relatively ineffective in reducing the motion-BOLD relationships both in high motion and low motion Subjects. Large motion-BOLD relationships are comparatively fewer in lowmotion subjects, as expected. GSR significantly increased negative motion–BOLD relationships in high motion subgroup, but not by much in the low motion subgroup. With motion censoring, GSR has a relatively negligible effect on the motion-BOLD relationships in both the subgroups.

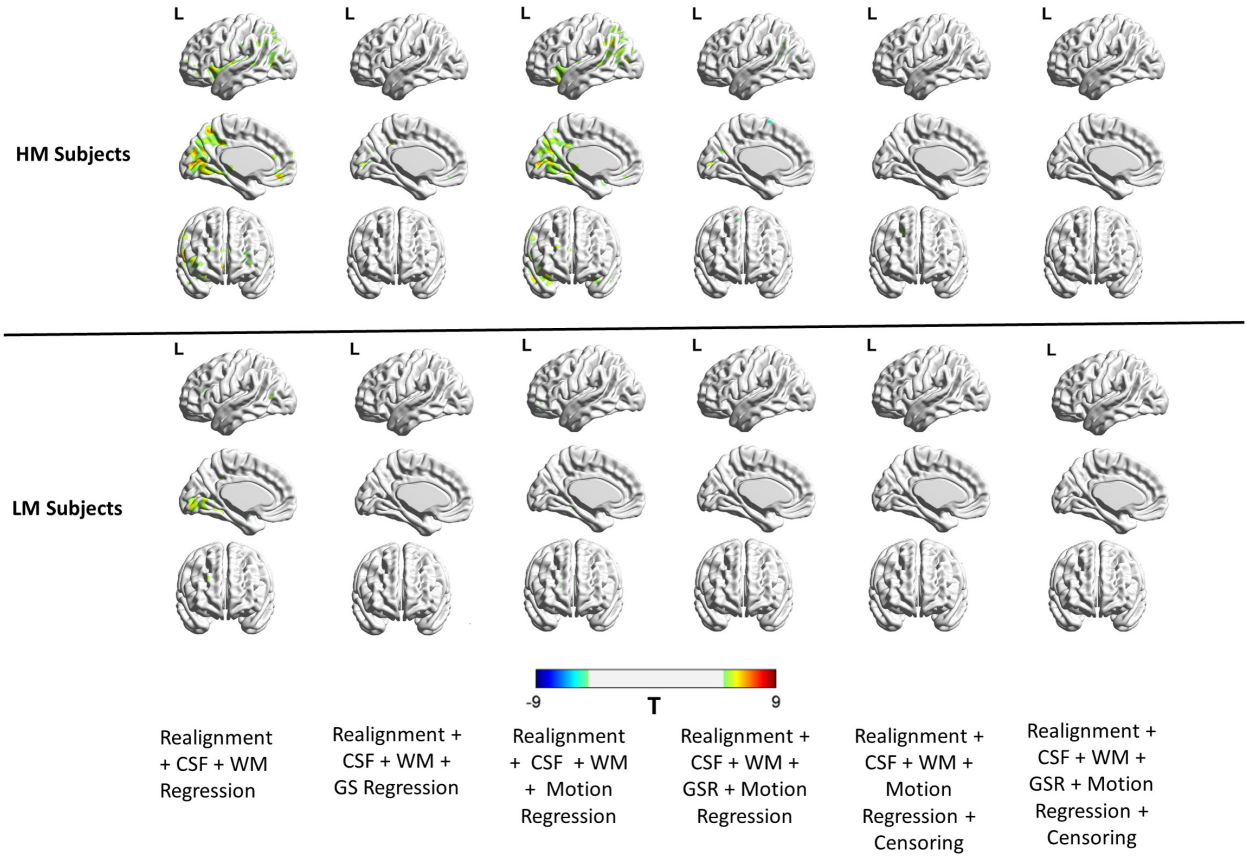


Figure 2.7. Thresholded correlation maps between the PACE-corrected BOLD signal and head motion (FD_{vox}) across the brain. The figure shows the relative absence of significant ($p<0.05$, FDR corrected) motion-BOLD relationships in low motion subjects compared to the high motion subjects. The reduction in motion-BOLD relationships after GS and motion regression in high motion subjects is stark, although residual correlations in the visual cortex are only eliminated after motion censoring.

Given the reported usefulness of including summary motion statistics of the subjects in group-level analyses to eliminate the group differences in the motion artifact [2, 3, 4], we show the variance (adjusted R^2) in the residual motion-BOLD relationships as explained by the subject summary motion statistic in Figure 2.8A. The correlation between summary motion statistic and the motion-BOLD maps is also shown in Figure 2.8B. In Figure 2.8A, the summary statistic explained a lot more variance with significant positive and negative correlations across the brain regions with only CSF, WM, GS and motion regression. With motion censoring, a lot less

variance was explained by the summary motion parameters as expected and a lot less correlation with the summary static was observed. This result explains why group level correction in unnecessary when motion censoring is performed at the subject level, thus confirming previous results that the benefits of group level correction are negated if motion censoring is included during preprocessing [2].

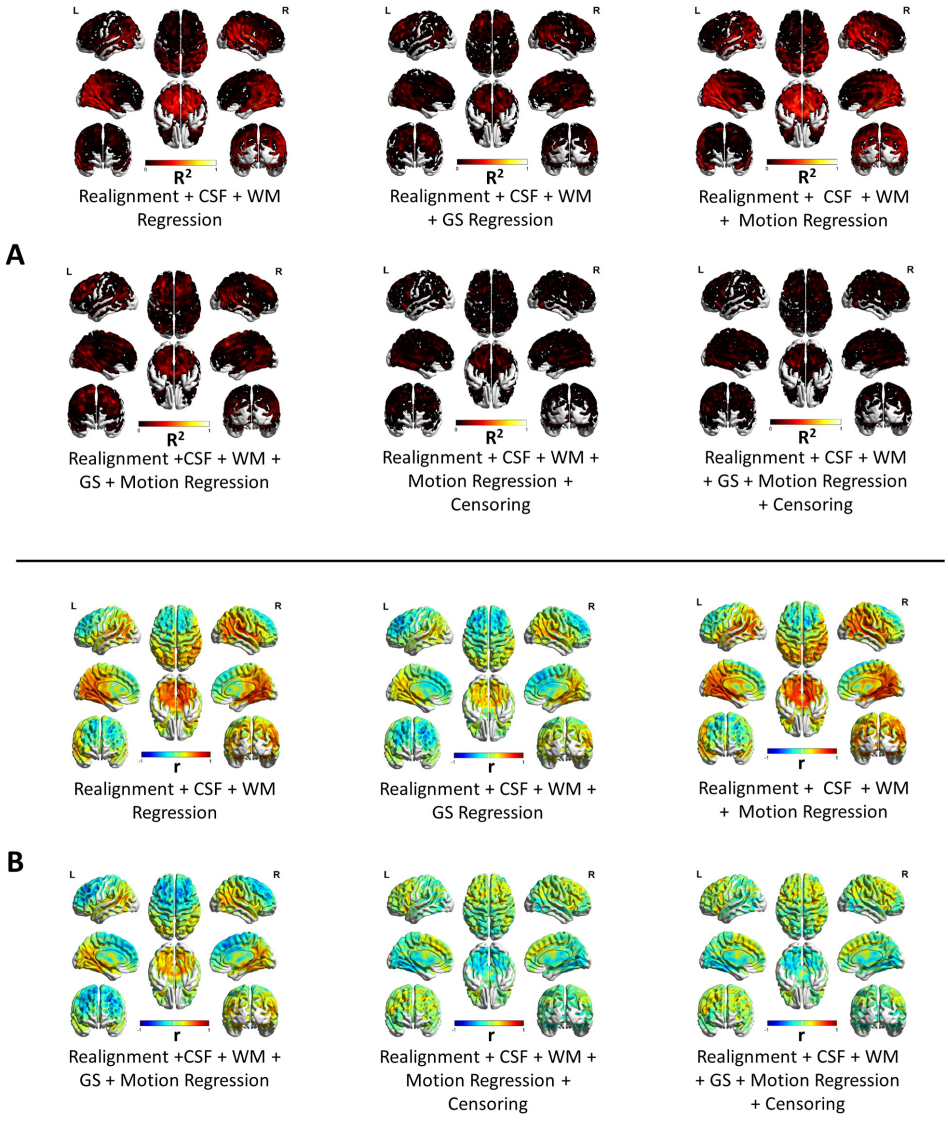


Figure 2.8. (A) The variance (R^2) in the residual motion-BOLD relationships as explained by the subject summary motion statistic after nuisance variable regression. (B) The correlation between summary motion statistic and the motion-BOLD relationships. The summary statistic explained a

lot more variance with significant positive and negative correlations across the brain regions with only CSF, WM, GS and motion regression. With censoring, a lot less variance is explained by the summary motion parameters as expected, confirming that group level correction is not necessary if motion censoring is included in the preprocessing.

2.3.3 Motion induced distance dependent artifact in resting-state functional connectivity

When functional connectivity is estimated with motion-corrupted data, connectivity strengths between two brain regions can be dependent on the relative location of the regions and the similarity in magnitude and the direction of the displacement experienced by head motion. This artifact helps us in evaluating the success of a motion correction strategy and the absence of the motion artifact in the data. We plot the 12720 connectivity values (obtained from PACE-corrected BOLD time series) which were correlated with each subject's summary head motion (mean [FD_{FSL}]) as a function of distance. This was done for all combination of nuisance regressors and motion censoring, and we show the results for all the subjects as well as for the high-motion and the low-motion subgroups separately in Figure 2.9. Ideally, if head motion was not artifactually modulating the connectivity values, we expect the plot to be a flat (zero slope) line. But as Figure 2.9 illustrates, the distance dependent artifact was present for all combinations of nuisance variable regression including WM, CSF, GS and Friston-24 motion regression. The correlation of head movement with the connectivity metrics exhibited positive values for all distances and only with the introduction of GSR, were the correlation with motion became negative for functional connectivity between farther regions. With motion censoring, this artifact did not seem to have been completely eliminated, especially in high-motion subjects, with a small slope and a positive intercept when fitted by a linear trend line. There was a positive correlation between FC and head motion at all distances in high-motion subjects. The artifact almost seems absent in low-motion subjects for all combinations of nuisance variable regression

and censoring as the slope is small and the line is relatively flat. In contrast, previous reports with non-PACE data indicate that the distance dependent artifact could not be eliminated (unless censoring thresholds were more severe than what we have used) even in low motion subjects [20]. When all subjects were used, a combined effect was noticed. A few more observations include that GSR appears to distort the distance dependent artifact and makes the artifact worse by increasing the slope in high-motion subjects and the variance in low-motion subjects. This result is in agreement with the observations made by Jo et al., that GSR distorts functional connectivity values [34]. However, when GSR was combined with censoring, it did seem to eliminate the distance dependent artifact even in subjects with high motion. Since we used a relatively modest threshold of 0.5 mm with a censoring window of one previous volume and two forward volumes, we wanted to see if a more severe threshold of FD_{Power}>0.2 mm, would have any additional benefits at the cost of substantial loss of data.

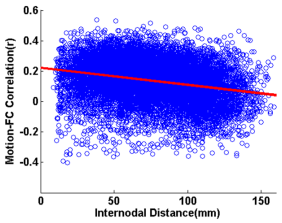
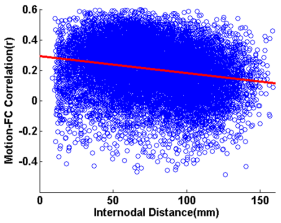
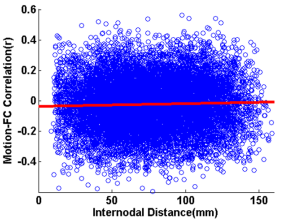
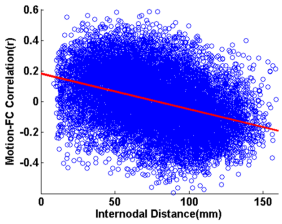
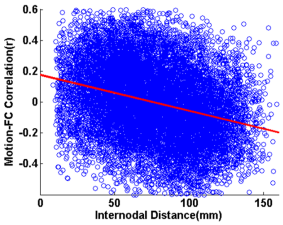
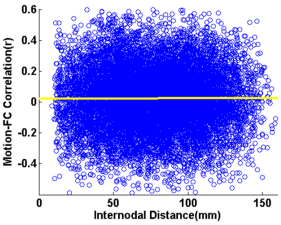
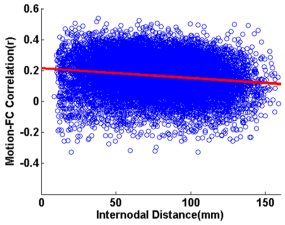
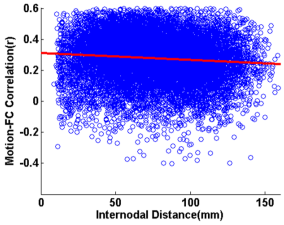
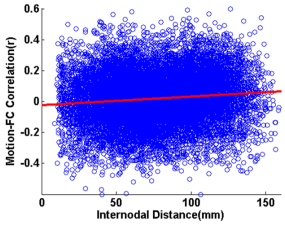
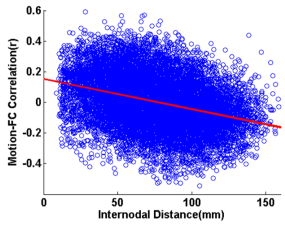
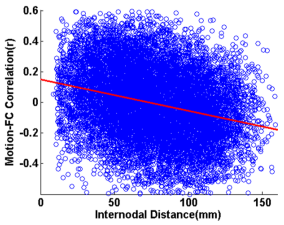
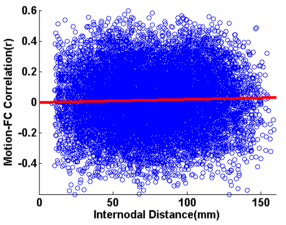
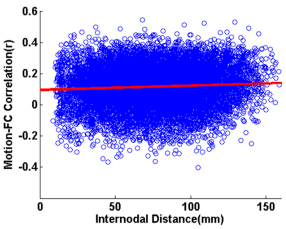
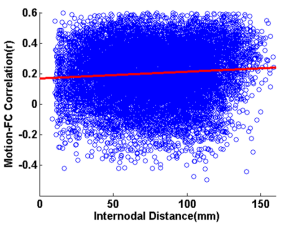
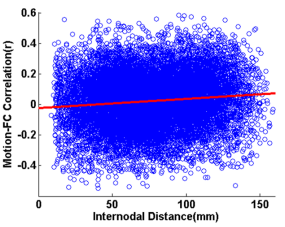
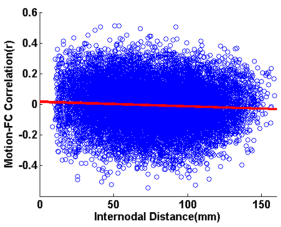
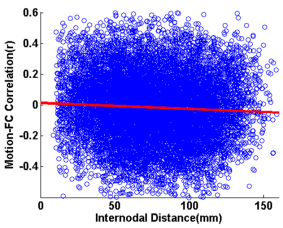
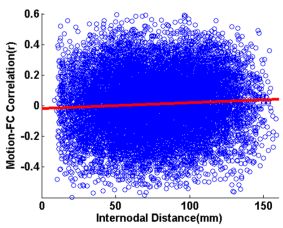
	All Subjects	High Motion Subjects	Low Motion Subjects
Realignment + CSF + WM Regression			
	p = 0 Slope = -0.001	p = 0 Slope = -0.001	p = 0 Slope = 2×10^{-4}
Realignment + CSF + WM + GS Regression			
	p = 0 Slope = -0.002	p = 0 Slope = -0.002	p = 0.82 Slope = 1.3×10^{-5}
Realignment + CSF + WM + Motion Regression			
	p = 3.6×10^{-65} Slope = -6.3×10^{-4}	p = 6.6×10^{-22} Slope = -4.5×10^{-4}	p = 2.6×10^{-29} Slope = 5.5×10^{-4}
Realignment + CSF + WM + GS + Motion Regression			
	p = 0 Slope = -0.002	p = 0 Slope = -0.002	p = 5×10^{-4} Slope = 2×10^{-4}
Realignment + CSF + WM + Motion Regression + Censoring			
	p = 1.2×10^{-12} Slope = 3×10^{-4}	p = 2.1×10^{-15} Slope = 4×10^{-4}	p = 5×10^{-32} Slope = 6×10^{-4}
Realignment + CSF + WM + GS + Motion Regression + Censoring			
	p = 1.5×10^{-11} Slope = -3×10^{-4}	p = 3.1×10^{-9} Slope = -4×10^{-4}	p = 7.4×10^{-11} Slope = 4×10^{-4}

Figure 2.9. The figure shows the FD-RSFC correlations for all the subjects as well as for the high and low-motion subgroups. The motion induced distance dependent FC artifact is almost absent in the low-motion subgroup for all stages and all combinations of nuisance signal regression. The color of the linear fit line indicates significance of the fit with red indicating significance ($p<0.05$, FDR corrected) and yellow indicating nonsignificance ($p>0.05$). The high motion subgroup does show the artifact which is only reduced after motion censoring. GSR distorts the FD-RSFC relationships significantly, especially in the high motion subgroup, though after motion censoring, the data is relatively free from the artifact in both the subgroups, with and without GSR.

2.3.4 Impact of censoring threshold on the existence of motion artifacts

In order to better understanding the dynamics between the removal of motion artifacts and preservation of non-corrupted PACE-corrected BOLD data, we experimented by using two different censoring thresholds (0.5 mm and 0.2 mm) and two censoring windows around the motion corrupted volumes (0B+1F and 1B+2F). This gave rise to four scenarios of motion censoring (i) $FD_{Power}> 0.5$, 0B+1F, (ii) $FD_{Power}> 0.5$, 1B+2F, (iii) $FD_{Power}> 0.2$, 0B+1F, (iv) $FD_{Power}> 0.2$, 1B+2F. Table 2.1 shows the total number of time-points, and the fraction of time-points censored for each of the above censoring scenarios. As expected there was huge loss in data when we used censoring at 0.2 mm compared to 0.5 mm. In fact, the number of subjects, who had at least 3 minutes of good data or 180 time points in our case was reduced from 47 to 24, when the threshold was greater than 0.2 mm and 1 volume before and two volumes after the motion were removed. We used the presence of the significant ($p<0.05$, FDR corrected) motion-BOLD relationships and the existence of the motion induced distance dependent functional connectivity artifact (FD-RSFC correlations) to assess the quality of non-motion corrupted data. To be fair in the comparison, we used the same 24 remaining subjects for all the four censoring cases. Since most of the subjects left had pretty low-motion, we expected results similar to those obtained by the low-motion dataset. It is noteworthy that several subjects were common to both

the subsets of data and there was relative absence of motion artifacts in the low-motion subgroup even with relatively less preprocessing. The unthresholded T-maps are shown in Figure 2.10A and the thresholded T-maps ($p<0.05$, FDR corrected) are shown in Figure 2.10B. As Figure 2.10B shows, motion-BOLD relationships were below significance for all voxels in all the four scenarios of motion censoring, though the values of few motion-BOLD relationships in the visual areas are removed with the more stringent threshold. The motion induced distance dependent connectivity artifact appeared to be appeared to be considerably reduced (Figure 2.10C) in the four cases as the slope was very small. But the slope was not significant ($p>0.05$) for the case with censoring $FD_{Power}> 0.5$, 0B+1F. The slope was small as well as significant ($p<0.05$) for other censoring scenarios, indicating that motion induced distance dependent connectivity artifact is eliminated after censoring the data at $FD_{Power}> 0.5$, 1B+2F. Increasing the censoring window size beyond the motion corrupted volume and a single volume after the corrupted volume, did not seem to have any effect on the data even after filtering the time series. Our results indicate that censoring volumes at a more stringent threshold of 0.2 mm or increasing the censoring window size to include more volumes did not have a detectable improvement in the data quality as the artifacts were almost eliminated at 0.5 mm, but it came at the cost of substantial loss of data.

			FD > 0.5 mm, 0B+1F			FD > 0.5 mm, 1B+2F			FD > 0.2 mm, 0B+1F			FD > 0.2 mm, 1B+2F		
Subject No.	mean FD_{PSL}(mm)	Total Timepoints	Timepoints Censored	Remaining Timepoints	% Timepoints Censored	Timepoints Censored	Remaining Timepoints	% Time points Censored	Timepoints Censored	Remaining Timepoints	% Time points Censored	Timepoints Censored	Remaining Timepoints	% Time points Censored
1	0.13	995	102	893	10.3	158	837	15.9	705	290	70.9	842	153	84.6
2	0.06	995	18	977	1.8	36	959	3.6	81	914	8.1	142	853	14.3
3	0.11	995	48	947	4.8	68	927	6.8	561	434	56.4	752	243	75.6
4	0.06	995	2	993	0.2	4	991	0.4	44	951	4.4	70	925	7.0
5	0.05	995	44	951	4.4	54	941	5.4	69	926	6.9	93	902	9.3
6	0.06	995	4	991	0.4	8	987	0.8	42	953	4.2	66	929	6.6
7	0.08	995	6	989	0.6	12	983	1.2	399	596	40.1	562	433	56.5
8	0.05	995	12	983	1.2	20	975	2.0	45	950	4.5	79	916	7.9
9	0.18	995	158	837	15.9	204	791	20.5	925	70	93.0	975	20	98.0
10	0.14	995	18	977	1.8	30	965	3.0	701	294	70.5	847	148	85.1
11	0.41	245	206	39	84.1	219	26	89.4	237	8	96.7	243	2	99.2
12	0.13	245	26	219	10.6	32	213	13.1	130	115	53.1	170	75	69.4
13	0.09	245	0	245	0.0	0	245	0.0	92	153	37.6	132	113	53.9
14	0.10	245	6	239	2.4	10	235	4.1	148	97	60.4	185	60	75.5
15	0.08	245	6	239	2.4	12	233	4.9	81	164	33.1	116	129	47.3
16	0.09	245	13	232	5.3	24	221	9.8	89	156	36.3	127	118	51.8
17	0.07	245	4	241	1.6	8	237	3.3	39	206	15.9	57	188	23.3
18	0.08	245	14	231	5.7	22	223	9.0	90	155	36.7	126	119	51.4
19	0.07	245	8	237	3.3	14	231	5.7	68	177	27.8	104	141	42.4
20	0.09	245	14	231	5.7	24	221	9.8	108	137	44.1	141	104	57.6
21	0.46	245	115	130	46.9	138	107	56.3	206	39	84.1	230	15	93.9
22	0.09	245	6	239	2.4	10	235	4.1	118	127	48.2	156	89	63.7
23	0.14	995	83	912	8.3	125	870	12.6	856	139	86.0	931	64	93.6
24	0.31	995	471	524	47.3	486	509	48.8	644	351	64.7	733	262	73.7
25	0.25	995	333	662	33.5	407	588	40.9	910	85	91.5	973	22	97.8
26	0.52	495	334	161	67.5	355	140	71.7	438	57	88.5	473	22	95.6
27	0.10	495	18	477	3.6	28	467	5.7	227	268	45.9	297	198	60.0
28	0.09	495	36	459	7.3	46	449	9.3	102	393	20.6	149	346	30.1
29	0.07	495	22	473	4.4	34	461	6.9	64	431	12.9	98	397	19.8
30	0.12	495	60	435	12.1	70	425	14.1	155	340	31.3	200	295	40.4
31	0.07	495	27	468	5.5	39	456	7.9	47	448	9.5	65	430	13.1
32	0.09	495	25	470	5.1	29	466	5.9	99	396	20.0	141	354	28.5
33	0.09	495	36	459	7.3	48	447	9.7	138	357	27.9	191	304	38.6
34	0.10	495	6	489	1.2	12	483	2.4	179	316	36.2	253	242	51.1
35	0.09	495	35	460	7.1	60	435	12.1	127	368	25.7	167	328	33.7
36	0.11	495	59	436	11.9	75	420	15.2	114	381	23.0	149	346	30.1
37	0.10	495	70	425	14.1	80	415	16.2	106	389	21.4	132	363	26.7
38	0.07	495	18	477	3.6	33	462	6.7	63	432	12.7	89	406	18.0
39	0.31	495	206	289	41.6	228	267	46.1	414	81	83.6	468	27	94.5
40	0.15	495	119	376	24.0	131	364	26.5	186	309	37.6	227	268	45.9
41	0.17	495	98	397	19.8	149	346	30.1	421	74	85.1	458	37	92.5
42	0.18	495	81	414	16.4	111	384	22.4	359	136	72.5	413	82	83.4
43	0.10	495	14	481	2.8	24	471	4.8	247	248	49.9	335	160	67.7
44	0.09	495	4	491	0.8	8	487	1.6	183	312	37.0	265	230	53.5
45	0.15	495	89	406	18.0	109	386	22.0	303	192	61.2	393	102	79.4
46	0.14	495	69	426	13.9	78	417	15.8	214	281	43.2	299	196	60.4
47	0.12	495	24	471	4.8	40	455	8.1	413	82	83.4	476	19	96.2

Table 2.1. Table showing motion statistics and the loss of data using two different censoring FD_{power} thresholds of 0.2 mm and 0.5 mm. Censoring at a higher threshold of FD_{power} >0.2 mm causes significant loss of data (subjects whose data is unusable due to less than 300 time points present per subject is highlighted in red) compared to a lower threshold of FD_{power} >0.5 mm. With a higher censoring threshold of 0.2 mm, almost half of the subjects would have to be eliminated from the study as they do not have the minimum required data of 3 minutes necessary for reliable estimation of RSFC metrics.

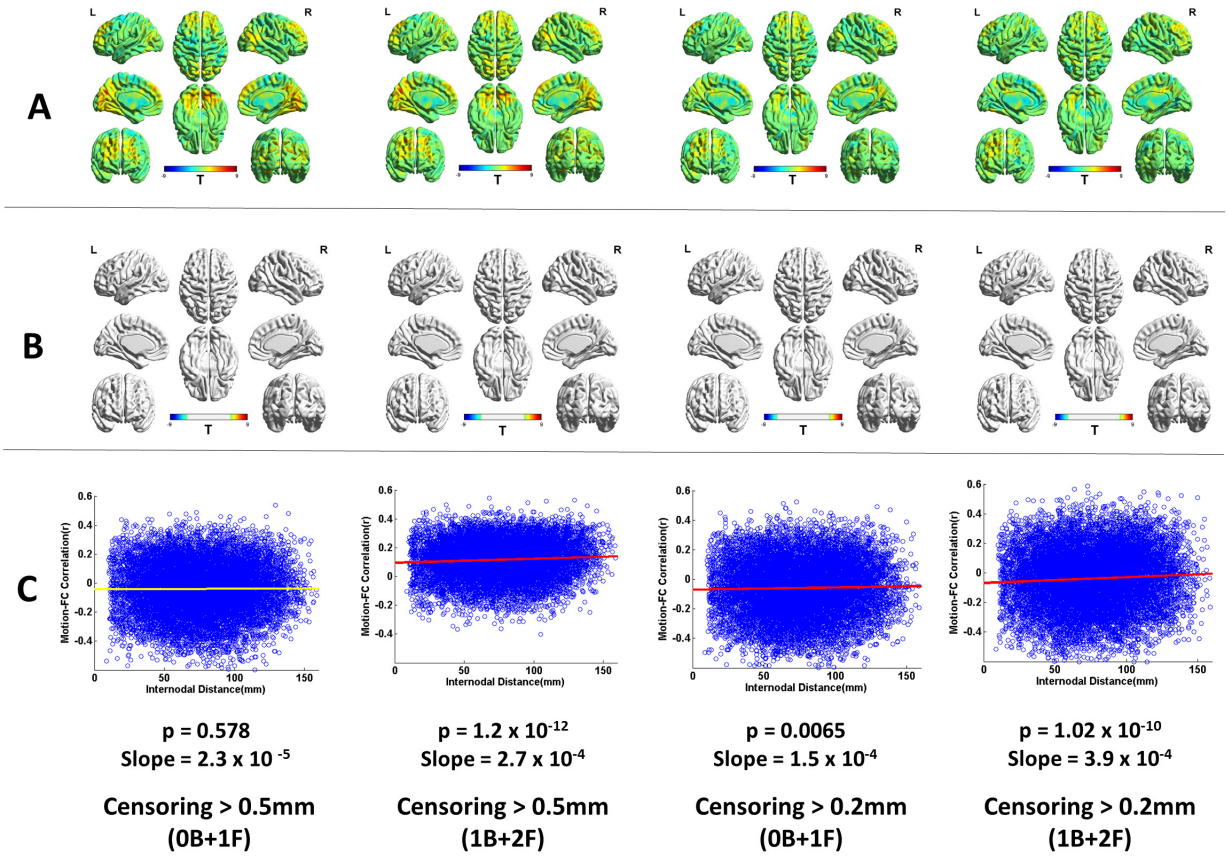


Figure 2.10. The absence of motion artifacts for the four cases of motion censoring. (A) The motion-BOLD relationships indicate very small positive motion–BOLD relationships in the visual cortex, which are removed by censoring the volumes at a lower (more stringent) threshold. (B) Thresholded motion-BOLD relationships for the figures shown in A. It must be noted that none of the volumes exhibited significant correlations for all the four scenarios of censoring. (C) The FD-RSFC correlations, which can be used to detect the presence of the motion induced distance dependent FC artifact, shows that for all the cases of censoring, the artifact was absent. The color of the linear fit line indicates significance of the fit with red indicating significance ($p<0.05$, FDR corrected) and yellow indicating nonsignificance ($p>0.05$). A stricter threshold for censoring or a larger censoring window does not seem to have a detectable improvement in data quality.

2.3.5 The impact of motion on functional connectivity estimates of degree centrality and PCC-FC

As seen earlier, motion does affect functional connectivity and other measures derived from it even with EPI-PACE acquisition. In order to understand the residual relationships between

functional connectivity metrics and motion, we calculated the Pearson's correlation coefficient between head motion, i.e. mean [FD_{vox}], and Degree Centrality (DC) (Figure 2.11) and between mean[FD_{vox}] and PCC seed based functional connectivity (Figure 2.12), for all the subjects as well as separately in the high-motion and the low-motion subgroups. As shown in Figure 2.11, degree centrality was relatively robust to the influence of motion artifact due to Z-standardization [2]. This implies that nuisance variable regression and censoring did not have much impact on the FD-DC correlations. However, we found large positive correlations in the sensorimotor cortex, and the correlations seemed to increase as motion artifacts were removed from the data via motion regression and censoring as shown in Figure 2.11. A more detailed image of motion-DC correlation in sensorimotor cortex is shown in Figure 2.13. This effect was observed both in the high-motion and the low-motion subgroups. A similar result was reported by Pujol et al., indicating that there is a component of motion related connectivity changes that may have a neural basis and may not be just a consequence of the motion artifact [33]. The FD-PCC functional connectivity correlation map (shown in Figure 2.12) identifies the regions whose correlation with PCC varies as a function of subject head motion. We observed significant ($p<0.05$, FDR corrected) negative correlations between residual motion and PCC-FC in the frontal regions in the low-motion subgroup and a significant reduction in the positive correlations especially in subjects with high motion as GS, motion regression, and censoring were performed. This highlights their relative effectiveness in reducing motion artifacts, particularly in subjects with high head motion.

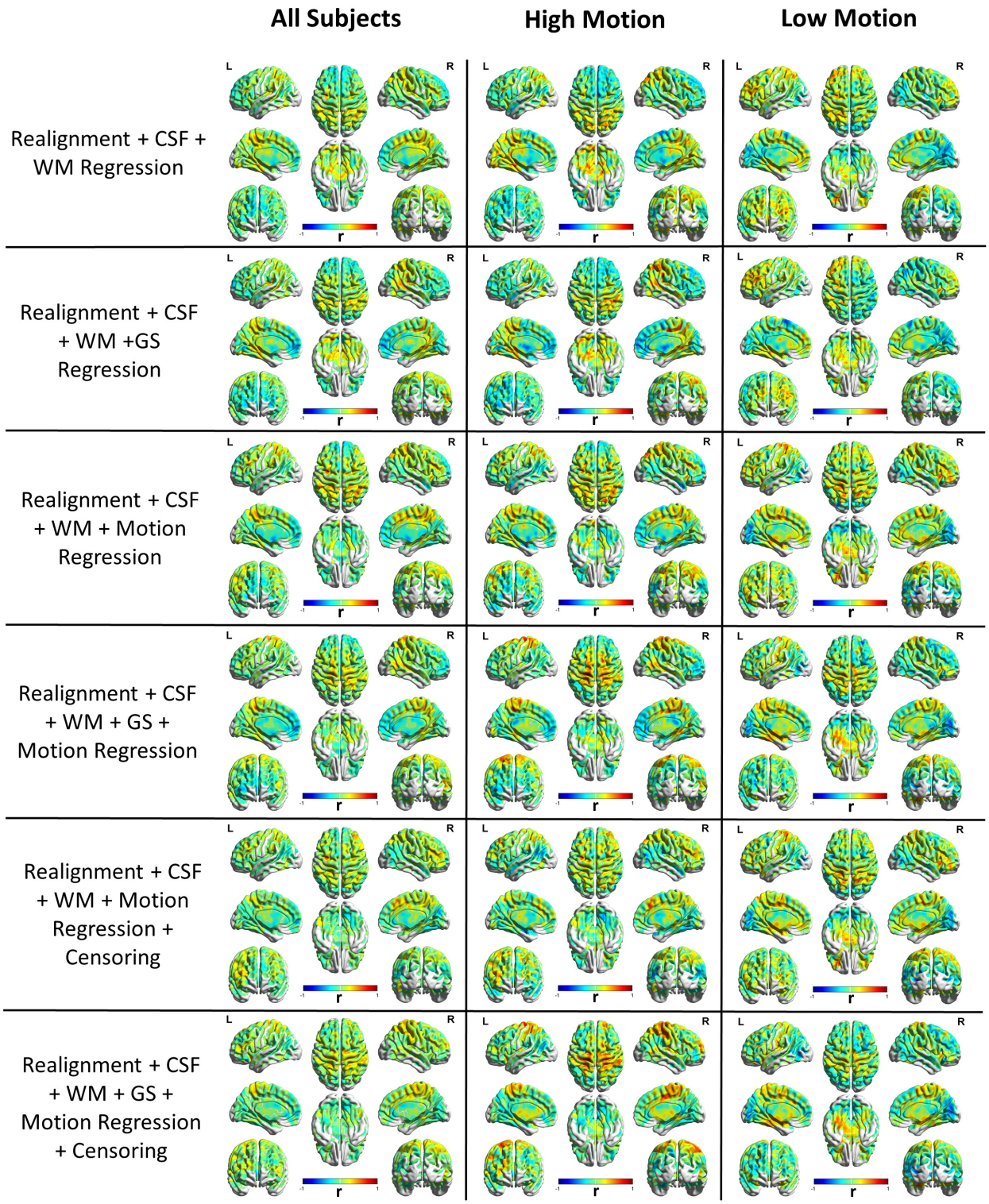


Figure 2.11. Unthresholded spatial map of the Pearson's correlation coefficient between the degree centrality (DC) obtained from PACE-corrected BOLD data and residual head motion as captured by mean [FD_{vox}] across subjects, shown for all subjects (left), in the high motion (middle) and low motion (right) groups separately. Large positive correlations were observed in

the sensorimotor cortex in the low-motion subgroup as well as in the high motion subgroup with nuisance variable regression and censoring. This illustrates that some changes in functional connectivity might have a neural origin and it could be confounded with changes due to motion artifact as even motion artifact causes changes in functional connectivity.

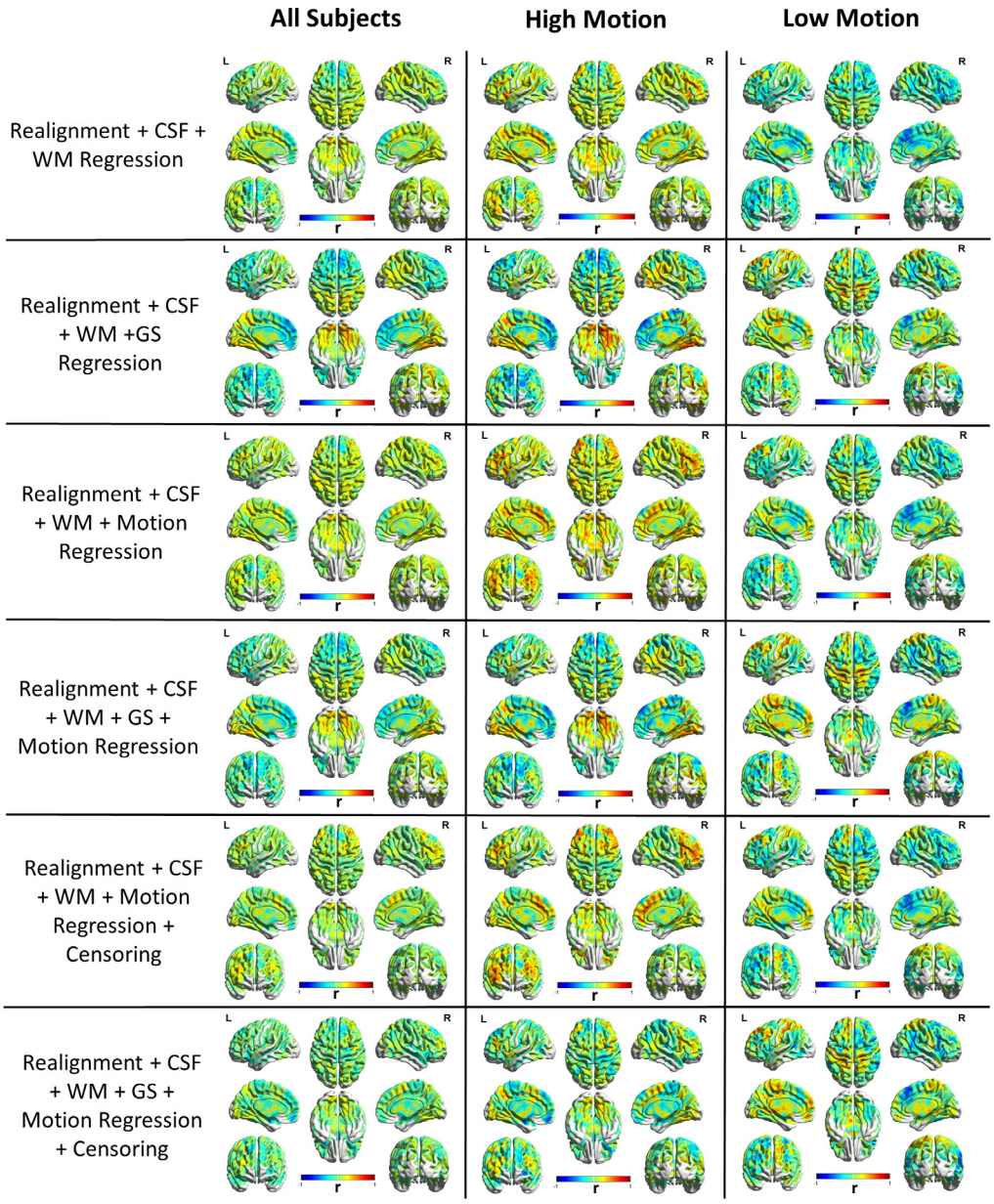


Figure 2.12. Correlation between seed based functional connectivity of PCC (PCC-FC) and head motion (as captured by mean [FD_{vox}] across subjects) shown for all subjects (left), in the high motion (middle) and low motion (right) groups separately. Large correlations were observed across the brain in both low-and high-motion subgroups. With motion censoring and GSR, the

correlations in the high motion group were reduced. This illustrates their relative effectiveness in reducing motion artifacts particularly in subjects with high head motion.

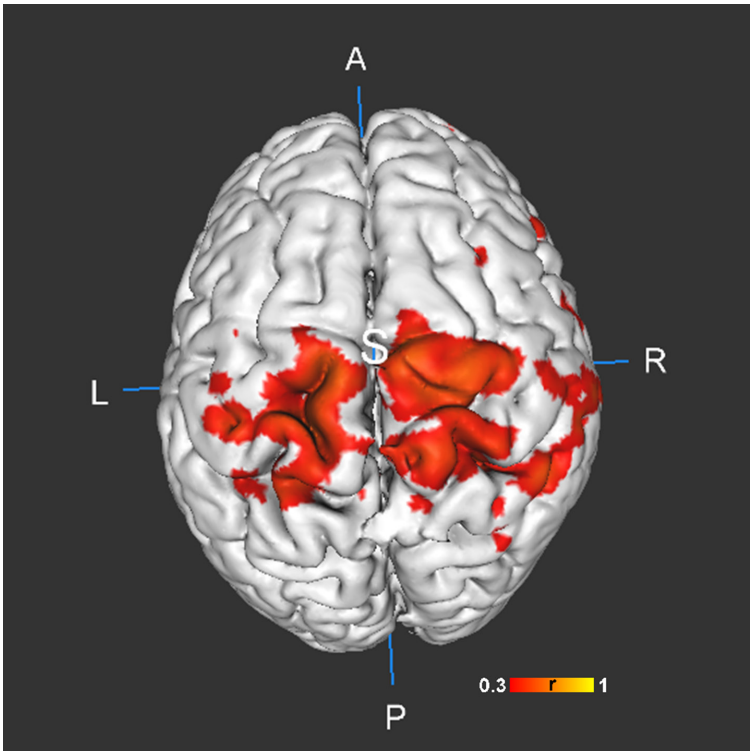


Figure 2.13. Figure showing the thresholded ($p<0.05$) correlation (R) map of degree centrality (DC) with head motion of the brain after nuisance variable regression including CSF, WM, Friston-24 motion regression and motion censoring in all the subjects. Significant positive correlations can been observed between residual head motion in PACE-corrected data and DC in the sensorimotor cortex. This shows that DC in the sensorimotor could possibly be attributed to neural processes responsible for head motion.

A two-tailed t-test was performed across subjects in the high and low motion subgroups separately by using individual subject DC maps as the sample to find consistent patterns across the motion subgroups. For both sub-groups, a similar trend is observed across all the different nuisance regressors used (Figure 2.14). This further demonstrates that degree centrality is robust to various motion correction strategies, and similar results can be obtained with different motion populations. In Figure 2.15, we show a similar result with PCC seed based functional

connectivity. The regions commonly associated with the the Default Mode Network (DMN) were observed in the PCC seed based FC map including regions such as the medial prefrontal cortex (mPFC), inferior parietal lobe (IPL), and lateral temporal cortex (LTC), without the GSR [35, 36, 37]. But with GSR, anti-correlated and task-positive networks such as the dorsal attention system and the hippocampal-cortical memory system were observed as expected [38]. However, it is important to note that in the high-motion subjects with GSR, the mPFC which is an integral part of DMN, was absent, whereas it was present in the low-motion subjects even after GSR. This illustrates that GSR is also likely removing neural components along with motion induced artifacts. Other than mPFC, other significant regions were commonly found in both the low-motion and the high-motion subgroups. Therefore, care must be taken when GSR is used in the preprocessing pipeline in the context of PACE-corrected BOLD data as well.

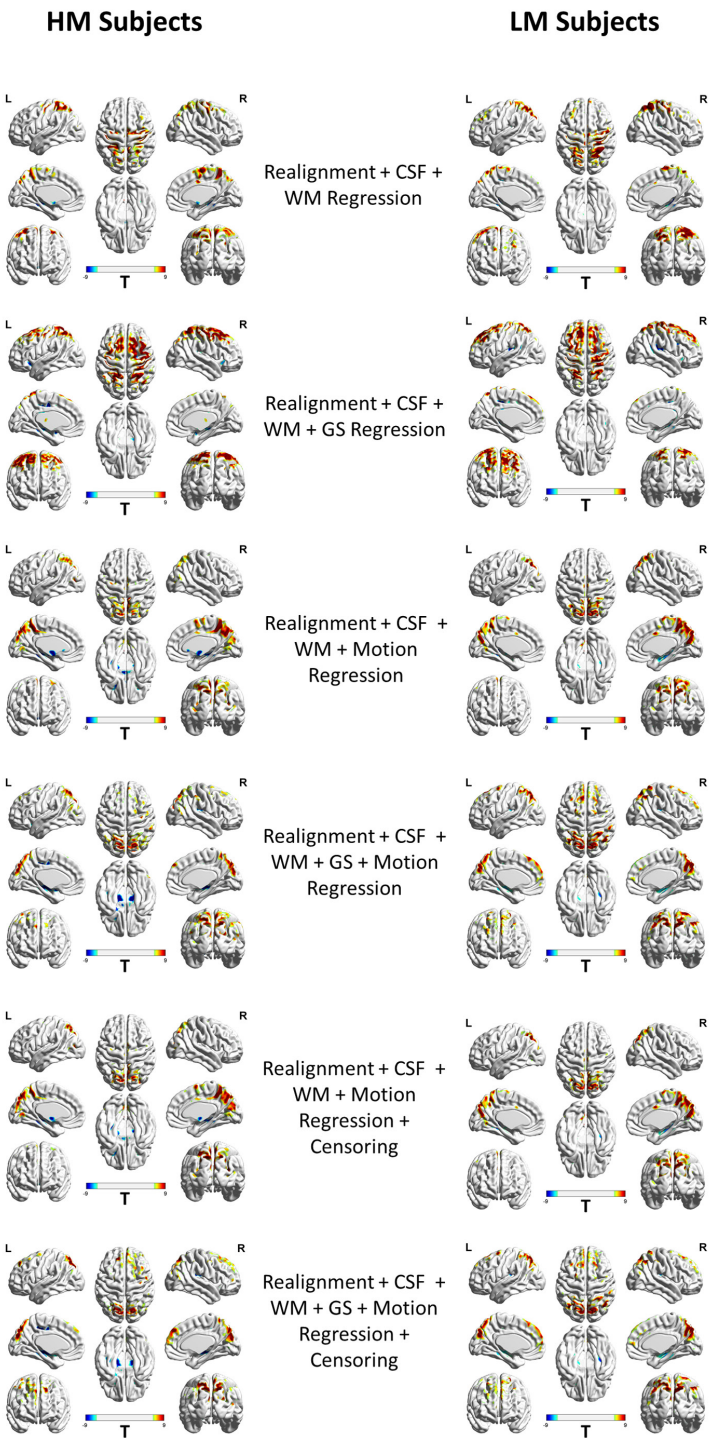


Figure 2.14. A comparison of cortical hubs ($p<0.05$, FDR corrected) as revealed by Degree Centrality (DC). This figure is shown for both high motion and low motion subjects across motion correction strategies. Similar regions as cortical hubs were observed in the high motion and low motion subgroups, with significant regions in the frontal areas, precuneus, cuneus, the mid-brain, sub-lobular regions and the limbic lobe. This shows that degree centrality is robust to

various motion correction strategies and similar results can be obtained with different motion populations.

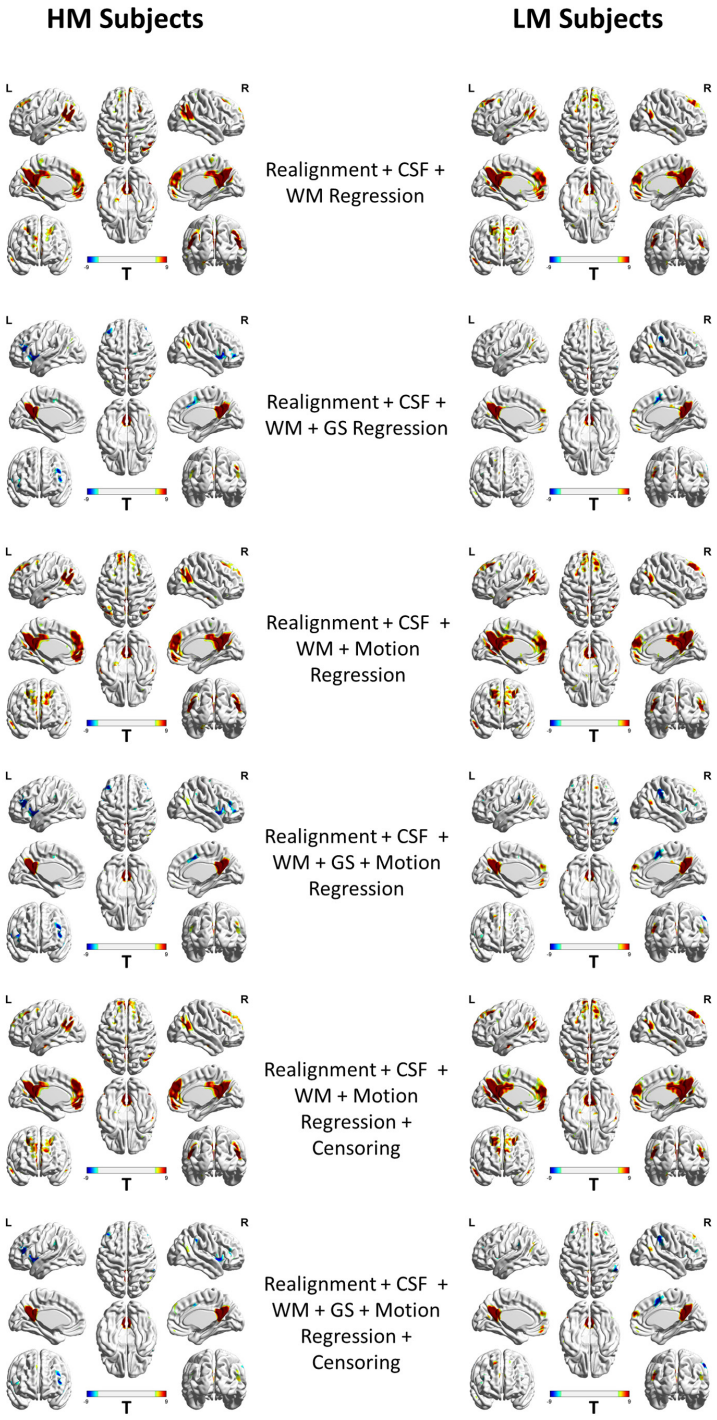


Figure 2.15. A comparison of regions with significant correlations ($p<0.05$, FDR corrected) with posterior cingulate cortex (PCC: 0,-53,26; 10 mm diameter sphere) as the seed region (PCC-FC).

This figure is shown for both high motion and low motion subjects across motion correction strategies. With the addition of GSR, anticorrelated networks were observed. In high-motion subjects with GSR, the correlation between medial prefrontal Cortex (mPFC) and PCC was reduced to chance levels, while it was still present in the low-motion subjects. This illustrates that GSR is also likely removing neural components along with motion induced noise signal.

2.3.6 Motion-BOLD relationships in deconvolved BOLD data

BOLD data used in the deconvolution model was corrected using PACE as well as retrospective techniques such as nuisance variable regression. However, the motion induced distance dependent functional connectivity artifact was present in deconvolved data estimated from raw BOLD data processed without GSR (Figure 2.16A) as well as with GSR (Figure 2.16B), mainly in the high-motion subgroup which is not surprising. However, the low-motion sub-group was relatively free from motion artifacts, as observed from the corresponding motion-BOLD correlations (Figure 2.17 as none of the motion-BOLD relationships achieved significance) as well as the distance dependent artifact (slope of the plot was very small) (Figure 2.16B). Since there are changes in the BOLD signal due to motion, it could potentially affect not just static functional connectivity (SFC) measures, but also other measures such as dynamic functional connectivity estimates (DFC), effective connectivity (EC) and multivariate pattern analyses (MVPA) results. A thorough investigation is required as to how the changes in signal intensity propagate into higher analyses to cause specific and structured artifacts.

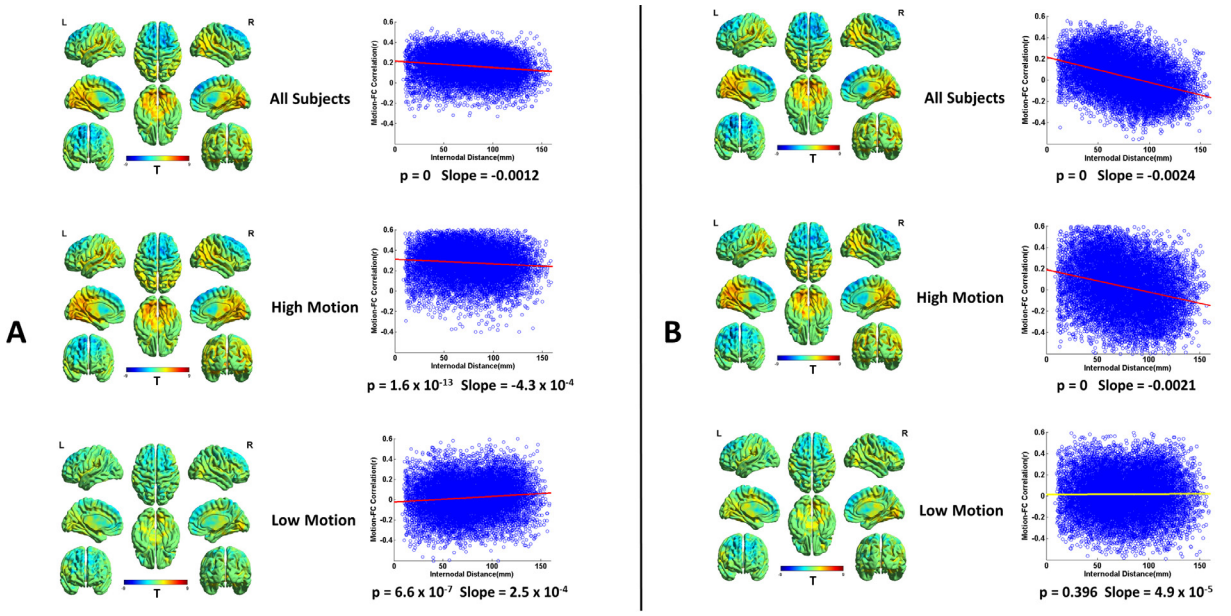


Figure 2.16. Unthresholded Motion-BOLD correlations and FD-RSFC correlations obtained from deconvolved BOLD data, i.e. from latent neural signals. Blind deconvolution was applied to PACE-corrected BOLD data subjected to the following pre-processing steps: realignment, detrending, CSF, WM and motion regression without GSR (A) and with GSR (B). The motion artifacts are still present in the data after deconvolution, mainly in subjects with high motion. GSR seems to distort the FD-RSFC correlations in subjects with high motion, though motion artifacts were absent in the low motion subgroup.

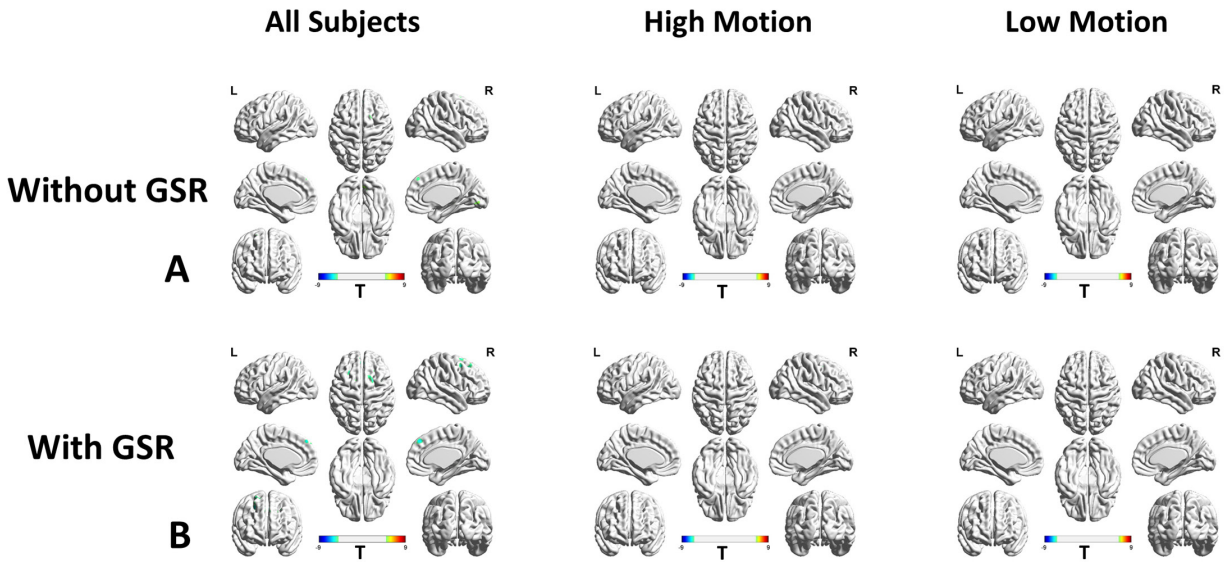


Figure 2.17. Thresholded Motion-BOLD correlations ($p<0.05$, FDR corrected) obtained from deconvolved BOLD data, i.e. from latent neural signals. Blind deconvolution was applied to

PACE-corrected BOLD data subjected to the following pre-processing steps: realignment, detrending, CSF, WM and motion regression without GSR (A) and with GSR (B).

2.3.7 Neural correlates of head motion

In order to identify neural correlates of head motion, we utilized latent neural signals obtained from hemodynamic deconvolution only in the low-motion subject sample, as the data for these subjects were relatively free from motion artifacts as seen in Figure 2.16 and Figure 2.17. In these subjects, we estimated the correlation coefficient between the deconvolved BOLD signal, i.e. latent neural signals, and FD_{vox}. A few significant voxels exceed the threshold ($p<0.001$, uncorrected) without GS regression in the cerebellum. But with GS, many regions which could potentially be associated with neural processes underlying head movements [39], were identified. These included areas in cerebellum, lateral globus pallidus, insula, thalamus, medial frontal gyrus, ventral anterior cingulate and parahippocampal gyrus as shown in Figure 2.18. Although it is possible that these correlations could be artifactual, it is unlikely given the relative absence of motion artifacts in the low-motion sub-group.

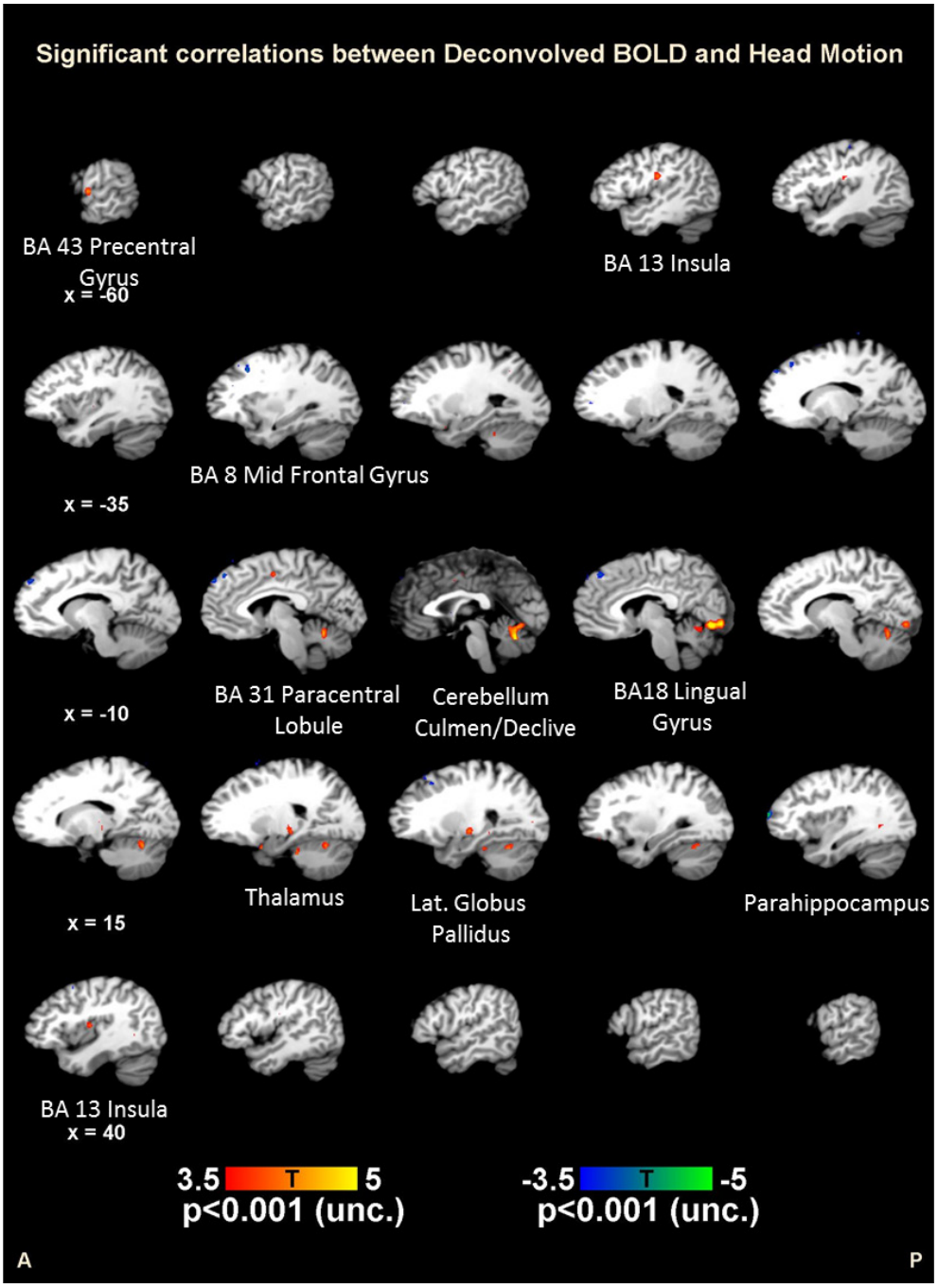


Figure 2.18 Thresholded T maps in the brain that represent significant correlation ($p<0.001$) between the deconvolved BOLD signal (i.e. latent neural signals) and head motion as measured by FD_{vox} across subjects in the low motion subgroup. Some areas such as Thalamus, Cerebellum, Insula, Globus Pallidus might play a role in controlling head movements and might indicate neural effects rather than motion artifacts

2.4 Discussion

This section is organized as follows. First, we discuss the principal advantages of using PACE for controlling head motion artifacts in resting state fMRI data. Next, we discuss the effectiveness of various retrospective motion correction strategies when used in combination with PACE. Subsequently, we discuss identifying and separating neural correlates of head motion from motion artifacts using deconvolved PACE-corrected BOLD data. This is followed by a discussion of other potential retrospective motion correction strategies which might benefit when used in combination with PACE, but which we have not been investigated here. Finally, we discuss some limitations of the current study which need to be kept in mind while interpreting our data.

2.4.1 The principal advantages of prospective motion correction (PACE)

In this study, we examined the effectiveness of PACE in reducing motion artifacts in resting state fMRI data. In combination with the retrospective motion correction methods, using PACE-corrected EPI sequence eliminated most of the motion artifacts. Specifically, we found that PACE provides two primary advantages over conventional EPI sequences. First, PACE was effective in eliminating significant negative motion-BOLD relationships. Significant voxel-wise negative motion-BOLD relationships are typically associated with large signal dropouts caused by relatively large head motion [2, 3] when scanned with a typical EPI sequence. Given the general difficulty in reducing these negative motion-BOLD replationships, PACE may provide a solution to this issue. Second, previous reports have suggested a stringent censoring threshold (FD_{Power}>0.5mm) for satisfactorily controlling the level of motion artifacts in resting state fMRI data [2, 3, 4]. However, with PACE, we found that censoring with a lower threshold (FD_{Power}>0.5mm) and a smaller window around the motion corrupted time-point, provided

qualitatively equivalent reductions in the motion artifact. In fact, with this liberal censoring strategy, we were able to reduce motion artifacts to almost chance levels even in subjects with relatively large motion. This will likely provide significant savings in data which would otherwise be lost to censoring. Practically speaking, the amount of data (in terms of total scan time for resting state fMRI) usually acquired from clinical populations is small given their difficulties in enduring longer scans and also given the fact that resting state scans are tagged onto to other studies which consume a bulk of the allotted imaging time. Given this scenario, acquiring data with PACE-EPI might result in larger amount of usable data and hence more robust analyses.

2.4.2 Effectiveness of retrospective motion correction methods when used in combination with PACE

Since PACE is a prospective motion correction sequence, the best advances in retrospective motion correction can still be used with equal or greater effectiveness when they are combined with PACE. The motion parameters captured are residual motion parameters after motion correction by PACE, not the actual subject motion. The ability of CSF and WM regression in removing the motion induced signal variance in resting state fMRI data is limited as reported by previous studies [2, 3], a fact confirmed by our results. We used motion regression by the Friston-24 model, which was shown to be the best performing model previously [3, 2] and our results confirm the same. Obviously, higher-order motion models might explain larger amount of variance for high-motion datasets across the brain, but it comes at the cost of significant loss of degrees of freedom and result in a drop in the BOLD sensitivity [40].

Though several previous studies recommend the use of GSR for reducing motion artifacts [4, 2, 33, 8], the effectiveness of GSR in reducing motion artifacts in the BOLD signal as well as

lowering FD-RSFC correlations is mixed. Our results (Figure 2.5) are in agreement with the previous studies indicating that GSR effectively reduces the positive motion-BOLD relationships but increases the negative motion-BOLD relationships [2]. GSR also distorted the FD-RSFC correlations (Figure 2.8) considerably [34]. GSR reduced the functional connectivity of medial prefrontal cortex (mPFC) with the PCC seed (Figure 2.15), a key component of the default mode network, in the high-motion subgroup. Other issues with GSR include the fact that it distorts the distribution of correlation values [5] and could alter inter-individual differences at the group Level [6, 41]. Given that PACE provides an additional strategy for motion correction, it could be used without GSR to achieve better quality data compared to conventional EPI coupled with no GSR. On the otherhand, for the proponents of GSR, PACE's tendency to remove negative motion-BOLD relationships may at least partially cancel out the negative motion-BOLD relationships introduced by GSR.

We found censoring high-motion time-points from the data to be the most effective retrospective motion correction. With censoring, spurious motion-BOLD relationships (Figure 2.7) and distance dependent functional connectivity artifacts (Figure 2.9) were almost eliminated in high-motion subjects. An extremely important issue, when performing censoring is to determine how much resting state data is sufficient for stable and reliable estimation of resting-state functional connectivity (RSFC) metrics. Some have suggested at least 4 minutes [3] and others believe that 3 minutes of RS-fMRI data to be sufficient [2]. While comparing usable data available after censoring with PACE and traditional EPI (Table 2.1), we have assumed that one has to have at least 3 minutes of data. In addition to scan time, the sampling period (TR) is also an important consideration. The value of the FD, used for identifying motion corrupted time-points is paramount while censoring as it is heavily dependent on TR. Sampling the brain at a smaller TR

tends to divide larger motion into smaller components, hence might have different effects on the presence of motion artifacts and motion correction. Also, censoring alters the temporal structure of the data even if the censored time points are interpolated. This affects frequency based analyses, moving window-based dynamic functional connectivity, and effective connectivity calculations. So all analyses which require an intact temporal structure of the data might want to avoid censoring. In such cases, PACE offers a way of obtaining relatively cleaner data without censoring, although motion artifacts cannot be completely eliminated without at least liberal censoring even when using PACE.

The effectiveness of group-level motion correction by including individual motion estimates in group-level analyses has been reported before [2, 4]. Group level regression with individual motion estimates effectively removes potential motion related artifacts, but may also remove changes related to neural activity [33]. Many pathological conditions are associated with changes in regional functional connectivity. These changes in connectivity might be biased by the group effects of the subject head motion especially in hyperkinetic populations. So it might be difficult to separate motion artifacts from disease effects, especially since the effect of interest is correlated with head motion. Unfortunately, in these cases group level motion correction cannot be performed, so motion correction has to be limited to subject-level motion correction methods.

As recommended by several previous papers, we think that there are merits to having different preprocessing pipelines for groups with different motion profiles as well as when performing different analyses, as no single preprocessing procedure is ideal for all cases. Some factors which need to be considered for the acquisition and processing of RS-fMRI data include the repetition time TR, use of slice time correction, the imaging sequence to capture the BOLD signal, head motion criteria to include a subject fMRI data in the study, the motion profile of the sample and

the population to be studied, the model complexity for modeling head motion, the use of global signal regression, the threshold used to decide motion corrupted volumes, the number of time-points left after motion censoring required for the stable estimation of RSFC metrics, the use of subject-level motion correction vs. group-level motion correction, and the use of group-level motion correction, if the variable of interest is correlated with head motion. The second important factor to consider is the objective and analysis of the study. As we have discussed earlier, motion censoring effectively precludes many types of analyses such as the ones that use hemodynamic deconvolution. Though interpolation has been suggested to reconstruct the removed timepoints, the fit could be unreliable as the neighboring points of a motion corrupted time point may also be corrupted by motion since multiple timepoints are affected by head motion. Another examples relates to the use of group-level motion correction in analyses involving clinical populations, especially in hyperkinetic populations where disease status is associated with head motion. Group level correction of head motion might remove some of the disease related variance. Therefore, a proper choice of the processing pipeline based on the motion profile and the planned analyses can reduce motion artifacts while still achieving study objectives.

2.4.3 Identifying and separating neural effects from motion artifacts

It is expected that in RS-fMRI experiments, some low-frequency BOLD fluctuations could potentially correlate with head motion because of the latter's neural origin [8]. In our results (Figure 2.11, Figure 2.13), we did observe correlations between head motion and degree centrality in the motor cortex, which is unlikely to be solely due to motion artifacts. Not just at the connectivity level but even at the BOLD signal level, we observed (Figure 2.18) that the deconvolved BOLD signal (i.e. latent neural signals which are devoid of hemodynamic delay

due to head motion) correlated with FD_{vox} in several regions in the brain that play an important role in the neural control/execution of head movements. The use of motion derived regressors in preprocessing, might reduce the signal variance associated with neural correlates of head motion. This could impact our ability to identify neural effects of head motion. The relationship between head motion and brain connectivity is bi-directional relationship, i.e. differences in brain connectivity could be associated with head motion in the scanner [10], just as head motion could cause artifactual changes in connectivity. Some have hypothesized that this might suggest that reduced connectivity in regions corresponding to the default mode network might predict how still the person can stay in the scanner [10]. These neural correlates of motion can cause functional connectivity changes that represent genuine variations of neural activity in certain regions, which can be mistaken for a motion artifact. Other areas such as the regions in the visual cortex have also been speculated to be a neural correlate of head motion [33]. Different clinical populations exhibit characteristic spatio-temporal motion patterns that can be associated with distinct motion artifacts for various pathological conditions [33], thus really complicating the distinction between disease changes in connectivity and motion artifacts and limiting the use of functional connectivity as effective disease biomarkers [42]. Given this scenario, it is all the more advantageous to prospectively correct for motion so that the resulting data undergoes as little retrospective correction as possible, so that the component of motion-related changes that may represent system-specific neural activity are preserved.

The observed BOLD signal is a convolution of the latent neural fluctuations with the Hemodynamic Response Function (HRF). Resting state BOLD data could be deconvolved [28] to remove the spatial heterogeneity in the latency of the HRF. The fidelity of deconvolution can be affected not only by motion censoring (scrubbing), but also when motion artifacts are present

in the data. Therefore, sufficient subject level motion correction must be performed at the individual level, and it should be ensured that the data is free from motion artifacts before deconvolution is performed to infer the latent neuronal activity. In subjects with low-head motion, PACE in combination with nuisance signal regression was successful in elimination of motion effects as spurious motion-BOLD relationships were eliminated (Figure 2.17) and the slope of the plot illustrating motion induced distance dependent connectivity artifact is very small (Figure 2.16). Since there are changes in the BOLD signal due to motion, it could potentially affect not just static functional connectivity (SFC) measures, but also other measures such as dynamic functional connectivity estimates (DFC), effective connectivity (EC) and multivariate pattern analyses (MVPA) results. A thorough investigation is required as to how the changes in signal intensity propagate into higher analyses to cause specific and structured artifacts.

2.4.4 Other motion correction methods

Many advances in retrospective motion correction methods, which involve slight modifications in the traditional preprocessing pipeline, have been reported to be beneficial in ameliorating motion artifacts. These methods can be used in combination with PACE for more effective reduction of motion artifacts. They include the usage of time series based or wavelet-based despiking [43], using aCompCor (anatomical ComCor) [44] for nuisance signal regression instead of mean CSF and WM signals, using edge voxel information rather than traditional motion parameters [45], ANATICOR, which uses local white matter regressors coupled with despiking [34] and ensures uniform smoothing in the entire data to further reduce the effects of inter-individual differences in head motion [46]. The voxel-wise estimates of head motion are derived from volume–based realignment parameters and their accuracy is limited by the accuracy

of the estimation of the volumetric realignment parameters. Therefore, slice wise parameter measures might give a better estimate of the actual voxel-wise motion for every voxel in the brain. As rapid head movements between TRs can have a differential effect on different slices in a single volume and cannot be adequately modeled by volume based realignment parameters, use of these slice-wise estimates may aid in the calculation of voxel-wise displacements and correction of motion induced signal changes [40]. While comparing motion-prone clinical populations with healthy controls at the group level, the use of Regional Displacement Interaction (RDI), which would encapsulate motion information in the voxel-wise metrics rather than use a summary motion statistic could further correct for motion artifacts and preserving neuronal effects [42].

2.5 Limitations

The number of subjects we have used for this study (N=47) is reasonable for typical fMRI studies, but small compared to other reports which have evaluated restrospective strategies using large databases (N>100). Due to the nature and effect sizes of motion artifacts, sample size can have a bearing on the results. Therefore, our results should be confirmed with a larger sample. Also, phenotypic factors such as age can have a bearing on motion artifacts. Our sample was homogeneous in this respect (20 male/27 females, age 25.1 ± 5 years) and hence did not sample the entire spectrum observed in the general population. Since we did not use external motion tracking devices to quantify head motion, the accuracy and reliability of image-based motion metrics used for prospective and retrospective correction of head motion could not be independently validated. Since PACE is a prospective motion correction method, we might not know the actual head movement of the subject, but only the residual motion of the subject on the

scanner coordinates. Therefore, it is impossible to directly compare data with and without PACE correction in a time-locked manner.

2.6 Bibliography

[1] K. Friston, S. Williams, R. Howard, R. Frackowiak and R. Turner, "Movement-related effects in fMRI time-series," *Magn Reson Med.,* vol. 35, no. 3, pp. 346-55, 1996.

[2] C.-G. Yan, B. Cheung, C. Kelly, S. Colcombe, R. C. Craddock, A. D. Martino, Q. Li, X.-N. Zuo, F. X. Castellanos and M. P. Milham, "A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics," *NeuroImage,* vol. 76, pp. 183-201, 2013.

[3] T. D. Satterthwaite, M. A. Elliott, R. T. Gerraty, K. Ruparel, J. Loughead, M. E. Calkins, S. B. Eickhoff, H. Hakonarson, R. C. Gur, R. E. Gur and D. H. Wolf, "An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data," *NeuroImage,* vol. 64, pp. 240-256, 2013.

[4] J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar and S. E. Petersen, "Methods to detect, characterize, and remove motion artifact in resting state fMRI," *NeuroImage,* vol. 84, pp. 320-341, 2014.

[5] K. Murphy, R. M. Birn, D. A. Handwerker, T. B. Jones and P. A. Bandettini, "The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced?," *NeuroImage,* vol. 44, no. 3, pp. 893-905, 2009.

[6] Z. S. Saad, S. J. Gotts, K. Murphy, G. Chen, H. J. Jo, A. Martin and R. W. Cox, "Trouble at Rest: How Correlation Patterns and Group Differences Become Distorted After Global Signal Regression," *Brain Connectivity,* vol. 2, no. 1, pp. 25-32, 2012.

[7] J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar and S. E. Petersen, "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion," *NeuroImage,* vol. 59, no. 3, pp. 2142-2154, 2012.

[8] J. D. Power, B. L. Schlaggar and S. E. Petersen, "Recent progress and outstanding issues in motion correction in resting state fMRI," *NeuroImage,* vol. 105, pp. 536-551, 2015.

[9] L. Lemieux, A. Salek-Haddadi, T. E. Lund, H. Laufs and D. Carmichael, "Modelling large motion events in fMRI studies of patients with epilepsy," *Magnetic Resonance Imaging,* vol. 25, no. 6, pp. 894-901, 2007.

[10] L.-L. Zeng, D. Wang, M. D. Fox, M. Sabuncu, D. Hu, M. Ge, R. L. Buckner and H. Liu, "Neurobiological basis of head motion in brain imaging," *PNAS,* vol. 111, no. 16, pp. 6058-

6062, 2014.

[11] D. A. Fair, J. T. Nigg, S. Iyer, D. Bathula, K. L. Mills, N. U. F. Dosenbach, B. L. Schlaggar, M. Mennes, D. Gutman, S. Bangaru, J. K. Buitelaar, D. P. Dickstein, A. Di Martino, D. N. Kennedy, C. Kelly, B. Luna, J. B. Schweitzer, K. Velanova, Y.-F. Wang, S. Mostofsky, F. X. Castellanos and M. P. Milham, "Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data," *Frontiers in Systems Neuroscience,* vol. 6, p. 80, 2013.

[12] M. GOTO, O. ABE, T. MIYATI, H. YAMASUE, T. GOMI and T. TAKEDA, "Head Motion and Correction Methods in Resting-state Functional MRI," *Magnetic Resonance in Medical Sciences,* vol. 15, no. 2, pp. 178-186, 2016.

[13] N. Todd, O. Josephs, M. F. Callaghan, A. Lutti and N. Weiskopf, "Prospective motion correction of 3D echo-planar imaging data for functional MRI using optical tracking," *NeuroImage,* vol. 113, pp. 1-12, 2015.

[14] M. Zaitsev, C. Dold, G. Sakas, J. Hennig and O. Speck, "Magnetic resonance imaging of freely moving objects: prospective real-time motion correction using an external optical motion tracking system," *NeuroImage,* vol. 31, no. 3, pp. 1038-1050, 2006.

[15] O. Melvyn B., K. Sascha, M. Jordan, W. J. Thomas and T. R. Brown, "Echo-planar imaging with prospective slice-by-slice motion correction using active markers," *Magn. Reson. Med.,* vol. 66, no. 1, pp. 73-81, 2011.

[16] D. Rotenberg, M. Chiew, S. Ranieri, F. Tam, R. Chopra and S. J. Graham, "Real-time correction by optical tracking with integrated geometric distortion correction for reducing motion artifacts in functional MRI," *Magn Reson Med,* vol. 69, no. 3, p. 734–748, 2013.

[17] C. Forman, M. Aksoy, J. Hornegger and R. Bammer, "Self-encoded marker for optical prospective head motion correction in MRI," *Medical Image Analysis,* vol. 15, no. 5, pp. 708-719, 2011.

[18] M. Zaitsev, B. Akin, P. LeVan and B. R. Knowles, "Prospective motion correction in functional MRI," *NeuroImage,* 2016.

[19] S. Thesen, O. Heid, E. Mueller and L. R. Schad, "Prospective acquisition correction for head motion with image-based tracking for real-time fMRI," *Magn. Reson. Med,* vol. 44, no. 3, pp. 457-465, 2000.

[20] T. D. Satterthwaite, D. H. Wolf, J. Loughead, K. Ruparel, R. Kosha, M. A. Elliott, H. Hakonarson, R. C. Gur and R. E. Gur, "Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth," *NeuroImage,* vol. 60, no. 1, pp. 623-632, 2012.

[21] C. Yan and Y. Zang, "DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI," *Frontiers in Systems Neuroscience,* vol. 4, p. 13, 2010.

[22] M. Jenkinson, P. Bannister, M. Brady and . S. Smith, "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images," *NeuroImage,* vol. 17, no. 2, pp. 825-841, 2002 October 2002.

[23] K. R. Van Dijk, M. R. Sabuncu and R. L. Buckner, "The influence of head motion on intrinsic functional connectivity MRI," *NeuroImage,* vol. 59, no. 1, pp. 431-438, 2012.

[24] N. U. F. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, K. A. Barnes, J. W. Dubis, E. Feczko, R. S. Coalson, J. R. Pruett, D. M. Barch, S. E. Petersen and B. L. Schlaggar, "Prediction of Individual Brain Maturity Using fMRI," *Science,* vol. 329, no. 5997, p. 1358–1361, 2010.

[25] X.-N. Zuo, R. Ehmke, M. Mennes, D. Imperati, X. F. Castellanos, O. Sporns and M. P. Milham, "Network Centrality in the Human Functional Connectome," *Cereb. Cortex,* vol. 22, no. 8, pp. 1862-1875, 2012.

[26] R. L. Buckner, J. Sepulcre, T. Tanveer, F. M. Krienen, H. Liu, T. Hedden, J. R. Andrews-Hanna, R. A. Sperling and K. A. Johnson, "Cortical Hubs Revealed by Intrinsic Functional Connectivity: Mapping, Assessment of Stability, and Relation to Alzheimer's Disease," *J Neurosci.,* vol. 29, p. 1860–1873, 2009.

[27] D. Rangaprakash, G. Deshpande, T. Daniel, A. Goodman, J. Katz, N. Salibi, T. Denney Jr. and M. M. Dretsch, "Static and Dynamic Functional Connectivity Impairments in Concussed Soldiers with and without PTSD," in *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine*, Toronto, 2015.

[28] G.-R. Wu, W. Liao, S. Stramaglia, J.-R. Ding, H. Chen and D. Marinazzo, "A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data," *Medical Image Analysis,* vol. 17, no. 3, pp. 365 - 374, 2013.

[29] G. K. Aguirre, E. Zarahn and M. D'Esposito, "The Variability of Human, BOLD Hemodynamic Responses," *NeuroImage,* vol. 8, p. 360–369, 1998.

[30] D. A. Handwerker, J. M. Ollinger and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *NeuroImage,* vol. 21, no. 4, pp. 1639-1651, 2004.

[31] M. B. Schippers, R. Renken and C. Keysers, "The effect of intra- and inter-subject variability of hemodynamic responses on group level Granger causality analyses," *NeuroImage,* vol. 57, no. 1, pp. 22-36, 2011.

[32] D. Ibai, E. Asier, E. Iñaki, B. Mateos, A. Cabrera, D. Marinazzo, E. J. Sanz-Arigita, S. Stramaglia, J. M. Cortes Diaz and f. t. A. D. N. Initiative, "Information Flow Between Resting-State Networks," *Brain Connectivity,* vol. 5, no. 9, pp. 554-564, 2015.

[33] J. Pujol, D. Macià, L. Blanco-Hinojo, G. Martínez-Vilavella, J. Sunyer, R. de la Torre, A. Caixàs, R. Martín-Santos, J. Deus and B. . J. Harrison, "Does motion-related brain functional connectivity reflect both artifacts and genuine neural activity?," *NeuroImage,* vol. 101, pp. 87-95, 2014.

[34] J. H. Jo, S. J. Gotts, R. C. Reynolds, P. A. Bandettini, A. Martin, R. W. Cox and Z. S. Saad, "Effective Preprocessing Procedures Virtually Eliminate Distance-Dependent Motion Artifacts in Resting State FMRI," *Journal of Applied Mathematics,* vol. 2013, 2013.

[35] H. Koshino, T. Minamoto, K. Yaoi, M. Osaka and N. Osaka, "Coactivation of the Default Mode Network regions and Working Memory Network regions during task preparation," *Scientific Reports,* vol. 4, no. 5954, 2014.

[36] R. L. Buckner, J. R. Andrews-Hanna and D. L. Schacter, "The Brain's Default Network," *Annals of the New York Academy of Sciences,* vol. 1124, no. 1, p. 1–38, 2008.

[37] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard and G. L. Shulman, "A default mode of brain function," *PNAS,* vol. 98, no. 2, pp. 676-682, 2001.

[38] M. D. Fox , A. Z. Snyder, J. . L. Vincent, M. Corbetta, D. C. Van Essen and M. E. Raichle, "The human brain is intrinsically organized into dynamic, anticorrelated functional networks," *PNAS,* vol. 102, no. 27, pp. 9673-9678, 2005.

[39] C. N. Prudente, R. Stilla, C. M. Buetefisch, S. Singh, E. J. Hess, X. Hu, K. Sathian and H. Jinnah, "Neural Substrates for Head Movements in Humans: A Functional Magnetic Resonance Imaging Study," *Journal of Neuroscience,* vol. 35, no. 24, pp. 9163-9172, 2015.

[40] E. B. Beall and M. J. Lowe, "SimPACE: Generating simulated motion corrupted BOLD data with synthetic-navigated acquisition for the development and evaluation of SLOMOCO: A new, highly effective slicewise motion correction," *NeuroImage,* vol. 101, pp. 21-34, 2014.

[41] S. J. Gotts, Z. S. Saad, H. J. Jo, G. L. Wallace, R. W. Cox and A. Martin, "The perils of global signal regression for group comparisons: a case study of Autism Spectrum Disorders," *Front. Hum. Neurosci.,* vol. 7, 2013.

[42] T. Spisák, A. Jakab, S. A. Kis, G. Opposits, C. Aranyi, E. Berényi and M. Emri, "Voxel-Wise Motion Artifacts in Population-Level Whole-Brain Connectivity Analysis of Resting-State fMRI," *PLOS ONE,* vol. 9, no. 9, p. e104947, 2014.

[43] A. X. Patel, P. Kundu, M. Rubinov, P. S. Jones, P. E. Vértes, K. D. Ersche, J. Suckling and E. T. Bullmore, "A wavelet method for modeling and despiking motion artifacts from

resting-state fMRI time series," *NeuroImage,* vol. 95, pp. 287-304, 2014.

[44] J. Muschelli, M. B. Nebel, B. S. Caffo, A. D. Barber, J. J. Pekar and S. H. Mostofsky, "Reduction of motion-related artifacts in resting state fMRI using aCompCor," *NeuroImage,* vol. 96, pp. 22-35, 2014.

[45] R. Patriat, E. K. Molloy and R. M. Birn, "Using Edge Voxel Information to Improve Motion Regression for rs-fMRI Connectivity Studies," *Brain Connectivity,* vol. 5, no. 9, pp. 582-595, 2015.

[46] D. Scheinost, X. Papademetris and R. T. Constable, "The impact of image smoothness on intrinsic functional connectivity and head motion confounds," *NeuroImage,* vol. 95, pp. 13-21, 2014.

Chapter 3

Supervised Machine Learning for Neuroimaging-based Diagnostic Classification

Abstract

With recent advances in neuroimaging, machine learning and the availability of large datasets, the field of neuroimaging has moved from identifying group differences of univariate measures to single subject predictions using multivariate analyses. With high classification accuracies reported using single-site data acquisitions and relatively lower classification accuracies with multisite acquisitions, there are growing concerns about the generalizability of machine learning classifiers across different demographic/phenotypic variables including acquisition sites and age groups. To evaluate the generalizability of machine learning classifiers across heterogeneous populations, we investigated four neurological diseases: Autism Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD), Post-traumatic Stress Disorder (PTSD) and Alzheimer's Disease (AD). We applied 18 different types of machine learning classifiers based on diverse principles to datasets where the training/validation and the hold-out test data belonged to samples with the same diagnosis but differing in either the age range or the acquisition site. Our results indicate that overfitting can be a huge problem in homogeneous datasets, especially with fewer samples, leading to inflated measures of accuracy that fail to generalize well to the general disease population. Further, different classifiers tend to perform well on different datasets. In order to address this, we propose a consensus classifier by combining the predictive power of all 18 classifiers. The consensus classifier was less sensitive to unmatched training/validation and hold-out test data. Finally, we combined feature importance scores obtained from all classifiers to infer the discriminative ability of connectivity features. The functional connectivities thus identified were robust to classification algorithm used, age and

acquisition site differences, had diagnostic predictive ability in addition to statistical separation between the groups. The connectivities thus identified could provide robust inferences about the neural basis of underlying disorders.

3.1 Introduction

Currently, the identification of many neurological disorders are based on subjective diagnostic criteria. The development of objective diagnostic tools is a work in progress in the field of neuroimaging with many promising leads. Univariate between-group differences in neuroimaging between healthy controls and clinical populations are not yet sufficiently predictive of disease states at the individual level. For automated disease diagnosis, a machine learning classifier is trained to model the relationship between features extracted from brain imaging data and the disease labels of individuals in the training dataset (the disease labels are typically determined via clinical assessment by a licensed physician) and the model is then used to predict the diagnostic label of a new and unseen subject drawn from a test dataset. However, many challenges to this paradigm being employed in practice remain. A few of them are: (i) Lack of availability of large clinical imaging datasets, (ii) Challenges in generalizing results across study populations, (iii) Difficulty in identifying reliable image-based biomarkers which are robust to progress and maturation of the disease, and (iv) Variability in classifier performance. Many of these issues are interrelated. In fact the ultimate goal of machine learning based diagnostic classification is not just to achieve high classification accuracy but also good generalizability to unseen data with varying characteristics [1]. To be useful in clinical settings, machine learning classifiers should be generalizable to the wider population and this can be achieved by including data from several imaging sites [2].

The main reason for the failure of identification of precise neuroimaging based disease biomarkers, despite high accuracies reported in many neuroimaging studies is that many of these studies use small, biologically homogenous samples and generalizing their results to larger heterogeneous disease populations is difficult. Single study analyses in which the training and the test data are from the same acquisition site gives higher classification accuracies than in the cases when they are from distinct imaging sites [3]. A classifier that works well on a particular dataset might fail to classify with good accuracy on a different dataset [2]. These prior findings indicate that a classifier may achieve high accuracy in a given data set even with cross-validation, but the accuracy may drop significantly when the classifier is used on a more general population which was not used in cross-validation as was observed with Autism [4]. Generalizability of the classifiers cannot be assessed using very few samples from a single site but can be shown by including data from various imaging sites. Classifiers which perform well on small training sets generalize poorly, and hidden correlations in the training and validation sets might lead to overoptimistic performance of the classifier [5]. This is also borne out by the observation that overall performance accuracy decreases with sample size [6]. Hence we should be extremely cautious in interpreting over-optimistic classification performance results from small datasets.

Classification across heterogeneous populations with considerable variation in demographic and phenotypic profiles, although desirable for generalizability, is extremely challenging, particularly when neuroimaging data is pooled from multiple acquisition sites [7]. Variance introduced in the data due to scanner hardware, imaging protocols, operator characteristics, demographics of the regions and other factors that are acquisition site specific, can affect the classification performance. The image-based biomarkers thus identified must be reliable and consistent across

imaging sites and age ranges to be useful clinically. The added difficulty with the plasticity, and the dynamic adaptability of the brain further complicates the use of brain-derived biomarkers for classification as the connectivity trajectories might diverge with disease maturation. This further increases the difficulty in determining imaging based biomarkers with certainty and reliability across populations [8].

Given the difficulties in disease classification with multisite studies, appropriate choice of features which are reliable and sensitive to underlying disease is the primary motivating factor in our choice of resting-state functional connectivity (RSFC) as features. RSFC measures the spontaneous low-frequency fluctuations between remote regions in the brain in baseline functional magnetic resonance imaging (fMRI) data and is typically estimated using the Pearson's correlation coefficient. It has been extensively used to characterize the functional architecture of the brain both in healthy and clinical populations. Consistency and reliability of RSFC measures across subjects and scanning sites are of prime importance for its use in disease classification. RSFC has been shown to have moderate to high reliability and reproducibility across healthy [9, 10, 11, 12, 13, 14, 15, 16, 17], clinical [18, 19], pediatric [19] and elderly [20, 21] populations. It has also been shown to have long-term test-retest reliability [22, 23, 24]. Resting-state functional connectivity is altered in clinical populations such as Attention Deficit Hyperactivity Disorder (ADHD), Depression, Autism, Schizophrenia, Post-traumatic Stress Disorder (PTSD) and Alzheimer's disease (AD). Hence there is growing optimism in the field that modulations in RSFC can help us understand the pathogenesis behind several neurological and psychiatric disorders due to its sensitivity to changes in development, aging and disease progression. These factors, combined with the ability to standardize protocols, have paved the way for data aggregation across multiple sites leading to increased statistical power and the

generalizability of the findings, and have catapulted RSFC into increasing prominence for diagnostic classification. Given the relatively lower prevalence of certain disorders and given the costs and time associated with aggregating large datasets, easier pooling of data from multiple sites is critical. RSFC protocols are simple to run with little overheads and hence it has been typically added into different imaging protocols employing various clinical populations. Also, the lack of necessity to comply with task instructions particularly in uncooperative clinical populations, there is a considerable rise in interest in the use of resting-state fMRI (Rs-fMRI) in patients with brain disorders [8].

With the advent of big data initiatives such as Autism Brain Imaging Database (ABIDE), where a large amount of data is collected from multiple sites, there is renewed optimism for accurate and reliable disease classification [2]. Generalizability of classifier performance can be increased, simulataneously avoiding overfitting, when we have large training data sizes. Another consequence of such big data initiatives and exploratory data analyses is that reliable and repeatable studies for testing novel hypotheses about the identification of relevant clinical biomarkers has taken ground. In this study we used RSFC measures to examine their efficacy in diagnostic classification using 18 different classifiers in 4 disease populations: (i) Autism Brain Imaging Data Exchange (ABIDE) for Autism Spectrum Disorder (ASD), (ii) ADHD-200 dataset for Attention Deficit Hyperactivity Disorder (ADHD), (iii) PTSD data which were acquired at the Auburn MRI Research Center for Post-Concussion Syndrome (PCS) and Post-traumatic Stress Disorder (PTSD) and, (iv) Alzheimer's Disease Neuroimaging Initiative (ADNI) for Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD). ABIDE and ADHD-200 datasets have more than 500 subjects whereas ADNI and PTSD data have around 100 subjects. This way, we were able to test the generalizability of classifiers under various conditions: (a) Using various

disorders whose origins are likely different, (b) Using both smaller and larger size of datasets, (c) Using data obtained from both multiple sites as well as single-site, and finally, (d) Using both homogeneous and heterogeneous samples from the population.

There are three primary goals of this paper. The first goal is to understand the generalizability of machine learning classifiers in the presence of disease and population heterogeneity, variability in disease across age, and variations in data caused by multisite acquisitions. We report a biased estimation of cross-validation accuracy and an unbiased estimate of performance on a completely independent and blind hold-out test dataset. The entire datasets were split into training/validation and hold-out test data (with both splits containing both controls and clinical populations) and the cross-validation accuracy was estimated using the training/validation data by splitting it further into training data and validation data. The hold-out test datasets were constructed under three different scenarios: (i) subjects with different, non-overlapping age range compared to training/validation data, (ii) subjects drawn from different imaging sites compared to training/validation data and, (iii) training/validation and hold-out test data matched on all demographics including age as well as acquisition site. We hypothesized that testing our classifiers on homogenous populations could give us optimistic estimates of classifier performance, which might not generalize well to the real world classification scenarios encountered in the clinic. Therefore, by comparing a holdout test data with the same disease diagnosis and matched in age and acquisition site as well as unmatched to the training/validation data, would give us a better idea of generalizability and robustness of the classifiers.

The second goal is to understand how overfitting can occur in the context of machine learning applied to neuroimaging-based diagnostic classification, whether in feature selection or performance estimation. We demonstrate how smaller datasets might give unreliable estimates of

classifier performance which could lead to improper model selection further leading to poor generalization across the larger population. The final goal of this study is to understand how specific functional connectivity patterns encode disease states and might possess predictive ability (as opposed to conventionally reported statistical separation) to distinguish between health and disease in novel individual subjects. We set out to identify these connectivity patterns which were not only statistically separated, but also were important for classification irrespective of age mismatch, acquisition site mismatch or the type of classifier used. These connectivity patterns must therefore be relatively robust to the age and acquisition site variations and their predictive ability must not be limited to a single classifier or a particular group of classifiers. In order to accomplish this, we propose feature ranking from multiple classifiers and data splits to construct a single score for the predictive ability of the connectivity features which can potentially be useful in clinical settings.

To achieve our goals, we applied 18 machine learning classifiers based on different principles including probabilistic/Bayesian classifiers, tree-based methods, kernel based methods, a few architectures of neural networks and nearest neighbor classifiers to RSFC metrics derived from ABIDE, ADNI, ADHD-200 and PCS/PTSD datasets described above. 7 of the 18 classifiers were implemented in a feature reduction framework called Recursive Cluster Elimination (RCE). We also built a consensus classifier which leverages the classifying power of all these classifiers to give reliable and robust predictions on the hold-out test dataset.

3.2 Materials and Methods

3.2.1 Data

Simulated Data

We first used simulated data to validate the classifiers using a known ground truth. We simulated 1500 normally distributed features for three classes, each feature with means of 0.2, 0.5 and 0.8 for each of the three classes. The standard deviation (SD) for each feature was incremented from 0.1 to 0.8 in steps of 0.1 to test the classifiers' robustness to noise. Since many classifiers model the data as multivariate Gaussians, we expect to get good performance from most classifiers at lower standard deviations.

Autism Spectrum Disorder (ASD)

ASD in a heterogeneous neurodevelopmental disorder in children characterized by impaired social communication, repeated behaviors and restricted interests. With a relatively high prevalence of 1 in 68 children, it is one of the most common developmental disorders in children [25]. According to DSM-V, ASD encompasses several disorders previously considered distinct including Autism and Asperger's Syndrome [26]. Asperger's Syndrome is considered to be a milder form of ASD, with patients in the higher functioning end of the spectrum. Autism is associated with large scale network disruptions of brain networks [27, 28, 29], thus making it an excellent candidate for disease classification using RSFC.

Resting state fMRI data from 988 individuals from the Autism Brain Imaging Data Exchange (ABIDE) database [29] was used for this study. The imaging data were acquired from 15 different acquisition sites including California Institute of Technology (CALTECH), Carnegie Mellon University (CMU), NYU Langone Medical Center (NYU), Kennedy Krieger Institute (KKI), University of Ludwig Maximilians University Munich (MAX-MUN), Pittsburgh School of Medicine (PITT), San Diego State University (SDSU), Olin Institute of Living at Hartford Hospital (OLIN), University of California, Los Angeles (UCLA), University of Leuven (LEUVEN), Trinity Centre for Health Sciences (TRINITY), University of Utah School of

Medicine (USM), Yale Child Study Center (YALE), University of Michigan (UM) and Social Brain Lab (SBL). The data consists of 556 healthy controls, 339 subjects diagnosed with Autism and 93 with Asperger's Syndrome. The distribution of the data used in this study with the acquisition site can be found in Table 3.1. Each subject's information was fully anonymized and was approved by the local Institutional Review Boards of the respective data acquisition sites. More details about the data including scanning parameters can be obtained from http://fcon_1000.projects.nitrc.org/indi/abide/index.html.

Imaging Site	Controls	Asperger's	Autism	Total
CALTECH	19	0	13	33
CMU	13	0	14	27
KKI	33	11	11	55
LEUVEN	35	0	29	64
MAX-MUN	33	22	2	57
NYU	105	21	53	179
OLIN	36	0	0	36
PITT	27	0	30	57
SBL	15	7	2	24
SDSU	22	7	3	32
TRINITY	25	7	10	42
UCLA	45	0	54	99
UM	77	10	55	142
USM	43	0	57	100
YALE	28	8	6	42
Total Subjects	556	93	339	988

Table 3.1. The site distribution for the ABIDE data set used in our study. We used a total of 988 subjects with 556 controls, 93 subjects with Asperger's syndrome and 339 with Autism.

Attention Deficit Hyperactivity Disorder (ADHD)

ADHD is one of the most common neurodevelopmental disorders in children with a childhood prevalence ratio as high as 11%, with significant increases in diagnoses every year [30]. ADHD diagnoses can be categorized into three subtypes based on the symptoms exhibited, including ADHD-I (Inattention) for persistent inattention, ADHD-H (Hyperactivity) for hyperactivity-impulsivity and ADHD–C (Combined) for a combination of both symptoms. There has been a massive increase in research efforts for automated detection of ADHD due to the ADHD-200 competition in 2011 [31].

930 subjects were selected from the ADHD-200 dataset, which was used for the ADHD-200 challenge [31]. The sample consists of 573 healthy controls, 208 subjects with ADHD-C (Combined), 13 subjects with subtype ADHD-H (Hyperactivity), and 136 subjects with ADHD-I (Inattentive). Imaging data for a few subjects were not included, as they did not pass the quality control (QC) thresholds. The subjects were scanned at seven different acquisition sites as shown in the Table 3.2. The acquisition parameters and other information about the scans be obtained from http://fcon_1000.projects.nitrc.org/indi/adhd200/.

Imaging Site	Controls	ADHD-C	ADHD-H	ADHD-I	TOTAL
Peking University	143	38	1	63	245
Kennedy Krieger Institute	69	19	1	5	94
NeuroIMAGE Sample	37	29	6	1	73
New York University Child Study Center	110	95	2	50	257
Oregon Health & Science University	70	27	3	13	113
University of Pittsburgh	94	0	0	4	98
Washington University	50	0	0	0	50
Total Subjects	573	208	13	136	930

Table 3.2. The site distribution for the ADHD-200 data across the seven imaging sites used in our study. We did not include the data from Brown University in our study since their diagnostic labels were not released.

Post-Traumatic Stress Disorder (PTSD) & Post-Concussion Syndrome (PCS)

PTSD is a debilitating condition which develops in individuals exposed to a traumatic or a life threatening situation. The estimated lifetime prevalence of PTSD among adult Americans is 6.8% [32]. Post-Concussion syndrome (PCS) consists of a set of symptoms that occur after a concussion, due to an injury to the head. PTSD can also be triggered by a traumatic brain injury, which is especially common in combat veterans. Such subjects display symptoms of both PCS and PTSD. Head injuries and traumatic experiences in the battlefield could be the main reasons for an unusually high prevalence rate of PTSD in combat veterans with a current prevalence of 12.1% in Gulf War Veteran population [33] and 13.8% in military veterans deployed in Afghanistan and Iraq during Operation Enduring Freedom and Operation Iraqi Freedom [34], respectively. Unfortunately, despite the serious nature of the problem, the current methods for diagnosis of the disease rely on subjective assessments of symptoms and psychological evaluations. An objective assessment of these disorders using image based biomarkers would facilitate reliable detection and diagnosis of PTSD and PCS.

While the three other datasets used in this study are publicly available, PTSD/PCS dataset was acquired in-house. 87 active duty male US Army soldiers were recruited to participate in this study from Fort Benning, GA and Fort Rucker, AL, USA. In the recruited subjects. 28 were combat controls, 17 were diagnosed with only PTSD, while 42 were diagnosed with both PCS and PTSD. All subject groups were matched for age, race, education and deployment history. The subjects were diagnosed as having PTSD if they had no history of mild Traumatic Brain Injury (mTBI), or symptoms of PCS in the past five years, with scores>38 on Checlist-5 (PCL5),

and <26 on Neurobehavioral Symptom Inventory (NSI). Subjects with medically documented mTBI, post-concussive symptoms, and scores ≥38 on PCL5 and ≥26 on NSI were grouped as PCS+PTSD. The procedure and the protocols in this study were approved by the Auburn University Institutional Review Board (IRB) and the Headquarters U.S. Army Medical Research and Material Command, IRB (HQ USAMRMC IRB).

The participants were scanned in a Siemens 3T MAGNETOM Verio Scanner (Siemens Erlangen, Germany) with a 32 channel head coil at Auburn University. The participants were instructed to keep their eyes open and fixated on a small white cross on a screen with a dark background. A T2* weighted multiband echo-planar imaging (EPI) sequence was used to acquire two runs of resting state data in each subject with the following sequence parameters: TR=600ms, TE=30ms, FA=55˚, multiband factor=2, Voxel size= 3×3×5 mm^3 and 1000 time points. Brain coverage was limited to the cerebral cortex, subcortical structures, midbrain and pons with the cerebellum excluded.

Mild Cognitive Impairment (MCI) & Alzheimer's disease (AD)

Mild Cognitive Impairment (MCI) can be defined as greater than the normal cognitive decline for a given age, but it does not significantly affect the activities of daily life [35]. It has a prevalence ranging from 3% to 19 % in adults older than 65 years. Alzheimer's disease (AD), on the other hand, does significantly affect daily activities of the person. It is the most common neurodegenerative disorder in adults aged 65 and older. It is characterized by cognitive decline, intellectual deficits, memory impairment and difficulty in social interactions. A large percentage of MCI patients slowly progress to Alzheimer's disease, yet the boundaries separating healthy aging from early/late MCI and AD is not very precise leading to diagnostic uncertainty in the

disease state [36]. Therefore, classifying MCI from AD and healthy older controls is extremely crucial and is particularly challenging.

Resting state functional brain imaging data of 132 subjects were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The sample consists of subjects in various stages of cognitive impairment and dementia, including 34 subjects with early mild cognitive impairment (EMCI), 34 with late mild cognitive impairment (LMCI), 29 with Alzheimer's disease (AD) and finally 35 matched healthy controls. More information about the data used for this study along with the image acquisition parameters can be obtained from http://adni.loni.usc.edu/.

3.2.2 Processing of the Rs-fMRI data

Standard preprocessing pipeline for Rs-fMRI data was implemented using Data Processing Assistant for Resting-State fMRI Toolbox (DPARSF) [37]. The preprocessing pipeline consisted of removal of first five volumes, slice timing correction, volume realignment to account for head motion, co-registration of the T1-weighed anatomical image to the mean functional image, nuisance variable regression which included linear detrending, mean global signal, white matter and cerebrospinal fluid signals and 6 motion parameters. After nuisance variable regression, the data were normalized to the MNI template. The Blood Oxygen Level Dependent (BOLD) time series from every voxel in the brain was deconvolved by estimating the voxel-specific Hemodynamic Response Function (HRF) using a blind deconvolution procedure to obtain the latent neural signals [38]. The data were then temporally filtered with a band pass filter of bandwidth 0.01-0.1 Hz. Mean time series were extracted from 200 functionally homogeneous brain regions as defined by the CC200 template [39]. After extracting the timeseries, functional connectivity (FC) between the 200 regions was calculated as the Pearson's correlation coefficient

between all region pairs giving us a total of 19,900 FC values. These were then used as features for the classification procedure. For ADHD and PTSD datasets, we did not have whole brain coverage. Therefore, we obtained time series from just 190 regions and 125 regions, respectively. The number of FC paths were accordingly lower for these datasets.

3.2.3 Data splits for training/validation and hold-out test data

In order to test the generalizability of the classifier models, we split all imaging data into two components. Approximately, 80% of the data was used for training/validation, and the remaining 20% was used as a hold-out test data set. The training/validation datasets were split even further for cross-validation in order to estimate the classifier models as we explain later. However, the hold-out test datasets were not used in cross-validation; instead, they were used only once with the classifier models obtained from cross-validation in order to obtain truly unbiased test accuracy on completely unseen data. In a few splits, the training/validation and test data came from homogeneous populations, i.e. they were matched for age and acquisition site. In some other splits, the training/validation and hold-out test data were not matched, i.e. they had different age range or acquisition site. With matched data, it is important to note that training/validation and the hold-out test data were matched in age, race, education and gender. In the unmatched splits, age/acquisition site was unmatched, while race, gender, education and acquisition site/age, respectively, were matched. All these splits on the four datasets are summarized in Figure 3.1 and will be elaborated below.

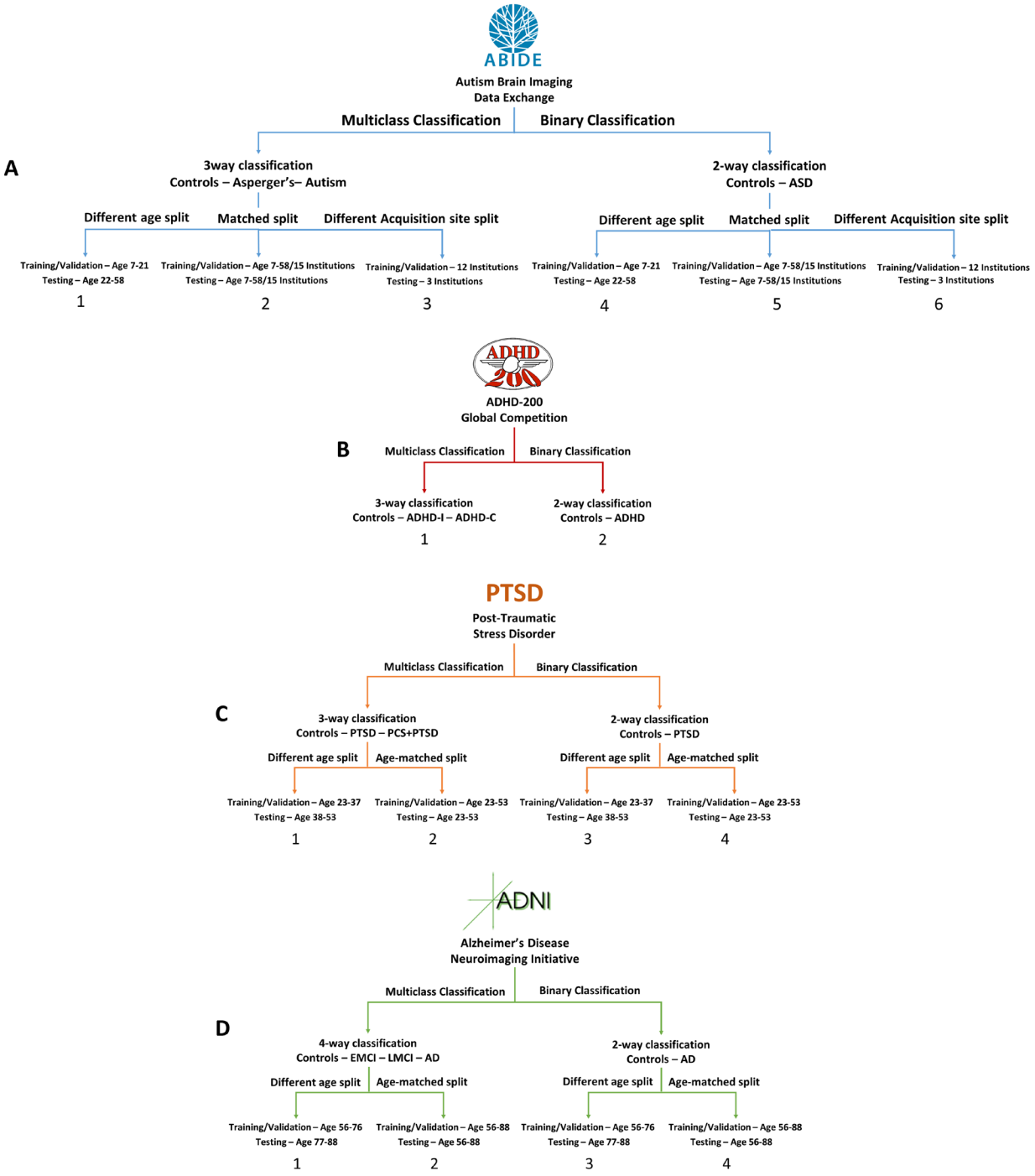


Figure 3.1. The age and imaging site split for the training/validation and the test data both for binary and multiclass classification scenarios. A) For the ABIDE dataset, we had age- and site-matched splits as well as unmatched splits for both 2-way and 4-way classifications. In the first split, subjects from an age range of 23-37 years were used in training/validation data and the subjects from the age range 38-53 years formed the hold-out test data. Second, we performed an imaging site split wherein the data from the 12 imaging sites (PITT, OLIN, SDSU, TRINITY, UM, USM, CMU, LEUVEN, NYU, MAXMUN, CALTECH, SBL) were used for the

training/validation data while the rest of the 3 imaging sites (Yale, KKI, UCLA) were used as a hold-out test dataset. In the third split, training/validation and hold-out test data were matched for age and acquisition site. B) For ADHD we directly used the training/validation and hold-out test data provided by the ADHD-200 Consortium for binary and multiclass classification. C) For binary and 3-way classification of the PTSD dataset, we followed an age split in which the training/validation data contained subjects from an age range of 23-37 years while the hold-out test data contained subjects from the age range 38-53 years. This was then compared with a matched training/validation data and hold-out test data with subjects in the age range of 23-53 years. D) For both 2-way and 4-way classification of ADNI dataset, we split the entire data by age wherein the training/validation data contained subjects from an age range of 56-76 years while the hold-out test data contained subjects from the age range 77-88 years. This scenario was compared with a matched training/validation data and hold-out test data with subjects in the age range of 56-88 years.

A

Diagnosis	Training/Validation Data	Hold-out Test Data
	Age 7-21	Age 22-58
Controls	429	127
Asperger's	63	30
Autism	279	60
Total Subjects	771	217

Diagnosis	Training/Validation Data	Hold-out Test Data
	12 imaging sites	3 imaging sites
Controls	450	106
Asperger's	74	19
Autism	268	71
Total Subjects	792	196

Diagnosis	Training/Validation Data	Hold-out Test Data
	Matched	Matched
Controls	445	111
Asperger's	74	19
Autism	271	68
Total Subjects	790	198

B

Diagnosis	Training/Validation Data	Hold-out Test Data
	7 imaging sites	6 imaging sites
Controls	479	94
ADHD-C	159	49
ADHD-H	11	2
ADHD-I	110	26
Total Subjects	759	171

C

Diagnosis	Training/Validation Data	Hold-out Test Data
	Age 23-37	Age 38-53
Controls	46	10
PCS+PTSD	66	18
PTSD	26	8
Total Subjects	138	36

Diagnosis	Training/Validation Data	Hold-out Test Data
	Age Matched (23-53)	
Controls	45	11
PCS+PTSD	67	17
PTSD	27	7
Total Subjects	139	35

D

Diagnosis	Training/Validation Data	Hold-out Test Data
	Age 56-76	Age 77-88
Controls	23	12
Early Mild Cognitive Impairment (EMCI)	28	6
Late Mild Cognitive Impairment (LMCI)	27	7
Alzheimer's Disease (AD)	22	7
Total Subjects	100	32

Diagnosis	Training/Validation Data	Hold-out Test Data
	Age Matched (56-88)	
Controls	28	7
Early Mild Cognitive Impairment (EMCI)	27	7
Late Mild Cognitive Impairment (LMCI)	27	7
Alzheimer's Disease (AD)	23	6
Total Subjects	105	27

Table 3.3. The data distributions for training/validation and hold-out test data for the age and imaging site splits for (A) ABIDE dataset (B) ADHD-200 dataset (C) PTSD dataset (D) ADNI dataset.

ABIDE: We split the ABIDE data into two heterogeneous sets for training/validation and testing, with differences in age group and imagining site: (i) The first heterogeneous split had the training/validation data from age range 23 -32 years while the holdout test data had both ASD and healthy controls in the age range 33-47 years. (ii) For the second split, the training validation data came from 12 imagining sites which participated in the study. The hold-out test data was drawn from the remaining three institutions. (iii) We also had a matched split with data for training/validation and testing drawn from the same age range and institutions. Since the ABIDE data has healthy controls and two subgroups of ASD in Autism and Asperger's syndrome, we performed both binary and multiclass classification with each of the three splits, giving us a total of six splits. The distribution of the subjects in each split is shown in Table 3.3A.

ADHD-200: The ADHD-200 global competition was structured in a way that training/validation data with diagnostic labels were first provided to the public and many groups around the world submitted their predictions on unlabeled hold-out test data dataset. The organizers of the competetion assessed the performance of the classification tools on the hold-out test data set based on the predicted diagnositc labels submitted by the groups. Following the completion of the competition, the labels for hold out test dataset was also publicly released. Therefore, we used the training/validation and hold-out test datasets originally provided by the organizers of the competition and no further splits were performed on the data by age or by acquisition site, as was done for other datasets used in this study. This also helps us stay true to the spirit of the ADHD-200 Global Competition. We performed binary classification between Controls and ADHD (data from all 3 ADHD subgroups were combined) as well as a three-way classification between controls, ADHD-C, and ADHD-I. ADHD-H was left out in multiclass classification because

only 11 subjects with ADHD-H were present in the data. The data distributions for the training/validation and hold-out test data is shown in Table 3.3B.

PTSD: Since the imaging data for PTSD was collected solely from our research site, we could not test the effects of the performance accuracy due to site variability. We performed binary (Controls vs. PTSD) as well as 3-way classification with Controls vs. PTSD vs. PCS+PTSD. Subjects in the age range from 23-35 years were used in training/validate data and ages 35-47 years were used in the hold-out test data for the heterogeneous split. Age matched training/validation and test data were also used. These two splits were used for each of the two classification scenarios (binary and 3-way), giving us a total of four splits. It is noteworthy that we had two runs from each of the 87 subjects in this dataset and we considered each run as a separate subject. Therefore, effectively, we had 174 subjects in this dataset. The data distributions of splits are shown in Table 3.3C.

ADNI: ADNI data contains subjects at various stages of cognitive impairment. Therefore, we tested a 4-way classification between healthy adults, EMCI, LMCI, and AD. We also performed binary classification using just healthy adults and AD subjects at the extreme ends of the spectrum. We tested the effect of age heterogeneity on the classification performance with subjects from the age range 34-67 years chosen for training/validation data and 45-78 years selected for hold-out test data. We also had a homogeneous split with training /validation and hold-out test data chosen randomly from the entire dataset with the age range of 56-88 years. The data distributions of each of the classes in these splits are shown in Table 3.3D.

We made no effort to balance the classes with unbalanced sample sizes in the four data sets because: (i) we wanted to identify classifiers which are robust to differences in class occurrences in the training data and, (ii) the number of healthy subjects are usually far greater than the

number of subjects with disorders in neuroimaging databases which are assembled retrospectively. While concerted efforts to acquire large and homogenized balanced datasets are currently underway [40], it will be many years before they become publicly available.

3.2.4 Classification procedure

The number of features obtained by resting state functional connectivity metrics are usually orders of magnitude larger than the number of subjects/samples available. Due to the "curse of dimensionality", when using high dimensional data, overfitting is a huge concern because the underlying distribution may be under-sampled [41, 7, 42]. Having an excess number of features compared to the number of data samples might lead to overfitting and give us poor generalization on the test data [43, 41]. The most useful strategies to deal with this issue include collecting more data, adding domain knowledge about the problem to the model or reduce the number of features, with ideally preserving class-discriminative information. Therefore, feature selection is a necessary step either before classification or as a part of the classification procedure, given the sample size of current neuroimaging databases. Most existing feature selection methods can be grouped into filter and wrapper methods. Filter methods are independent of the classification strategy. A simple univariate score such as a T-score can be used to rank the features and the top ranked features can be utilized for classification [44]. Although computationally quick, this univariate approach does not take into consideration the relationships between different features and the classifier performance when retaining features. A wrapper method selects subsets of features which give good classification performance and contain class-discriminative information. Hence the classifier is embedded with feature selection in the wrapper method framework. A combination of wrapper and filter methods have been shown to perform well with minimum resources [45, 46, 47]. Therefore, we have adopted this

strategy in the current study. As our filter method, we used a two-sample T-test/ANOVA and selected the features whose means were significantly different between the groups ($p<0.05$, FDR corrected), after controlling for confounding factors such race, gender and education for the age unmatched splits. However for the age-matched splits, age was also controlled for along with race, gender and education. When selecting significant features in the age-unmatched splits, age was not included because including it in the model would have removed age-related variance from the data.

The Rs-fMRI data was divided into training/validation data and hold-out testing data with approximately 80% of the data used for training and cross-validation, and the remaining 20% of the data was used as a separate hold-out test dataset as was mentioned in the previous section on the data splits. In many cases, the training/validation data and hold-out test data differed in a few factors as mentioned previously such as age and acquisition site. As mentioned above, an initial "feature-filtering" was performed wherein only the connectivity paths that were significantly different between the groups ($p<0.05$, FDR corrected) in the training/validation data were retained (after controlling for head motion, age, race and education) thereby reducing the number of features from 19,100 to around 1000. No statistical tests were performed on the independent test data to avoid introducing any bias. Therefore, the features with $p<0.05$ (FDR corrected) in the training/validation dataset were also removed from the hold-out test dataset. Please note that the hold-out test dataset was not used in feature or model selection and thus can be expected to give an unbiased estimate of the generalization accuracy. This is contrary to the cross-validation accuracy estimate because using t-test filtering in reducing features on the entire training/validation data will lead to optimistic accuracy estimates given that the training data and the validation data are not completely separated. Even if t-test was not performed on the

validation data during cross-validation, cross-validation accuracy, by definition, is the average accuracy obtained from different splits. Therefore, it does not provide a conservative estimate of the classifier's performance. To further reduce the number of features while retaining discriminative information, some of the classifiers were embedded in the recursive cluster elimination (RCE) framework [45] for feature section (Figure 3.2). As we describe later, some of the classifiers had some form of feature selection embedded within them and hence such methods were implemented without the RCE framework (Figure 3.3).

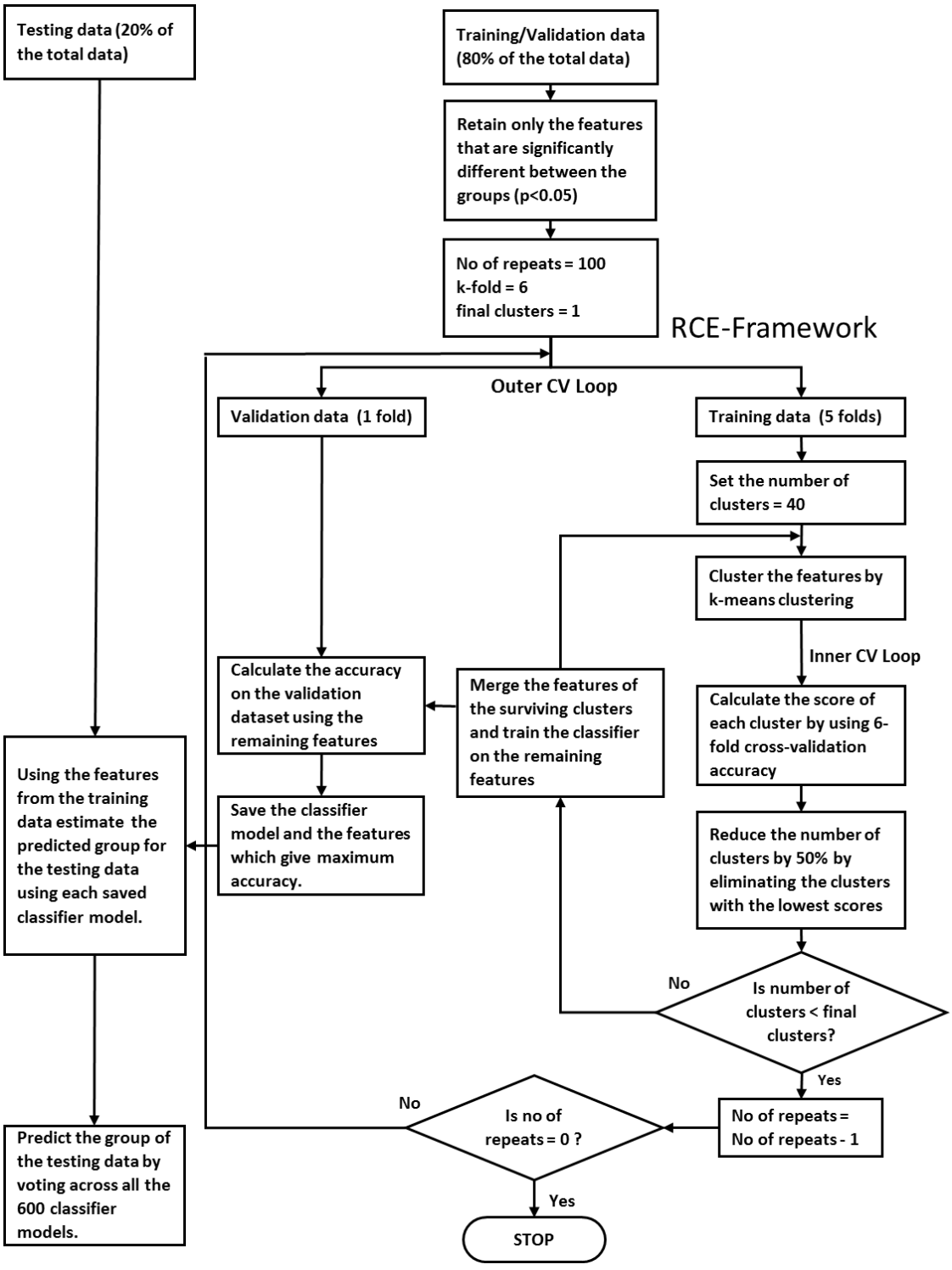


Figure 3.2. A schematic of the classifiers implemented within the RCE framework. RCE selects features in the inner cross-validation loop while the outer cross-validation loop estimates the performance of the classifier.

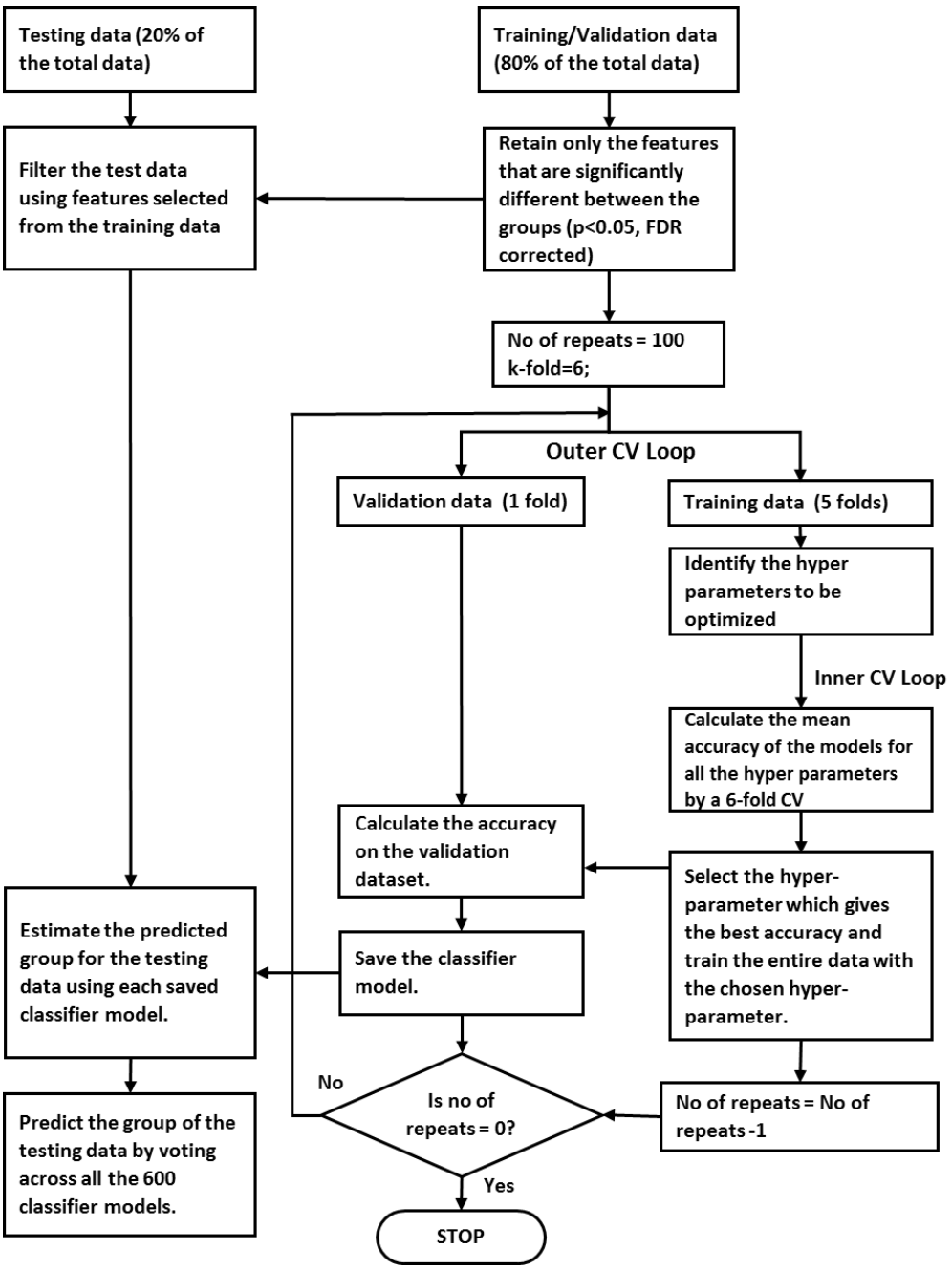


Figure 3.3. A schematic of the classifiers implemented outside the RCE framework. We used a two-level cross-validation for parameter optimization and performance estimation for the training/validation data.

Recursive Cluster Elimination (RCE) framework

Recursive Cluster Elimination (RCE) is a heuristic method for identifying a subset of features that have class-discriminative information. Recursive Cluster Elimination is a wrapper method that combines K-means feature clustering with a machine learning classifier to score the discriminative ability of clusters of features and helps retain only features with good discriminative power and remove the ones without any discriminative power [45, 48]. RCE exploits the fact that features (functional connectivities in our case) are often correlated with each other and hence their discriminative abilities can be ascertained together by clustering the feature space. This provides an order of magnitude increase in speed compared to eliminating each feature individually [48]. We implemented classifiers in a nested cross-validation procedure, with the inner cross-validation loop performing feature selection via RCE and the outer cross-validation loop was used for performance estimation (Figure 3.2). We first started with all features after t-test filtering and clustered these features using the K-means algorithm. The correlation coefficient was used as the distance metric while clustering. Each cluster of features was then used to train a machine learning classifier and a score was assigned to the cluster based on the performance of the cluster on the validation data. The clusters were ranked according to their classification performance, and the clusters with lowest scores were eliminated, and the features in the remaining clusters were merged. This process was iteratively repeated until any further removal of clusters decreased the classification accuracy. This ensured that the best set of feature clusters were identified. This optimal set of feature clusters for each k-fold and partitioning of data of the cross-validation loop, and the final decision surface (or hyper-plane in higher dimension) which gave the best cross-validation performance were saved and used for calculating the accuracy from hold-out test data. For each repetition, a different model, with distinct hyperparameters and features were selected. These models were then used to assess

the cross-validation accuracy in the outer k-fold. This ensures that separation is maintained between feature selection by RCE and performance estimation.

Using FC features from training/validation data, classification accuracy was calculated using repeated 6-fold cross-validation. The classifier models obtained from the differences in the partitioning of the training data (repeats × folds) were saved. Test accuracy was calculated on the independent hold-out test data using the saved classifier models by a voting procedure. Each classifier would vote towards a decision on test subjects (accuracy was the percentage of correct votes). This is the voting test accuracy reported. The w/o voting accuracy refers to the mean accuracy and standard deviation for the test data obtained by each of the individual 600 classifier models obtained in each iteration during the cross-validation. The classification procedure was identical for simulated and experimental imaging data.

Classifier models

We used a number of classifier models to address the issues in performance estimation and generalizability so that our results are not specific to any particular classifier or type of classifiers. The classifiers we implemented can be broadly divided into the following categories (i) Probabilistic/Bayesian methods: Gaussian Naïve Bayes (GNB), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Sparse Logistic Regression (SLR), Ridge Logistic Regression (RLR), (ii) Kernel methods: Linear and Radial Basis Function (RBF) kernel Support Vector Machines (SVM), Relevance Vector Machines (RVM), (iii) Artificial Neural Networks: MLP-Net (Multilayer Perceptron Neural Net), FC-Net (Fully Connected Neural Net), ELM (Extreme Learning Machines), LVQNET (Linear Vector Quantization Net), (iv) Instance-based learning: K-Nearest Neighbors (KNN), (v) Decision Tree based Ensemble Methods: Bagged trees, Boosted Trees, Boosted Stumps, Random Forest, Rotation Forest. We now provide

a brief introduction to the machine learning classifiers used in this paper. This is by no means a comprehensive description of the classifiers and the readers are referred to sources cited in these descriptions for more detailed information. For classifiers with hyper-parameters in them that needed to be optimized, we performed a grid search to estimate an optimum value. Therefore, it may be possible to further optimize these parameters using more advanced methods. However, a concern with fine-tuning the parameters and testing a large number of models in cases with limited data is that it might lead to overfitting [49]. All the classifiers were implemented in MATLAB environment (Natick, MA). Also note that in this paper, the terms parameters and weights are used interchangeably. Further details about the algorithms implemented in this study can be found in Appendix A.

We implemented Linear- & RBF-kernel SVM, GNB, LDA, QDA, KNN and ELM in the RCE framework. Many other classifiers we used, such as SLR, RLR, RVM, FC-NN and MLP-NN have built in regularization to control model complexity. Ensemble methods such as Bagged Trees, Random Forests, Boosted Stumps, Boosted Trees and Rotation Forests are not as sensitive to classification problems with a large number of features. Therefore we did not implement classifiers with built-in regularization as well as ensemble methods in the RCE-framework. KNN was implemented both within and outside the RCE-framework.

3.2.5 Classification performance metrics

Since many of the datasets which are used in this study are unbalanced in class labels (i.e. each class contains an unequal number of instances), it is important to investigate individual class accuracies. In such cases where one class has more observations in the dataset than the other class, the classifier reports a high accuracy even if the classifier just assigns the majority class label to all instances in the test dataset [7]. In these cases, the overall/unbalanced accuracy is not

indicative of the actual performance of the classifier. Therefore, in addition to presenting the overall/unbalanced accuracy, we also report individual class accuracies as well as the balanced accuracy. The individual class accuracies report the ratio of correctly classified instances of a particular class to the total number of instances of the class in the data. The mean of individual class accuracies obtained from both the training/validation data and the hold-out test dataset represents the balanced cross-validation accuracy and the balanced hold-out test accuracy, respectively.

For all the classification problems we considered, we report: (i) The unbalanced cross-validation (CV) accuracy and its standard deviation (in parenthesis), (ii) CV class accuracies of the individual groups (iii) The balanced CV accuracy obtained by the mean of individual CV class accuracies, (iv) Hold-out test accuracy by voting (unbalanced hold-out test accuracy), (v) Mean hold-out test accuracy, which is obtained by using mean of the test accuracies calculated from individual classifier models and its standard deviation (in parenthesis), (vi) Individual class accuracies of the groups obtainted from the hold-out test data, and (vii) The balanced hold-out test accuracy as an average of individual class hold-out test accuracies. A schematic illustrating the derivation of the classification performance metrics from the confusion matrix is shown in Figure 3.4.

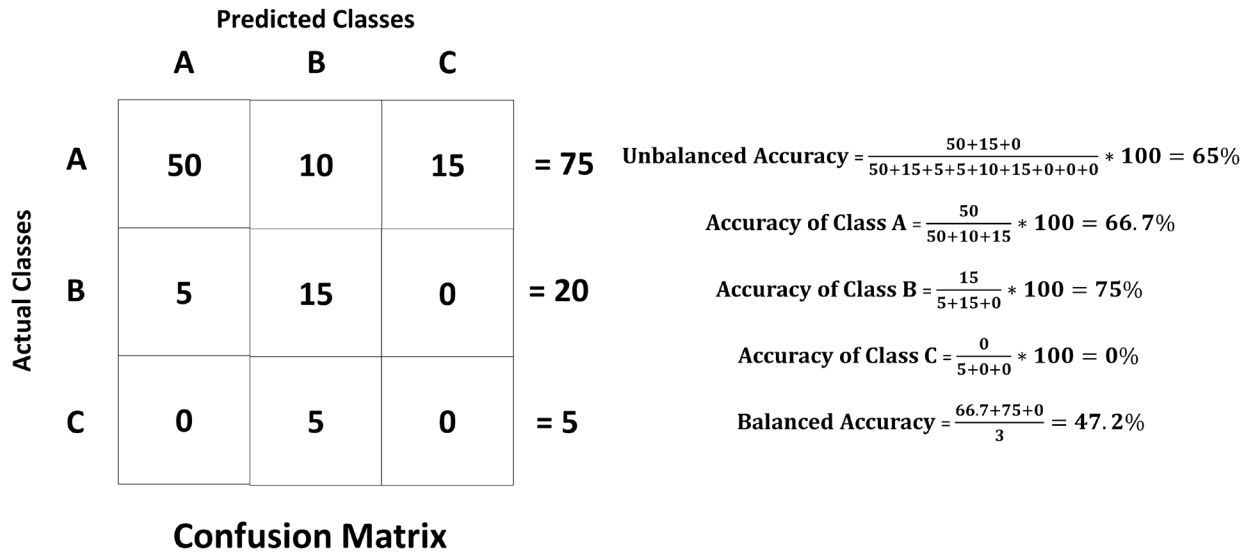


Figure 3.4. A schematic showing the derivation of the classification performance metrics from the confusion matrix. We report balanced accuracy, unbalanced accuracy and individual class accuracies.

The evaluation of the classification performance and the diagnostic utility of the classifier must be made taking into consideration all the above performance metrics as well as the classification scenario. It should be noted that in datasets in which some classes have very few instances, classifiers can find it extremely difficult to learn those patterns. The balanced accuracy might also suffer because some diseases such as Asperger's have a tiny number of samples compared to other groups in their dataset, thereby making any reliable classification extremely difficult and giving a low balanced accuracy. The holdout accuracy is a pessimistic estimator of the generalization accuracy because only a portion of the data was given to the classifier for training and the holdout test dataset in our study was chosen to be from a slightly different population than training data. As we demonstrate in this study, the high accuracies commonly reported for Leave-one-out cross-validation (LOOCV) and k-fold CV in neuroimaging studies [50, 46] are misleading, especially when there is significant heterogeneity in the population. Therefore, one

has to evaluate the performance of a classifier using multiple metrics presented above, in order to assess its performance under both optimistic and pessimistic scenarios.

3.2.6 Calculation of Feature Importance

Recursive Cluster Elimination (RCE) procedure provides us with a feature ranking that indicates the importance of a particular feature in discriminating between the classes. For every step of the RCE loop, we kept the count of the features retained and used the count to assign higher feature importance scores (FIS) to features that were retained by the classification procedure while assigning lower scores to features eliminated early in the feature elimination process. We repeated this for every partitioning of data in the outer k-fold, thereby obtaining the FIS for every classifier implemented in the RCE-framework. We combined the feature importance score of all the classifiers implemented in the RCE-Framework, weighted by their balanced cross-validation accuracy, to obtain a combined score of feature importance (CFIS) for the classification problem. Multiple splits of the entire data into training/validation and hold-out test data gave a slightly different ranking to most classifiers across different splits. We plotted the CFIS of the features commonly found in all the data splits as a scatter plot. We repeated this procedure separately for multiclass and binary classification problems for every dataset. To obtain features which are generalizable across age groups and data acquisition sites, we identified a subset of features in each split, which have high feature importance scores (top 100), implying that they play a significant role in class discriminative ability as well as have significantly different means between the groups ($p<0.05$, corrected for multiple comparisons using permutation test [51] by modeling the null distribution of maximum t-scores of features by permuting the class labels of the data). The features or connectivity paths thus identified were then visualized in BrainNet Viewer [52]. Similarly, we also ranked brain regions based on the sum of the CFIS of

connectivity paths associated with them. A list of the top 20 brain regions was obtained for every neurological disorder considered in this study.

3.2.7 Consensus classifier

We have employed 18 different classifiers in this study. Many of them are based on entirely different principles, yet they all attempt to achieve the same result of determining the decision boundary which separates the groups. When multiple classifiers are used in neuroimaging, it is customary to report and emphasize on the one which gave highest classification accuracy [53, 54]. This might give an optimistic estimate of the accuracy and the result might not be repeatable even for data from the same population. Alternatively, we developed a simple consensus based approach wherein the performance of all 18 classifiers were combined to provide a consensus estimate.

For every classifier, during cross-validation, we resampled the data 600 times (6-fold x 100 repetitions), to get 600 different classifier models for each resampling. We used these 600 models for each classifier to predict the class of the observations in the validation data, giving us a total of 600 predictions for every observation in the hold-out test data. We then calculated individual class probabilities for the hold-out test data by estimating the relative frequency of the 600 target class predictions for the hold-out test data. In this way, the relative frequency of the target class was estimated for each test observation. Then the final class probabilities of the consensus classifiers were calculated by weighing the predicted class frequencies of each classifier with its balanced cross validation accuracy. The test observation was assigned to the class with highest probability. This way multiple classifiers can be combined to provide a consensus classifier which greatly improves the reliability and robustness of inferences made

from them. A schematic depicting the predictions of the consensus classifier on the hold-out test data is shown in Figure 3.5.

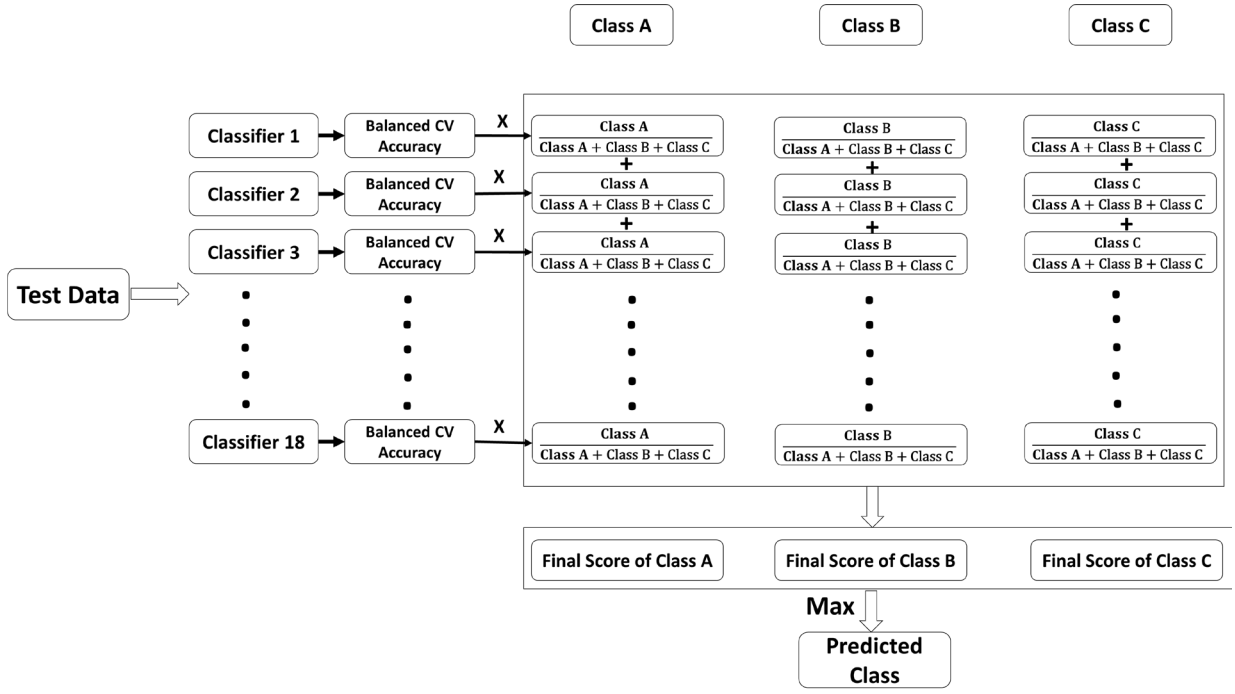


Figure 3.5. A schematic illustrating the consensus classifier used to combine predictions from multiple classifiers on the hold-out test dataset. The consensus classifier combines the balanced cross-validation accuracy of each classifier with the predictions of its 600 decision models obtained by cross-validation, to predict the class label for the hold-out test data.

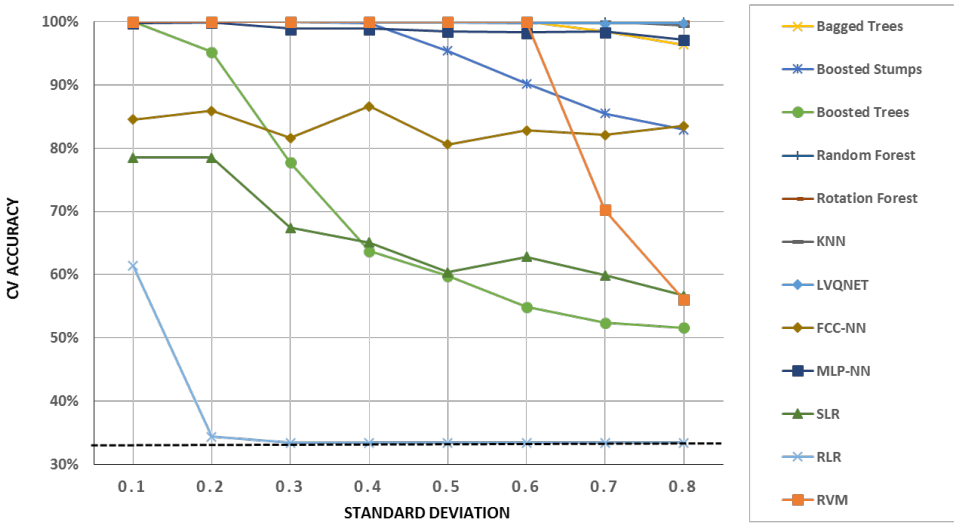
3.3 Results

3.3.1 Simulation results

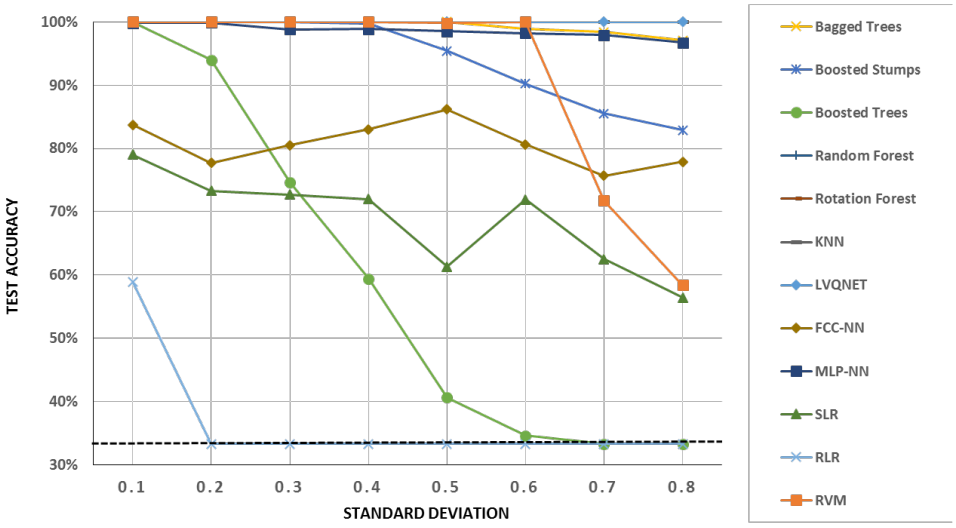
The simulation results indicate that most classifiers performed well at lower standard deviations and their performance became worse as the standard deviation increased (Figure 3.6). This is to be expected given that higher standard deviations of simulated features reduces the separation between the groups. The hold out test accuracy with voting was significantly better than the average test accuracy of the individual cases. Classifiers such as SLR and RLR did not perform

well probably because the features selected at the cross-validation stage were not representative of the actual feature importance. This is because the features were generated randomly and they have not had an actual discriminative value across the subject groups. The feature weights which are learned by these two classifiers are thus not generalizable. This issue is exacerbated in classifiers such as SLR and RLR which have in-built feature selection. But the simulation results do indicate that most classifiers were successful in separating the classes when each feature was represented by univariate Gaussian distributions. The results for classifiers implemented within RCE framework are not included in Figure 3.6 because all classifiers gave 100% accuracy at all the standard deviations tested.

A


CV ACCURACY
100%
90%
80%
70%
60%
50%
40%
30%
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
STANDARD DEVIATION
Bagged Trees
Boosted Stumps
Boosted Trees
Random Forest
Rotation Forest
KNN
LVQNET
FCC-NN
MLP-NN
SLR
RLR
RVM

B


TEST ACCURACY
100%
90%
80%
70%
60%
50%
40%
30%
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
STANDARD DEVIATION
Bagged Trees
Boosted Stumps
Boosted Trees
Random Forest
Rotation Forest
KNN
LVQNET
FCC-NN
MLP-NN
SLR
RLR
RVM

C

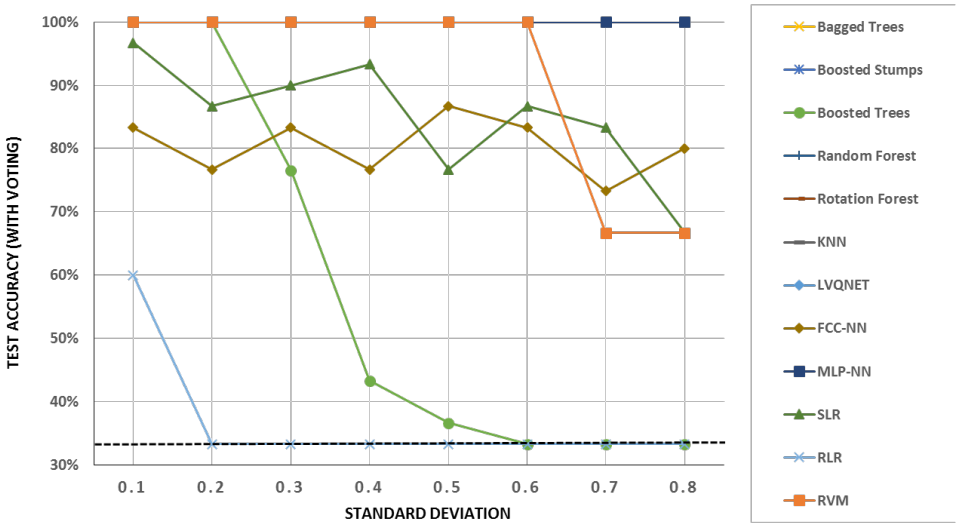

TEST ACCURACY (WITH VOTING)
100%
90%
80%
70%
60%
50%
40%
30%
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
STANDARD DEVIATION
Bagged Trees
Boosted Stumps
Boosted Trees
Random Forest
Rotation Forest
KNN
LVQNET
FCC-NN
MLP-NN
SLR
RLR
RVM

Figure 3.619. Performance estimates from various classifiers obtained with the simulated dataset for the (A) training/cross-validation (B) hold-out test data without voting (C) hold-out test data with voting. Most classifiers performed well with the simulated data for both the cross-validation accuracy as well as the accuracy of the hold-out test data. As expected voting on the test data by using multiple classifier models obtained by multiple partitionings of the data did boost the accuracies of the classifiers. Classifiers with inbuilt feature selection such as RLR and SLR performed terribly, probably because the features were generated randomly and were not indicative of the predictive power of the classifier. It is noteworthy that the results for classifiers implemented within RCE framework are not included here because all of them gave 100% accuracy at all the standard deviations tested.

3.3.2 ABIDE

Classification results: The classification results for the binary classification scenario between healthy controls and subjects with ASD for the age split, site split as well as for the matched data are shown in Figure 3.7, Figure 3.8, and Figure 3.9 respectively. The corresponding tables showing the detailed individual class accuracies are shown in Table 3.4, Table 3.5, and Table 3.6 respectively. The corresponding results for the multiclass classification scenario are presented in Figure 3.10, Figure 3.11 and Figure 3.12 with detailed accuracy performance presented in Table 3.7, Table 3.8, and Table 3.9, respectively. The results indicate that there is considerable difference in accuracy between the biased cross-validation accuracy and the hold-out test accuracy across all the classifiers, with the former being consistently greater than the latter. The performance was above the accuracy that could be obtained for a majority classifier (50% for balanced, 54% to 58.5% for unbalanced depending on the splits) for binary as well as multiclass class (33% for balanced, 54% to 58.5% for unbalanced depending on the splits) scenarios for all classifiers. The majority classifier is a primitive classifier which assigns the most frequently occurring class label to all the instances in the test data. Even in multiclass classification scenarios, no classifier was able to reliably classify Asperger's syndrome, which was the reason for lower balanced accuracy for all three multiclass scenarios.

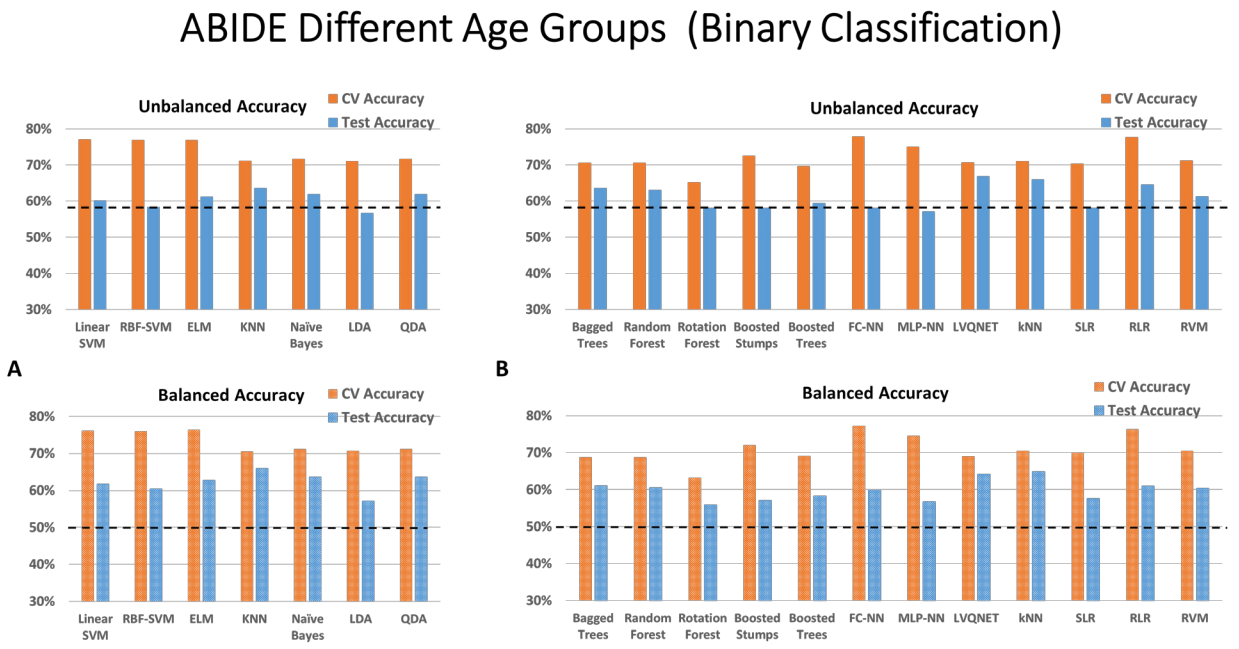


Figure 3.7. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ABIDE data when the training/validation data and the hold-out test data are from different age groups for the binary classification between healthy controls and subjects with ASD. The training/validation data is from an age range of 7-21 years while the data from the age range of 22-58 years was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 58.5% since healthy controls formed 58.5% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 66.8% obtained by LVQNET while the best balanced hold-out test accuracy was 64.9% obtained for KNN implemented outside the RCE framework.

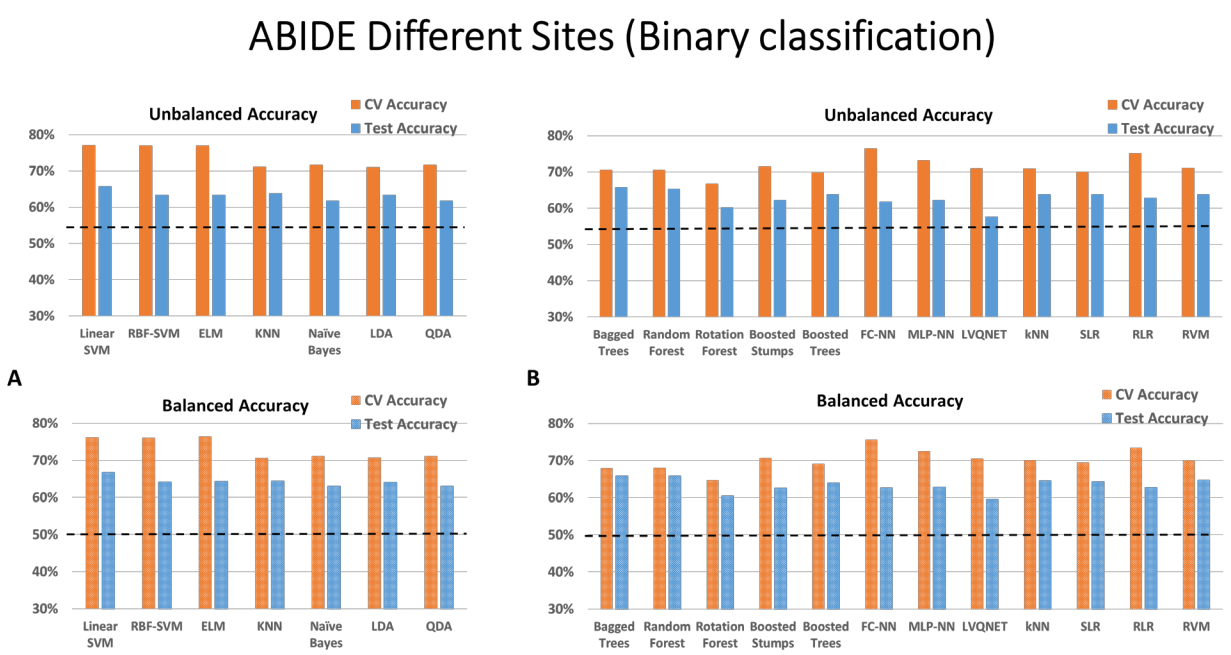


Figure 3.8. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ABIDE data when the training/validation data and the hold-out test data are from different acquisition sites for the binary classification between healthy controls and subjects with ASD. The training/validation data are from 12 institutions while the data for the remaining three institutions was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 54% since healthy controls formed 54% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 66% obtained for Bagged trees while the best balanced hold-out test accuracy was 66.8% obtained for Linear SVM implemented within the RCE framework.

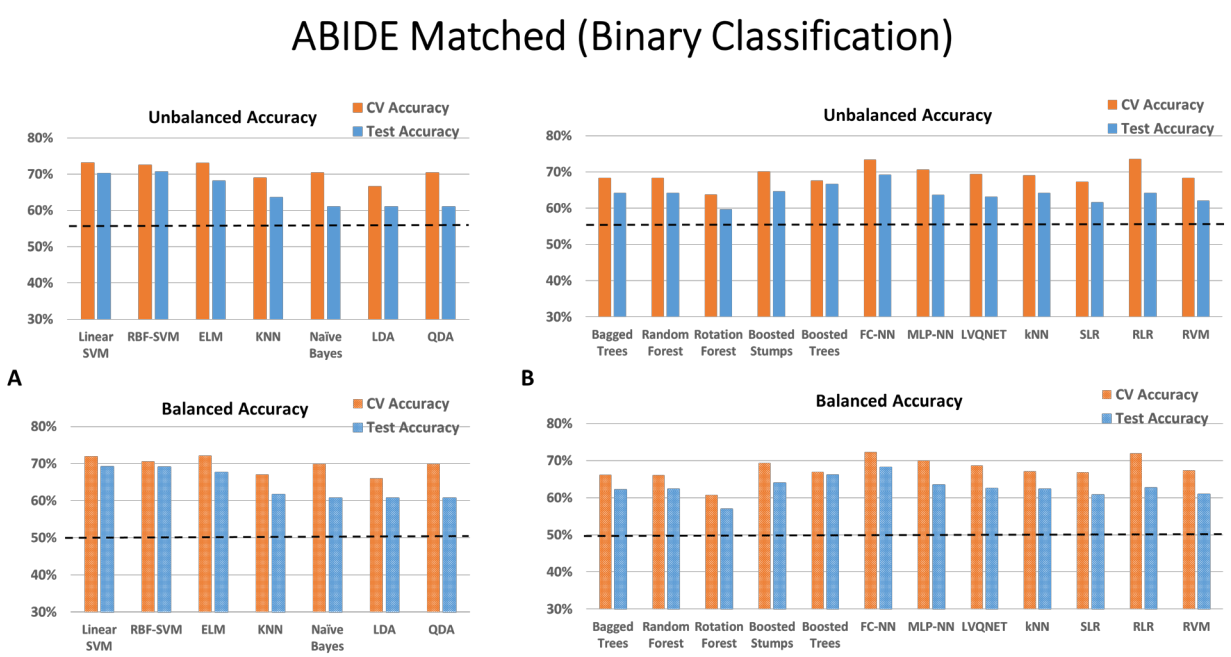


Figure 3.9. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ABIDE data when the training/validation data and the hold-out test data are matched in imaging sites as well as age group for the binary classification problem between healthy controls and subjects with ASD. The training/validation and the hold-out test data are from all 15 imaging sites and age range of 7-58 years. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 56% since healthy controls formed 56% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. Some of the discrepancies between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 70.7% obtained by RBF-SVM within the RCE framework while the best balanced hold-out test accuracy was 69.2% obtained for Linear SVM implemented within the RCE framework.

ABIDE Different Age Groups (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ASD	w/o Voting	Voting	Balanced	Controls	ASD
Linear SVM	77% (3.3%)	76.2%	83.0%	69.4%	60% (2.4%)	61.7%	60.1%	70.1%	50.0%
RBF-SVM	76.9% (3.5%)	76.1%	83.9%	68.2%	59.7% (1.9%)	60.4%	58.4%	70.1%	46.7%
ELM	76.9% (3.7%)	76.5%	80.7%	72.2%	61.3% (1.9%)	62.7%	61.2%	70.1%	52.2%
KNN	71.1% (3.8%)	70.6%	74.5%	66.7%	63.6% (3.1%)	65.9%	63.6%	77.2%	50.0%
Naïve Bayes	71.6% (3.7%)	71.2%	74.7%	67.7%	62.8% (2.8%)	63.6%	61.9%	71.6%	52.2%
LDA	71% (3.5%)	70.7%	73.4%	68.0%	56.3% (2.9%)	57.1%	56.7%	59.0%	54.4%
QDA	71.6% (3.8%)	71.2%	74.7%	67.7%	62.8% (2.9%)	63.6%	61.9%	71.6%	52.2%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ASD	w/o Voting	Voting	Balanced	Controls	ASD
Bagged Trees	70.5% (4%)	68.8%	84.5%	53.1%	63.2% (1.6%)	63.6%	61.0%	82.7%	36.7%
Random Forest	70.5% (3.7%)	68.8%	84.3%	53.3%	63.2% (1.6%)	63.1%	60.5%	82.7%	35.6%
Rotation Forest	65.1% (3.6%)	63.2%	79.9%	46.5%	57.6% (1.8%)	58.1%	55.8%	74.8%	34.4%
Boosted Stumps	72.5% (3.6%)	72.1%	76.3%	67.8%	57.5% (2.3%)	58.1%	57.1%	65.3%	47.8%
Boosted Trees	69.6% (4%)	69.2%	73.5%	64.8%	58% (2.7%)	59.4%	58.3%	67.8%	47.8%
FC-NN	77.8% (3.6%)	77.3%	82.4%	72.1%	60.4% (1.3%)	58.1%	59.9%	73.2%	46.7%
MLP-NN	75.0% (3.6%)	74.6%	78.1%	71.1%	58.4% (2.2%)	57.1%	56.7%	59.8%	53.3%
LVQNET	70.6% (3.8%)	69.0%	83.1%	54.9%	64.7% (2.4%)	66.8%	64.2%	85.8%	40.0%
NN	71% (3.7%)	70.5%	74.3%	66.7%	65% (1.8%)	65.9%	64.9%	73.2%	55.6%
SLR	70.3% (4%)	70.0%	73.4%	66.5%	55.8% (2.7%)	58.1%	57.6%	61.4%	53.3%
RLR	77.6% (3.4%)	76.3%	87.8%	64.8%	64.2% (1%)	64.5%	60.9%	90.5%	27.8%
RVM	71.2% (3.8%)	70.6%	76.2%	64.9%	62% (1.9%)	61.3%	60.3%	68.5%	51.1%

Table 3.4. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ABIDE data when the training/validation and the hold-out test data are from different age groups for the binary classification problem between healthy controls and subjects with ASD. The training/validation data is from an age range of 7-21 years while the data from the age range of 22-58 years was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test

observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 66.8% obtained for LVQNET while the best balanced hold-out test accuracy was 64.9% obtained for KNN implemented outside the RCE framework.

ABIDE Different Sites (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ASD	w/o Voting	Voting	Balanced	Controls	ASD
Linear SVM	75.7% (3.4%)	74.7%	82.2%	67.1%	63.1% (2.5%)	65.8%	66.8%	54.7%	78.9%
RBF-SVM	75.1% (3.6%)	74.2%	80.8%	67.5%	62.8% (2.2%)	63.3%	64.1%	53.8%	74.4%
ELM	76.6 (3.3%)	75.9%	81.1%	70.7%	63.4% (1.7%)	63.3%	64.3%	51.9%	76.7%
KNN	71.1% (3.5%)	70.3%	76.7%	63.8%	62.9% (2.4%)	63.8%	64.4%	56.6%	72.2%
Naïve Bayes	72.3% (3.5%)	72.4%	71.9%	72.8%	61.5% (2.4%)	61.7%	63.1%	46.2%	80.0%
LDA	69.5% (4%)	68.8%	73.8%	63.8%	56.9% (3.3%)	63.3%	64.0%	54.7%	73.3%
QDA	72.3% (3.5%)	72.4%	71.9%	72.8%	61.5% (2.4%)	61.7%	63.1%	46.2%	80.0%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ASD	w/o Voting	Voting	Balanced	Controls	ASD
Bagged Trees	70.5% (3.5%)	68.0%	86.0%	49.9%	64.7% (2.1%)	66%	65.9%	65.0%	66.7%
Random Forest	70.5% (3.4%)	68.1%	86.2%	49.9%	64.6% (2%)	65%	65.9%	64.1%	67.7%
Rotation Forest	66.7% (3.5%)	64.7%	79.4%	50.0%	59.9% (1.8%)	60%	60.5%	56.6%	64.4%
Boosted Stumps	71.4% (3.6%)	70.7%	76.1%	65.3%	60.1% (2.7%)	62%	62.6%	57.5%	67.7%
Boosted Trees	69.8% (3.9%)	69.1%	74.3%	63.9%	58.9% (3.2%)	64%	64.0%	62.3%	65.6%
FC-NN	76.4% (3.4%)	75.6%	81.6%	69.6%	63.2% (1.9%)	62%	62.7%	50.9%	74.4%
MLP-NN	73.1% (3.8%)	72.6%	76.3%	68.8%	61.7% (2.3%)	62%	62.9%	55.7%	70.0%
LVQNET	71% (3.8%)	70.5%	74.0%	67.0%	59.7% (2.4%)	58%	59.5%	36.8%	82.2%
KNN	70.9% (3.6%)	70.1%	76.2%	64.0%	62.9% (1.4%)	64%	64.6%	54.7%	74.4%
SLR	70% (4.1%)	69.5%	73.5%	65.5%	59.8 (3%)	64%	64.3%	58.5%	70.0%
RLR	75.1% (3.3%)	73.5%	85.6%	61.3%	64% (1.7%)	63%	62.8%	62.3%	63.3%
RVM	71.1% (3.4%)	70.0%	78.1%	61.8%	63.2% (2.2%)	64%	64.7%	53.8%	75.6%

Table 3.5. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ABIDE data when the training /validation and the hold-out test data are from different imaging sites for the binary classification problem between healthy controls and subjects with ASD. The training/validation data are from 12 institutions while the data for

the remaining three institutions was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 66% obtained by Bagged trees while the best balanced hold-out test accuracy was 66.8% obtained for Linear SVM implemented within the RCE framework.

ABIDE Matched (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ASD	w/o Voting	Voting	Balanced	Controls	ASD
Linear SVM	73.1% (3.6%)	71.9%	82%	61.7%	67.5% (2.6%)	70.2%	69.2%	77.5%	60.9%
RBF-SVM	72.6% (3.5%)	70.6%	86.5%	54.7%	68.9% (2.2%)	70.7%	69.1%	82%	56.3%
ELM	73% (3.4%)	72.1%	79.4%	64.8%	66.7% (1.9%)	68.2%	67.6%	72.1%	63.2%
KNN	69% (3.6%)	67.1%	82.3%	51.8%	62.5% (2%)	63.6%	61.8%	77.5%	46%
Naïve Bayes	70.4% (3.5%)	69.9%	74%	65.7%	61.8% (1.7%)	61.1%	60.8%	64%	57.5%
LDA	66.6% (3.8%)	66%	70.6%	61.4%	59.3% (3%)	61.1%	60.8%	64%	57.5%
QDA	70.4% (3.7%)	69.9%	73.9%	65.8%	61.9% (1.5%)	61.1%	60.8%	64%	57.5%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ASD	w/o Voting	Voting	Balanced	Controls	ASD
Bagged Trees	68.3% (3.4%)	66.3%	82.7%	49.8%	64% (1.4%)	64.1%	62.2%	78.4%	46.0%
Random Forest	68.3% (3.5%)	66.2%	82.7%	49.6%	63.8% (1.4%)	64.1%	62.3%	77.5%	47.1%
Rotation Forest	63.7% (3.7%)	60.8%	84.6%	36.9%	57.8% (1.8%)	59.6%	57%	78.4%	35.6%
Boosted Stumps	70.1% (3.7%)	69.4%	75%	63.8%	62.1% (2.5%)	64.6%	64%	69.4%	58.6%
Boosted Trees	67.6% (4%)	67%	71.8%	62.1%	60.7% (2.9%)	66.7%	66.2%	70.3%	62.1%
FC-NN	73.3% (3.5%)	72.3%	80.8%	63.7%	66.9% (1.7%)	69.2%	68.2%	76.6%	59.8%
MLP-NN	70.6% (3.5%)	70%	75.6%	64.3%	62.7% (2.2%)	63.6%	63.5%	64.9%	62.1%
LVQNET	69.4% (3.9%)	68.7%	74.2%	63.1%	61.7% (2.3%)	63.1%	62.6%	67.6%	57.5%
KNN	69% (3.6%)	67.1%	82.5%	51.6%	62.4% (1.3%)	64.1%	62.3%	77.5%	47.1%
SLR	67.3% (3.8%)	66.8%	70.9%	62.7%	59.5% (2.8%)	61.6%	60.8%	67.6%	54%
RLR	73.5% (3.6%)	71.9%	84.4%	59.4%	64.7% (1.2%)	64.1%	62.7%	74.8%	50.6%
RVM	68.3% (3.5%)	67.4%	75.1%	59.6%	61.8% (1.7%)	62.1%	61%	70.3%	51.7%

Table 3.6. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ABIDE data when the training/validation and the hold-out test data are matched in imaging sites as well as age group for the binary classification problem between healthy controls and subjects with ASD. The training/validation and the hold-out test data are from all 15 imaging sites and age range of 7-58 years. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 70.7% obtained by RBF-SVM within the RCE framework while the best balanced hold-out test accuracy was 69.2% obtained for Linear SVM implemented within the RCE framework.

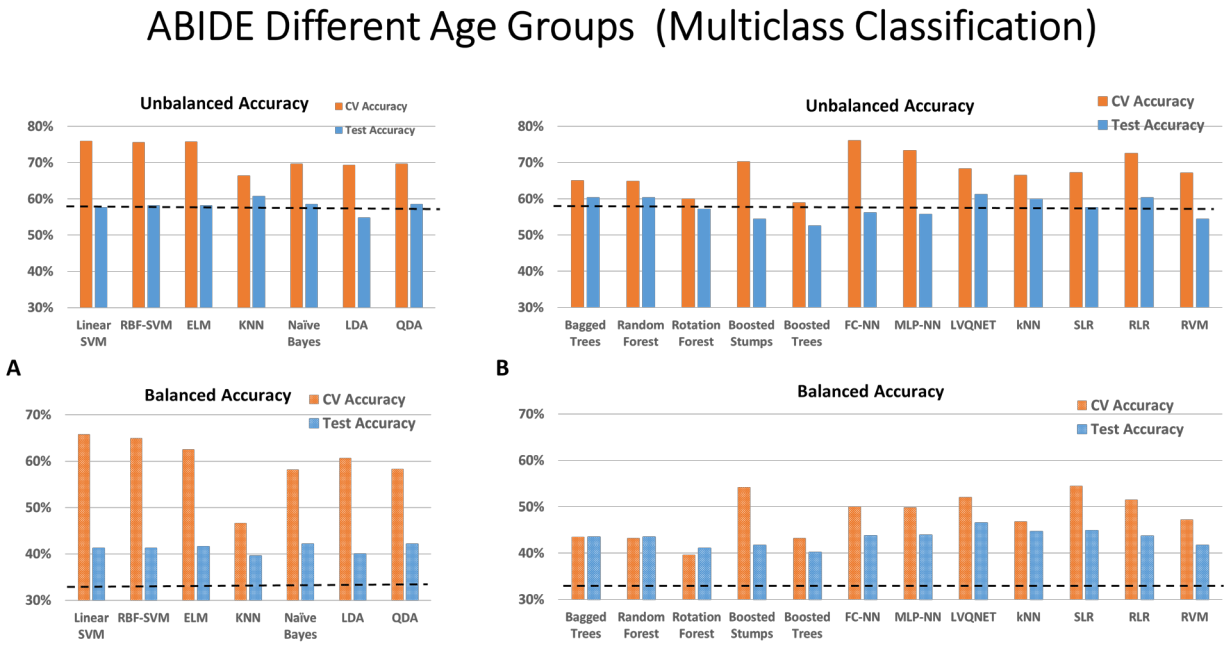


Figure 3.10. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ABIDE data when the training/validation data and the hold-out test data are from different age groups for the multiclass classification between healthy controls, subject with Asperger's syndrome and Autism. The training/validation data is from an age range of 7-21 years while the data from the age range of 22-58 years was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 58.5% since healthy controls formed 58.5% of the total size of the hold-out test data. For balanced accuracy, this is 33.3%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the

training data. The considerable difference between the unbalanced and the balanced accuracies can be attributed to the fact that all classifiers were unsuccessful in classifying subjects with Asperger's Syndrome due to their relative lower number of observations in the dataset. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 66.8% obtained by LVQNET while the best balanced hold-out test accuracy was 64.9% obtained for KNN implemented outside the RCE framework.

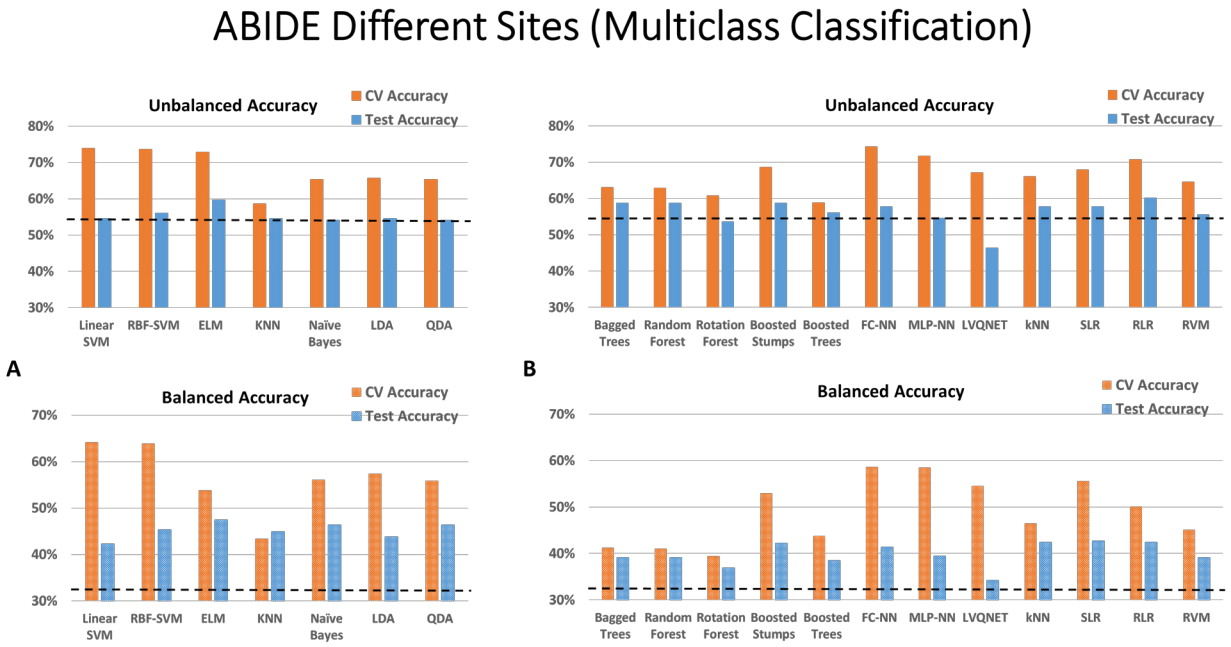


Figure 3.11. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ABIDE data when the training/validation data and the hold-out test data are from different imaging sites for the multiclass classification between healthy controls, subject with Asperger's syndrome and Autism. The training/validation data are from 12 institutions while the data for the remaining three institutions was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 54% since healthy controls formed 54% of the total size of the hold-out test data. For balanced accuracy, this is 33.3%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The considerable difference between the unbalanced and the balanced accuracies can be attributed to the fact that all classifiers were unsuccessful in classifying subjects with Asperger's Syndrome due to their relative lower number of observations in the dataset. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 66% obtained for Regularized Logistic Regression (RLR) while the best balanced hold-out test

accuracy was 66.8% obtained for Extreme Learning Machines (ELM) implemented within the RCE framework.

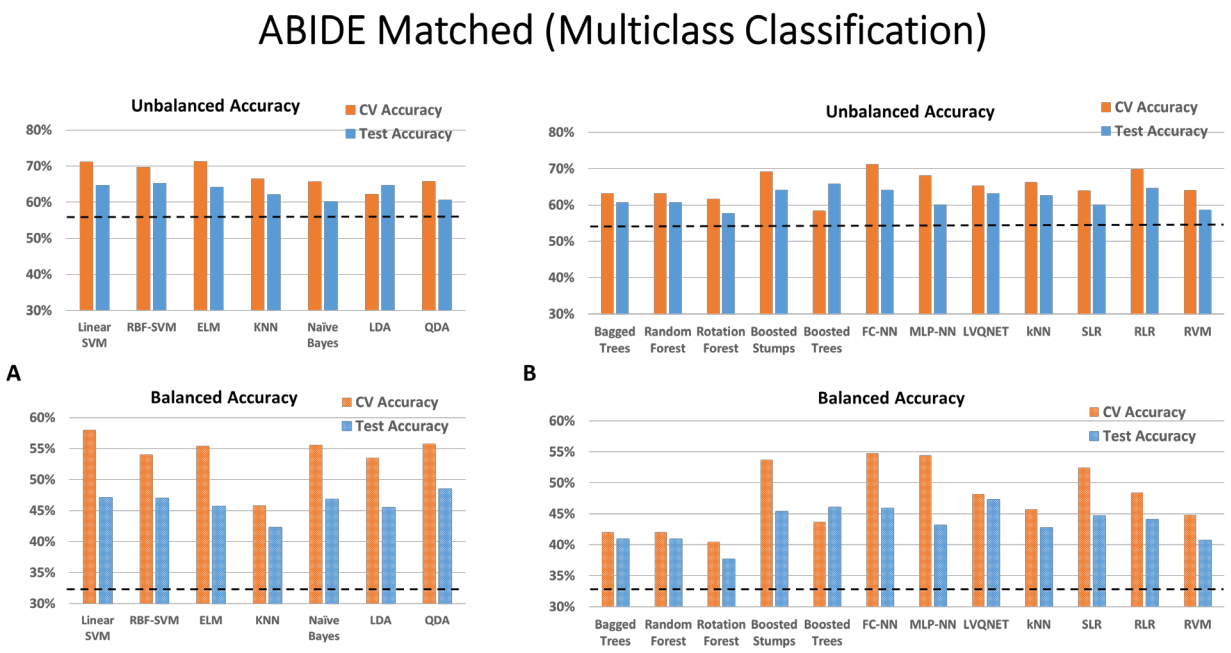


Figure 3.12. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ABIDE data when the training/validation data and the hold-out test data are matched in imaging sites as well as age group for the multiclass classification problem between healthy controls and subjects with Asperger's syndrome and Autism. The training/validation data are from 12 institutions while the data for the remaining three institutions was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 56% since healthy controls formed 56% of the total size of the hold-out test data. For balanced accuracy, this is 33.3%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The considerable difference between the unbalanced and the balanced accuracies can be attributed to the fact that all classifiers were unsuccessful in classifying subjects with Asperger's Syndrome due to their relative lower number of observations in the dataset. Some of the discrepancies between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 70.7% obtained by Boosted Trees while the best balanced hold-out test accuracy was 47.3% obtained for LVQNET.

ABIDE Different Age Groups (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	Asperger's	Autism	w/o Voting	Voting	Balanced	Controls	Asperger's	Autism
Linear SVM	76% (3.3%)	65.8%	84.0%	42.0%	71.4%	55.3% (2%)	57.6%	41.3%	75.6%	0.0%	48.3%
RBF-SVM	75.6% (3.6%)	64.9%	84.9%	40.6%	69.3%	55% (2%)	58.1%	41.3%	77.2%	0.0%	46.7%
ELM	75.8% (3.2%)	62.5%	84.9%	30.3%	72.2%	56.3% (1.6%)	58.1%	41.6%	76.4%	0.0%	48.3%
KNN	66.4% (3.7%)	46.6%	81.5%	0.0%	58.3%	57.9% (2.7%)	60.8%	39.6%	90.5%	0.0%	28.3%
Naïve Bayes	69.7% (3.8%)	58.2%	76.4%	30.0%	68.3%	57.5% (1.9%)	58.5%	42.1%	76.4%	0.0%	50.0%
LDA	69.3% (3.8%)	60.7%	75.8%	40.5%	65.8%	52.2% (2.6%)	54.8%	40.0%	70.1%	0.0%	50.0%
QDA	69.7% (4%)	58.3%	76.3%	30.3%	68.4%	57.5% (2%)	58.5%	42.1%	76.4%	0.0%	50.0%

B

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	Asperger's	Autism	w/o Voting	Voting	Balanced	Controls	Asperger's	Autism
Bagged Trees	65.1% (3%)	43.5%	92.3%	0.0%	38.1%	59.7% (1.2%)	60.4%	43.5%	93.7%	0.0%	20.0%
Random Forest	64.9% (3.1%)	43.2%	92.1%	0.0%	37.6%	60% (1.3%)	60.4%	43.5%	93.7%	0.0%	20.0%
Rotation Forest	60% (2.9%)	39.6%	88.1%	0.2%	30.4%	56.3% (2.1%)	57.1%	41.1%	89.0%	0.0%	18.3%
Boosted Stumps	70.3% (3.6%)	54.2%	82.2%	16.5%	64.0%	53.2% (2.1%)	54.4%	41.7%	75.6%	0.0%	36.7%
Boosted Trees	58.9% (4%)	43.2%	68.3%	4.9%	56.5%	50.1% (3.2%)	52.5%	40.2%	73.2%	0.0%	35.0%
FC-NN	76.1% (3.2%)	61.3%	86.3%	26.0%	71.7%	55.5% (1.8%)	56.2%	43.8%	75.6%	0.0%	43.3%
MLP-NN	73.3% (3.6%)	61.2%	81.6%	31.9%	70.0%	54.1% (2.2%)	55.8%	43.9%	73.2%	0.0%	46.7%
LVQNET	68.3% (3.6%)	52.1%	84.8%	17.2%	54.4%	58.8% (3.5%)	61.3%	46.5%	86.6%	0.0%	38.3%
KNN	66.5% (3.4%)	46.8%	81.1%	0.0%	59.2%	58.6% (2.3%)	59.9%	44.7%	87.4%	0.0%	31.7%
SLR	67.2% (3.7%)	54.5%	76.4%	24.3%	62.8%	53.2% (2.6%)	57.6%	44.9%	77.2%	0.0%	45.0%
RLR	72.5% (3.1%)	51.5%	93.0%	5.2%	56.2%	60.6% (0.8%)	60.4%	43.7%	92.9%	0.0%	21.7%
RVM	67.1% (3.4%)	47.2%	81.8%	0.0%	59.8%	54.9% (2%)	54.4%	41.7%	75.6%	0.0%	36.7%

Table 3.7. The table shows the cross-validation and the hold out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ABIDE data when the training/validation and the hold-out test data are from different age groups for the multiclass classification between healthy controls, subjects with Asperger's syndrome and Autism. The training/validation data is from an age range of 7-21 years while the data from the age range of 22-58 years was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 61.3% while the best balanced hold-out test accuracy obtained was 46.5%, both for LVQNET.

ABIDE Different Sites (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	Asperger's	Autism	w/o Voting	Voting	Balanced	Controls	Asperger's	Autism
Linear SVM	73.9% (3.2%)	64.2%	82.3%	41.7%	68.5%	54.1% (2.3%)	54.6%	42.3%	67.9%	5.3%	47.9%
RBF-SVM	73.7% (3.2%)	63.9%	83.5%	41.9%	66.2%	53.7% (2.3%)	56.1%	45.3%	68.9%	5.3%	50.7%
ELM	72.9% (3.2%)	53.8%	89.9%	9.9%	61.6%	59.7% (1.5%)	59.7%	47.4%	70.7%	0%	59.1%
KNN	58.6% (3.8%)	43.4%	67.4%	3.7%	59.1%	51.8% (3.4%)	54.6%	44.9%	51.9%	0%	73.2%
Naïve Bayes	65.4% (3.9%)	56.1%	76.7%	37.5%	54.1%	49% (5.9%)	54.1%	46.4%	67.0%	21%	43.7%
LDA	65.7% (4%)	57.4%	73.8%	38.7%	59.7%	47.7% (3.5%)	54.6%	43.8%	68.9%	5.3%	46.5%
QDA	65.4% (4.1%)	55.9%	76.6%	37.0%	54.2%	49% (6%)	54.1%	46.4%	67.0%	21%	43.7%

B

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	Asperger's	Autism	w/o Voting	Voting	Balanced	Controls	Asperger's	Autism
Bagged Trees	63.1% (2.8%)	41.2%	92.6%	0%	31.1%	58.5% (1.8%)	58.7%	39.1%	90.6%	0%	26.8%
Random Forest	62.9% (2.8%)	41%	92.8%	0%	30.2%	58.9% (1.5%)	58.7%	39.1%	90.6%	0%	26.8%
Rotation Forest	60.8% (3%)	39.5%	90%	0%	28.6%	54% (1.9%)	53.6%	36.9%	75.5%	0%	35.2%
Boosted Stumps	68.6% (3.5%)	53%	82.4%	16.9%	59.7%	53.7% (2.6%)	58.7%	42.2%	71.7%	0%	54.9%
Boosted Trees	58.8% (3.9%)	43.8%	70.4%	7.7%	53.4%	49.6% (3.5%)	56.1%	38.5%	80.2%	0%	35.2%
FC-NN	74.3% (5.5%)	58.6%	87.4%	21.5%	66.9%	57.4% (2%)	57.7%	41.4%	70.7%	0%	53.5%
MLP-NN	71.7% (3.6%)	58.5%	81.8%	26.4%	67.2%	54% (3.1%)	54.6%	39.5%	65.1%	0%	53.5%
LVQNET	67.1% (3.5%)	54.5%	82.5%	29.1%	51.8%	47.6% (3.1%)	46.4%	34.2%	59.4%	5.3%	38%
KNN	66% (3.3%)	46.5%	82%	0%	57.4%	57.6% (1.7%)	57.7%	42.4%	65.1%	0%	62%
SLR	67.9% (3.6%)	55.6%	78.2%	26.4%	62.2%	51.3% (2%)	57.7%	42.7%	63.2%	0%	64.8%
RLR	70.7% (3.1%)	50.1%	90.9%	4%	55.3%	60.8% (1.2%)	60.2%	42.4%	79.2%	0%	47.9%
RVM	64.6% (3.4%)	45.1%	83.6%	1.5%	50.2%	55 % (1.3%)	55.6%	39.1%	73.6%	0%	43.7%

Table 3.8. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ABIDE data when the training/validation and the hold-out test data are from different imaging sites for the multiclass classification problem between healthy controls and subjects with Asperger's syndrome and Autism. The training/validation data are from 12 institutions while the data for the remaining three institutions was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 60.2% obtained for Regularized Logistic Regression (RLR) while the best balanced hold-out test accuracy was 47.4% obtained for Extreme Learning Machines (ELM) implemented within the RCE framework.

ABIDE Matched (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	Asperger's	Autism	w/o Voting	Voting	Balanced	Controls	Asperger's	Autism
Linear SVM	71.2% (3.3%)	58%	84.5%	28.5%	61.1%	62.9% (2.3%)	64.6%	47.1%	80.2%	5.3%	55.9%
RBF-SVM	69.6% (3.1%)	54%	86.6%	20.2%	55.3%	64% (2.4%)	65.2%	47.0%	82.9%	5.3%	52.9%
ELM	71.3% (3.1%)	55.4%	85.6%	18.4%	62.3%	63% (1.7%)	64.1%	45.7%	78.4%	0%	58.8%
KNN	66.5% (3.2%)	45.8%	87.5%	0%	50%	60.4% (1.7%)	62.1%	42.3%	85.6%	0%	41.2%
Naïve Bayes	65.6% (3.8%)	55.6%	74.5%	32.3%	60.1%	57.5% (1.8%)	60.1%	46.8%	73%	15.8%	51.5%
LDA	62.2% (3.5%)	53.5%	69.6%	32.8%	58%	54.9% (3.1%)	64.6%	45.5%	82%	0%	54.4%
QDA	65.7% (3.7%)	55.8%	74.5%	32.6%	60.3%	57.6% (1.9%)	60.6%	48.5%	73%	21.1%	51.5%

B

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	Asperger's	Autism	w/o Voting	Voting	Balanced	Controls	Asperger's	Autism
Bagged Trees	63.1% (3.1%)	42.0%	90.1%	0%	36%	60.8% (1.2%)	60.6%	41%	84.7%	0%	38.2%
Random Forest	63.1% (3.1%)	42.0%	90.3%	0%	35.6%	60.4% (1.2%)	60.6%	41%	84.7%	0%	38.2%
Rotation Forest	61.6% (2.9%)	40.5%	90.4%	0%	31%	57.6% (1.3%)	57.6%	37.7%	86.5%	0%	26.5%
Boosted Stumps	69.1% (3.4%)	53.7%	83.1%	17.7%	60.2%	62.2% (2%)	64.1%	45.4%	80.2%	0%	55.9%
Boosted Trees	58.4% (3.9%)	43.7%	70.7%	8.3%	52%	54.5% (3%)	65.7%	46.1%	83.8%	0%	54.4%
FC-NN	71.1% (3.1%)	54.7%	86.5%	16.8%	60.7%	63.7% (1.6%)	64.1%	45.9%	77.5%	0%	60.3%
MLP-NN	68.1% (3.7%)	54.4%	79.3%	21.5%	62.5%	59.6% (2.5%)	60.1%	43.2%	72.1%	0%	57.4%
LVQNET	65.2% (3.7%)	48.2%	79.6%	8.1%	57%	59.7% (2.2%)	63.1%	47.3%	78.4%	10.5%	52.9%
KNN	66.2% (3.1%)	45.7%	87.5%	0.1%	49.4%	60.7% (1.3%)	62.6%	42.7%	85.6%	0%	42.6%
SLR	63.9% (3.6%)	52.4%	74.2%	25.5%	57.4%	57.2% (2.6%)	60.1%	44.6%	71.2%	5.3%	57.4%
RLR	69.8% (3%)	48.4%	91.0%	0.5%	53.8%	63.4% (1%)	64.6%	44.1%	88.3%	0%	44.1%
RVM	64% (3.5%)	44.8%	81.5%	0%	52.8%	58% (1.3%)	58.6%	40.7%	76.6%	0%	45.6%

Table 3.9. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ABIDE data when the training/validation and the hold-out test data are matched in imaging sites as well as age group for the multiclass classification problem between healthy controls and subjects with Asperger's syndrome and Autism. The training/validation data and the hold-out test data are from all 15 imaging sites and age range of 7-58 years. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 70.7% obtained by Boosted Trees within the RCE framework while the best balanced hold-out test accuracy was 47.3% obtained for LVQNET.

In the binary classification scenario for the split in which the training/validation and the hold-out test data belong to different age ranges, the best hold-out test accuracy was 66.8% obtained by LVQNET while the best-balanced hold-out test accuracy was 64.9% obtained from KNN implemented outside the RCE framework. In the multiclass classification for the same split, the best hold-out test accuracy was 66.8% obtained by LVQNET while the best-balanced hold-out test accuracy was 64.9% obtained for KNN implemented outside the RCE framework. When the training/validation data is from 12 imaging sites, and the hold-out test data is from the remaining three imaging sites, the best accuracy on the hold-out test data was 66% obtained for Bagged trees while the best-balanced hold-out test accuracy was 66.8% obtained for Linear SVM implemented within the RCE framework. In the multiclass scenario between healthy controls, subjects with Asperger's syndrome and Autism, the best hold-out test accuracy was 66% obtained for RLR while the best-balanced hold-out test accuracy was 66.8% obtained for ELM implemented within the RCE framework. Finally in the third split wherein the training/validation and the holdout test data are matched for age and imaging site, the binary classification results were higher compared to the unmatched cases with the best hold-out test accuracy at 70.7% obtained by RBF-SVM within the RCE framework while the best-balanced hold-out test accuracy was 69.2% obtained for Linear SVM implemented within the RCE framework. For the 3-way classification, the best hold-out test accuracy was 70.7% achieved by Boosted Trees while the best-balanced hold-out test accuracy was 47.3% obtained for LVQNET.

Feature importance: The combined feature importance scores (CFIS) were calculated for all classifiers implemented within the RCE framework. These combined feature importance scores for various splits are plotted in a scatter plot as shown in Figure 3.13. The figure indicates that, though there is significant ($p<10^{-10}$) agreement in the feature importance scores across the splits,

age range, and scanner variability do contribute to the increase variance in these score estimates. Using the feature importance scores, we identified the top connectivity paths whose means were significantly different between the groups ($p<0.05$, FDR corrected) as well as have high feature importance scores. These paths are visualized in Figure 3.14. Along with these connectivity paths, we also identified the top 20 regions associated with altered and discriminative connectivity paths as shown in the Table 3.10.

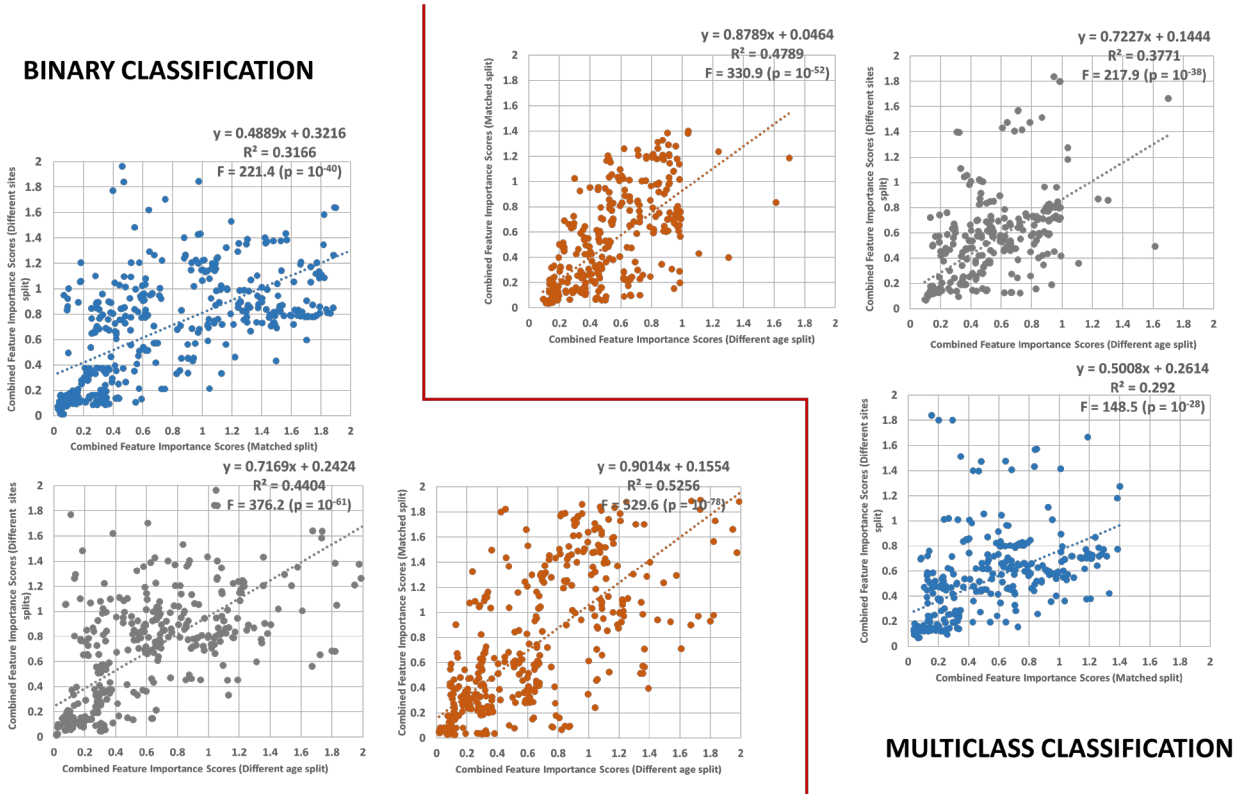
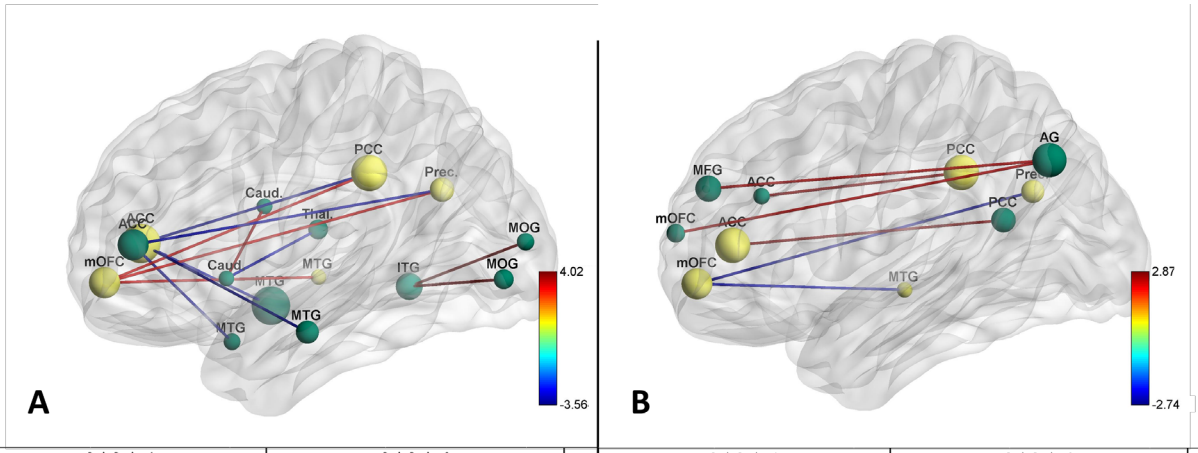


Figure 3.13. Scatter plots of combined feature importance scores (CFIS) for the three splits performed on the ABIDE data. We selected the common features obtained in all three splits and plotted them as scatter plots with two splits at a time for both binary and multiclass classification scenarios. The plot illustrates that there is significant agreement in the CFIS across splits. However a lot of variability is present as well and this can be attributed to age and site variability of the training/validation data obtained from the three splits.



Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(48.00,-52.70,-16.70)	Inferior temporal gyrus	ITG	(-39.60,-85.40,1.40)	Middle occipital gyrus	MOG	3.50
(1.40,55.90,-7.20)	Medial orbitofrontal cortex	mOFC	(13.30,-64.40,24.00)	Precuneus	Prec.	3.46
(1.40,55.90,-7.20)	Medial orbitofrontal cortex	mOFC	(1.90,-38.20,32.20)	Posterior cingulate cortex	PCC	3.41
(1.40,55.90,-7.20)	Medial orbitofrontal cortex	mOFC	(61.90,-21.10,-15.60)	Middle temporal gyrus	MTG	3.56
(-25.50,-92.90,12.00)	Middle occipital gyrus	MOG	(48.00,-52.70,-16.70)	Inferior temporal gyrus	ITG	3.54
(55.10,-3.60,-25.40)	Middle temporal gyrus	MTG	(6.70,42.60,6.10)	Anterior cingulate cortex	ACC	-3.35
(43.50,9.90,-36.20)	Middle temporal gyrus	MTG	(-6.80,45.70,7.80)	Anterior cingulate cortex	ACC	-3.30
(1.90,-38.20,32.20)	Posterior cingulate cortex	PCC	(6.70,42.60,6.10)	Anterior cingulate cortex	ACC	-3.56
(-13.80,12.60,-2.90)	Caudate	Caud.	(11.60,-20.00,9.50)	Thalamus	Thal.	-3.42
(-6.80,45.70,7.80)	Anterior cingulate cortex	ACC	(13.30,-64.40,24.00)	Precuneus	Prec.	-3.49
(-56.40,-15.10,-15.30)	Middle temporal gyrus	MTG	(6.70,42.60,6.10)	Anterior cingulate cortex	ACC	-3.51
(14.20,-0.50,17.50)	Caudate	Caud.	(-13.80,12.60,-2.90)	Caudate	Caud.	4.02

Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(1.40,55.90,-7.20)	Medial orbitofrontal cortex	mOFC	(13.30,-64.40,24.00)	Precuneus	Prec.	-2.74
(1.40,55.90,-7.20)	Medial orbitofrontal cortex	mOFC	(61.90,-21.10,-15.60)	Middle temporal gyrus	MTG	-2.58
(10.70,63.00,10.00)	Medial orbitofrontal cortex	mOFC	(-42.90,-67.00,40.70)	Angular gyrus	AG	2.42
(1.90,-38.20,32.20)	Posterior cingulate cortex	PCC	(1.60,33.30,24.30)	Anterior cingulate cortex	ACC	2.86
(1.50,-52.80,14.80)	Posterior cingulate cortex	PCC	(6.70,42.60,6.10)	Anterior cingulate cortex	ACC	2.88
(0.30,51.60,26.50)	Middle frontal gyrus	MFG	(-42.90,-67.00,40.70)	Angular gyrus	AG	2.50

Figure 3.14. The figure illustrates the connectivity paths which have significantly different means between the groups ($p<0.05$, corrected for multiple comparisons using permutation test) as well as are among the top hundred most discriminative paths in ABIDE dataset for (A) binary classification between controls and ASD. (B) 3-way classification between healthy controls, Asperger's syndrome and Autism. The size of the nodes indicates the relative importance of the region (Table 3.10). Common nodes between binary and multiclass classification are indicated in yellow while other nodes are indicated in green. The sign of the paths indicates over-connectivity (positive) or under-connectivity (negative) in healthy controls compared to clinical populations. Consequently, red represents a higher connectivity between controls compared to the diseased populations and blue represents a lower connectivity. The numerical values in the color bar denote the combined feature importance score of the path (CFIS) obtained from classification. A higher absolute number indicates more discriminative ability for the functional connectivity path. The table below the figure tabulates the brain regions involved in the paths visualized above along with the abbreviations of the two regions and the CFIS (combined feature importance score) for the connectivity paths.

A

Rank	X	Y	Z	Brain Regions	Abbr.
1	7	43	6	Anterior cingulate cortex	ACC
2	55	-4	-25	Middle temporal gyrus	MTG
3	2	-38	32	Posterior cingulate cortex	PCC
4	-43	-67	41	Angular gyrus	AG
5	0	39	-12	Medial orbitofrontal cortex	mOFC
6	-42	41	13	Inferior frontal gyrus	IFG
7	1	56	-7	Medial orbitofrontal cortex	mOFC
8	-7	46	8	Anterior cingulate cortex	ACC
9	54	-55	24	Angular gyrus	AG
10	2	-53	15	Precuneus	Prec.
11	48	-53	-17	Inferior temporal gyrus	ITG
12	45	41	15	Middle frontal gyrus	MFG
13	-10	-67	22	Cuneus	Cun.
14	3	12	49	Supplementary motor area	SMA
15	0	52	27	Middle frontal gyrus	MFG
16	2	-17	35	Cingulate gyrus	CG
17	-57	-43	27	Supramarginal gyrus	SMG
18	-52	-50	42	Inferior parietal lobule	IPL
19	-9	62	14	Middle frontal gyrus	MFG
20	13	-64	24	Precuneus	Prec.

B

Rank	X	Y	Z	Brain Regions	Abbr.
1	-7	46	8	Anterior cingulate cortex	ACC
2	55	-4	-25	Middle temporal gyrus	MTG
3	2	-38	32	Posterior cingulate cortex	PCC
4	7	43	6	Anterior cingulate cortex	ACC
5	-43	-67	41	Angular gyrus	AG
6	1	56	-7	Medial orbitofrontal cortex	mOFC
7	-42	41	13	Inferior frontal gyrus	IFG
8	0	52	27	Middle frontal gyrus	MFG
9	2	-17	35	Cingulate gyrus	CG
10	2	-53	15	Cingulate gyrus	CG
11	0	39	-12	Medial orbitofrontal cortex	mOFC
12	45	41	15	Middle frontal gyrus	MFG
13	3	12	49	Supplementary motor area	SMA
14	54	-55	24	Angular gyrus	AG
15	13	-64	24	Precuneus	Prec.
16	-10	-67	22	Cuneus	Cun.
17	60	-45	11	Supramarginal gyrus	SMG
18	-39	6	-39	Inferior temporal gyrus	ITG
19	-57	-43	27	Supramarginal gyrus	SMG
20	12	-20	10	Thalamus	Thal.

Table 3.10. This table illustrates the top 20 regions for ASD as identified by (A) Binary classification between Healthy Controls and ASD (B) Multiclass classification between Controls, Asperger's syndrome and Autism, using ABIDE data.

3.3.3 ADHD-200

Classification results: For the ADHD-200 dataset, the classification results for the binary classification scenario between healthy controls and subjects with ADHD are shown in Figure 3.15. Table 3.11 provides corresponding detailed individual class accuracies. Results for the 3-way classification scenario between healthy controls and subjects with ADHD-I and ADHD-C are provided in Figure 3.16 with detailed accuracy performance presented in Table 3.12. The results indicate the apparent difficulty in classifying controls from ADHD as reported by several papers which used the same data with reported performances similar to our own results [31]. For binary classification, the best hold-out test accuracy was 61.4% while the best balanced hold-out test accuracy obtained was 59.6% using Boosted Stumps. Similarly, for the multiclass classification, the best hold-out test accuracy was 58% using Boosted Trees while the best balanced hold-out test accuracy obtained was 38.7% using RBF-SVM implemented within RCE

framework. These results indicate the difficulty of multiclass classification with ADHD-200 data compared to a binary classification.

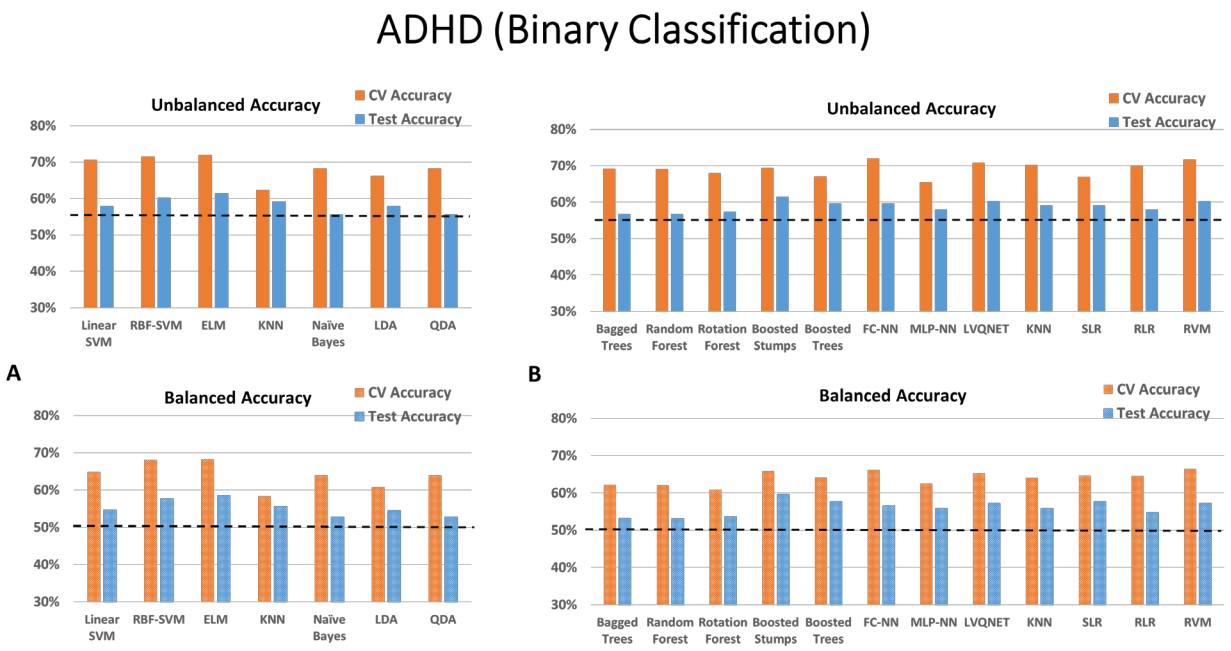


Figure 3.1520. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for the ADHD-200 data between healthy controls and subjects with ADHD. The training/validation data and the hold-out test data are from 7 imaging sites as released by ADHD-200 consortium. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 55% since healthy controls formed 55% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The best hold-out test accuracy was 61.4% while the best balanced hold- out test accuracy obtained was 59.6% obtained for Boosted Stumps.

ADHD (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ADHD	w/o Voting	Voting	Balanced	Controls	ADHD
Linear SVM	70.6% (3.3%)	64.9%	86.4%	43.5%	58.1% (1.9%)	57.9%	54.6%	87.2%	22.1%
RBF-SVM	71.4% (3.4%)	68.1%	80.7%	55.5%	59.3% (2.1%)	60.2%	57.7%	83.0%	32.5%
ELM	71.9% (3.3%)	68.2%	82.3%	54.1%	59.9% (1.4%)	61.4%	58.5%	87.2%	29.9%
KNN	62.3% (4.6%)	58.3%	73.8%	42.8%	55% (3.9%)	59.1%	55.6%	90.4%	20.8%
Naïve Bayes	68.2% (3.7%)	63.9%	80.4%	47.5%	55.8% (1.7%)	55.6%	52.8%	80.9%	24.7%
LDA	66.2% (4.2%)	60.8%	81.3%	40.4%	56.1% (3%)	57.9%	54.4%	89.4%	19.5%
QDA	68.2% (3.7%)	63.9%	80.4%	47.5%	55.8% (1.7%)	55.6%	52.8%	80.9%	24.7%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	ADHD	w/o Voting	Voting	Balanced	Controls	ADHD
Bagged Trees	69.1% (3.3%)	62.2%	88.6%	35.8%	57% (1%)	56.7%	53.2%	89.4%	16.9%
Random Forest	69% (3.1%)	62.1%	88.4%	35.9%	56.9% (1.1%)	56.7%	53.1%	89.4%	16.9%
Rotation Forest	67.9% (3%)	60.8%	87.8%	33.8%	56.2% (1.2%)	57.3%	53.6%	91.5%	15.6%
Boosted Stumps	69.4% (3.7%)	65.9%	79.1%	52.7%	58.5% (2.6%)	61.4%	59.6%	77.7%	41.6%
Boosted Trees	67% (4%)	64.1%	75.3%	52.8%	57.8% (2.9%)	59.6%	57.7%	77.7%	37.7%
FC-NN	71.9% (3.3%)	66.2%	88.1%	44.3%	58.7% (1.3%)	59.6%	56.5%	88.3%	24.7%
MLP-NN	65.4% (4%)	62.5%	73.8%	51.2%	57.2% (2.7%)	57.9%	55.8%	76.6%	35.1%
LVQNET	70.8% (3.2%)	65.2%	86.5%	43.9%	60.2% (0.9%)	60.2%	57.2%	88.3%	26.0%
KNN	70.2% (3.4%)	64.0%	87.7%	40.3%	57.6% (1.3%)	59.1%	55.8%	88.3%	23.4%
SLR	66.9% (4%)	64.6%	73.4%	55.8%	56.6% (2.8%)	59.1%	57.7%	71.3%	44.2%
RLR	70% (3.4%)	64.5%	85.5%	43.6%	57.5% (1.2%)	57.9%	54.8%	86.2%	23.4%
RVM	71.7% (3.4%)	66.5%	86.3%	46.7%	60.6% (1.4%)	60.2%	57.2%	87.2%	27.3%

Table 3.11. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ADHD-200 data for the binary classification problem between healthy controls and subjects with ASD. The training/validation and the hold-out test data are from 7 imaging sites. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are

highlighted. The best hold-out test accuracy was 61.4% while the best balanced hold- out test accuracy obtained was 59.6% for Boosted Stumps.

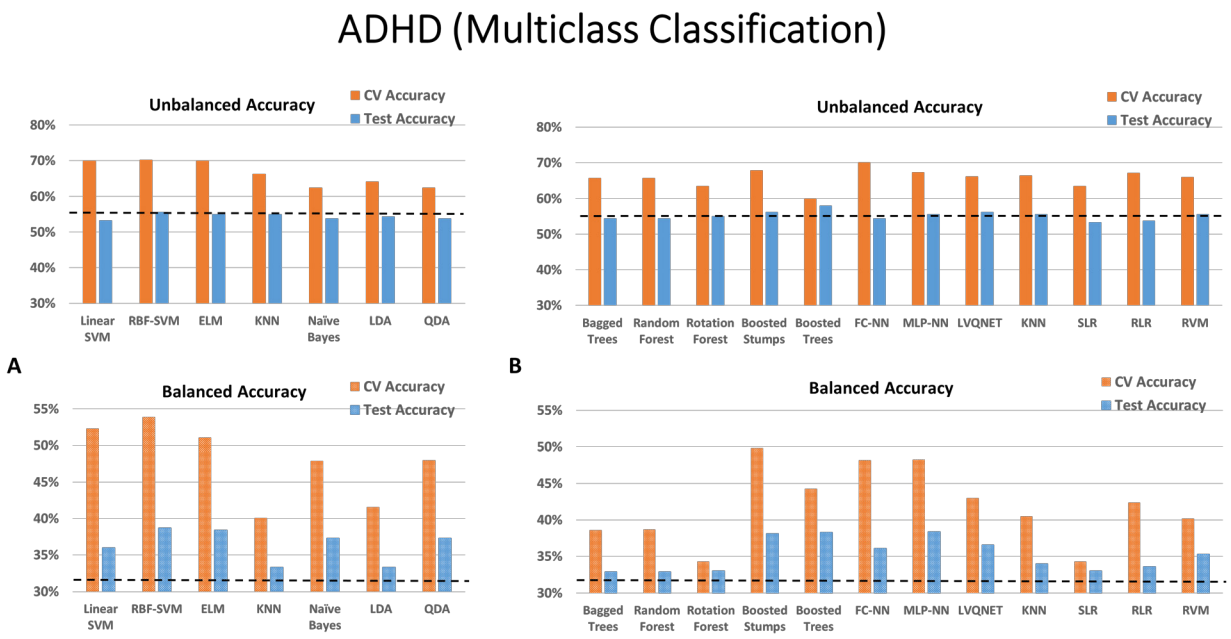


Figure 3.16. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for the 3-way classification scenario for the ADHD-200 data between healthy controls and subjects with ADHD-I and ADHD-C. The training/validation data and the hold-out test data are from 7 imaging sites as released by the ADHD-200 consortium. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 55.6% since healthy controls formed 55.6% of the total size of the hold-out test data. For balanced accuracy, this is 33.3%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The best hold-out test accuracy was 58% for Boosted Trees while the best balanced hold- out test accuracy obtained was 38.7% obtained for RBF-SVM implemented within RCE framework.

ADHD (Multiclass Classification)

A

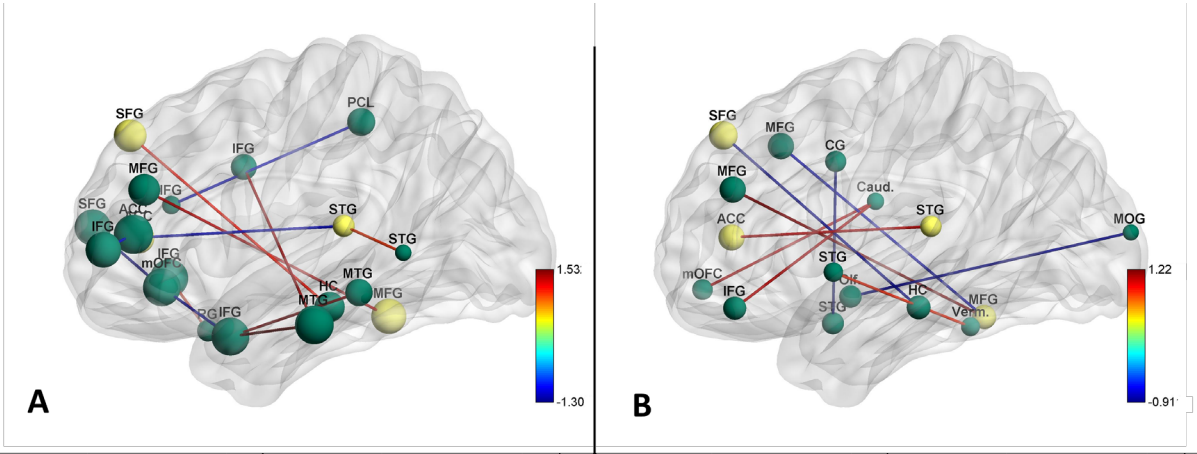
Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	ADHD-C	ADHD-I	w/o Voting	Voting	Balanced	Controls	ADHD-C	ADHD-I
Linear SVM	69.9% (3%)	52.3%	89.4%	41.4%	26.1%	54.6% (1.5%)	53.3%	36%	88.3%	8.2%	11.5%
RBF-SVM	70.2% (3.1%)	53.9%	88.4%	43.8%	29.3%	54.6% (1.7%)	55.6%	38.7%	88.3%	16.3%	11.5%
ELM	69.9% (3%)	51.1%	90.1%	43.3%	19.9%	55.5% (1.2%)	55%	38.4%	87.2%	16.3%	11.5%
KNN	66.2% (2.3%)	40.1%	95.3%	23.4%	1.7%	55.4% (1.1%)	55%	33.3%	97.9%	2%	0%
Naïve Bayes	62.5% (3.7%)	47.9%	78.3%	40.7%	24.8%	52.3% (2.3%)	53.8%	37.3%	86.2%	14.3%	11.5%
LDA	64.2% (3.1%)	41.6%	89.5%	26.2%	9%	52.5% (2.7%)	54.4%	33.3%	95.7%	4.1%	0%
QDA	62.5% (3.8%)	48%	78.5%	40.4%	25.1%	52.4% (2.3%)	53.8%	37.3%	86.2%	14.3%	11.5%

B

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	ADHD-C	ADHD-I	w/o Voting	Voting	Balanced	Controls	ADHD-C	ADHD-I
Bagged Trees	65.7% (2.1%)	38.6%	96.1%	19.7%	0.1%	54.8% (0.8%)	54.4%	32.9%	96.8%	2.0%	0.0%
Random Forest	65.7% (2.1%)	38.7%	96.0%	19.8%	0.1%	54.7% (0.8%)	54.4%	32.9%	96.8%	2.0%	0.0%
Rotation Forest	63.5% (1.5%)	34.3%	97.4%	5.4%	0.1%	55.1% (0.4%)	55.0%	33.0%	98.9%	0.0%	0.0%
Boosted Stumps	67.8% (3.3%)	49.8%	87.6%	40.1%	21.7%	55.3% (2.1%)	56.2%	38.1%	90.4%	16.3%	7.7%
Boosted Trees	60% (3.8%)	44.3%	77.3%	36.6%	18.9%	51.7% (3.2%)	58.0%	38.3%	94.7%	16.3%	3.8%
FC-NN	70.1% (2.8%)	48.2%	94.1%	36.3%	14.3%	54.8% (1.1%)	54.4%	36.1%	90.4%	10.2%	7.7%
MLP-NN	67.3% (3.5%)	48.3%	87.4%	43.7%	13.7%	53.8% (1.9%)	55.6%	38.4%	87.2%	20.4%	7.7%
LVQNET	66.1% (2.8%)	43.0%	91.6%	28.9%	8.4%	56.2% (1.6%)	56.2%	36.6%	93.6%	12.2%	3.8%
KNN	66.4% (2.2%)	40.5%	95.2%	24.2%	2.0%	55.1% (1.1%)	55.6%	34.0%	97.9%	4.1%	0.0%
SLR	63.5% (4.1%)	34.3%	97.4%	5.4%	0.1%	51.7% (2.5%)	53.3%	33.0%	98.9%	0.0%	0.0%
RLR	67.1% (2.4%)	42.4%	94.0%	30.6%	2.6%	53.8% (0.5%)	53.8%	33.6%	92.6%	8.2%	0.0%
RVM	65.9% (2.3%)	40.2%	94.2%	26.5%	0.0%	55.9% (0.8%)	55.6%	35.3%	93.6%	12.2%	0.0%

Table 3.12. The table shows the cross-validation and the hold-out test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ADHD-200 data for the three-way classification problem between healthy controls, ADHD-I and ADHD-C. The training/validation and the hold-out test data are from 7 imaging sites. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 58% for Boosted Trees while the best balanced hold- out test accuracy obtained was 38.7% for RBF-SVM implemented within RCE framework.

Feature importance: Since we did not perform multiple splits on the ADHD-200 dataset, we did not plot the feature importance scores for the splits as a scatter plot, as was done with other datasets in this study. After calculating the combined feature importance scores (CFIS) for all classifiers implemented within the RCE framework, we used the combined feature importance scores to identify the top connectivity paths whose means were significantly different between the groups ($p<0.05$, corrected) as well as have high CFIS. These paths are shown in Figure 3.17 and the top 20 regions in the brain whose connectivity paths were altered in the disease are shown in Table 3.13.



Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(-32.82,53.61,6.90)	Middle frontal gyrus	MFG	(-9.51,44.53,10.07)	Anterior cingulate cortex	ACC	-1.13
(-63.84,-23.81,-16.11)	Middle temporal gyrus	MTG	(-19.83,9.20,-25.15)	Orbitofrontal cortex	OFC	-1.53
(0.02,35.07,-9.84)	Medial orbitofrontal cortex	mOFC	(-19.83,9.20,-25.15)	Orbitofrontal cortex	OFC	-1.29
(43.93,-42.84,-24.22)	Fusiform gyrus	FG	(-31.43,39.14,28.32)	Middle frontal gyrus	MFG	1.33
(39.56,8.58,29.26)	Inferior frontal gyrus	IFG	(-63.84,-23.81,-16.11)	Middle temporal gyrus	MTG	1.39
(-40.09,-32.63,16.86)	Superior temporal gyrus	STG	(-61.32,-55.35,9.65)	Middle temporal gyrus	MTG	-1.02
(3.42,18.86,-25.53)	Rectus gyrus	RG	(50.31,35.96,-11.84)	Orbitofrontal cortex	OFC	-1.09
(27.43,61.62,8.53)	Superior frontal gyrus	SFG	(0.02,35.07,-9.84)	Medial orbitofrontal cortex	mOFC	-1.31
(-58.00,-39.32,-5.05)	Middle temporal gyrus	MTG	(-19.83,9.20,-25.15)	Orbitofrontal cortex	OFC	-1.38
(13.26,44.25,6.47)	Anterior cingulate cortex	ACC	(-40.09,-32.63,16.86)	Superior temporal gyrus	STG	1.00
(14.90,-35.03,48.29)	Paracentral lobule	PCL	(49.15,35.12,14.47)	Inferior frontal gyrus	IFG	1.01
(15.34,47.41,42.71)	Superior frontal gyrus	SFG	(-20.96,-26.24,-14.20)	Hippocampus	HC	1.12

Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(15.97,-6.95,19.68)	Caudate	Caud.	(-46.14,36.88,-12.56)	Orbitofrontal cortex	OFC	1.15
(15.97,-6.95,19.68)	Caudate	Caud.	(33.15,56.72,-14.39)	Medial orbitofrontal cortex	mOFC	1.02
(-1.59,-42.74,-22.95)	Cerebellum	Cbl.	(-55.53,0.62,1.72)	Superior temporal gyrus	STG	0.90
(43.93,-42.84,-24.22)	Fusiform gyrus	FG	(-0.97,25.10,40.84)	Middle frontal gyrus	MFG	-0.85
(43.93,-42.84,-24.22)	Fusiform gyrus	FG	(-31.43,39.14,28.32)	Middle frontal gyrus	MFG	1.23
(-9.60,-100.93,12.25)	Superior occipital gyrus	SOG	(-0.48,0.29,-12.76)	Olfactory	Olf.	-0.91
(28.48,9.58,-25.46)	Superior temporal gyrus	STG	(1.91,5.81,35.25)	Cingulate gyrus	CG	-0.91
(13.26,44.25,6.47)	Anterior cingulate cortex	ACC	(-40.09,-32.63,16.86)	Superior temporal gyrus	STG	0.97
(15.34,47.41,42.71)	Superior frontal gyrus	SFG	(-20.96,-26.24,-14.20)	Hippocampus	HC	-0.90

Figure 3.17. The figure illustrates the connectivity paths which have significantly different means between the groups ($p<0.05$, corrected for multiple comparisons using permutation test) as well as are among the top hundred most discriminative paths in ADHD for (A) binary classification between controls and ADHD. (B) 3-way classification between healthy controls, ADHD-I and ADHD-C. The size of the nodes indicates the relative importance of the region (Table 3.13). Common nodes between binary and multiclass classification are indicated by yellow while other nodes are indicated by green. The sign of the paths indicates over-connectivity (positive) or under-connectivity (negative) in healthy controls compared to clinical populations. So, red represents a higher connectivity between controls compared to the diseased

populations and blue represents a lower connectivity. The numerical values in the color bar denote the combined feature importance score of the path (CFIS) obtained from classification. A higher absolute number indicates more discriminative ability for the functional connectivity path. The table below the figure tabulates the brain regions involved in the paths visualized above along with the abbreviations of the two regions and the CFIS (combined feature importance score) for the connectivity paths.

A

Rank	X	Y	Z	Brain Regions	Abbr.
1	36	20	1	Insula	Ins.
2	-10	45	10	Anterior cingulate cortex	ACC
3	-64	-24	-16	Middle temporal gyrus	MTG
4	50	36	-12	Orbitofrontal cortex	OFC
5	-20	9	-25	Orbitofrontal cortex	OFC
6	0	35	-10	Medial orbitofrontal cortex	mOFC
7	-13	46	42	Superior frontal gyrus	SFG
8	42	50	4	Middle frontal gyrus	MFG
9	27	62	9	Superior frontal gyrus	SFG
10	15	14	2	Caudate	Caud.
11	-33	54	7	Middle frontal gyrus	MFG
12	-16	-59	1	Lingual gyrus	LG
13	44	-43	-24	Fusiform gyrus	FG
14	-25	-5	-8	Putamen	Put.
15	0	-18	33	Cingulate gyrus	CG
16	15	47	43	Superior frontal gyrus	SFG
17	-31	39	28	Middle frontal gyrus	MFG
18	13	44	6	Anterior cingulate cortex	ACC
19	42	-7	-14	Insula	Ins.
20	-11	12	4	Caudate	Caud.

B

Rank	X	Y	Z	Brain Regions	Abbr.
1	-13	46	42	Superior frontal gyrus	SFG
2	-10	45	10	Anterior cingulate cortex	ACC
3	0	-18	33	Cingulate gyrus	CG
4	-16	58	26	Superior frontal gyrus	SFG
5	15	47	43	Superior frontal gyrus	SFG
6	-45	23	36	Middle frontal gyrus	MFG
7	42	50	4	Middle frontal gyrus	MFG
8	-16	-59	1	Lingual gyrus	LG
9	0	35	-10	Medial orbitofrontal cortex	mOFC
10	-3	-25	-38	Pons	Pons
11	-33	54	7	Middle frontal gyrus	MFG
12	-31	39	28	Middle frontal gyrus	MFG
13	-1	25	41	Middle frontal gyrus	MFG
14	36	20	1	Insula	Ins.
15	13	44	6	Anterior cingulate cortex	ACC
16	-64	-24	-16	Middle temporal gyrus	MTG
17	-13	-80	-49	Cerebellum	Cbl.
18	10	60	25	Middle frontal gyrus	MFG
19	55	11	-6	Superior temporal gyrus	STG
20	15	14	2	Caudate	Caud.

Table 3.13. This table lists the top 20 regions for ADHD as identified by (A) Binary classification between Healthy Controls and ADHD (B) Multiclass classification between Controls, ADHD-I, and ADHD-C using ADHD-200 data.

3.3.4 PTSD

Classification results: The classification results for the binary classification scenario between healthy Soldiers and Soldiers diagnosed with PTSD, for the age split as well as the age-matched data are shown in Figure 3.18 and Figure 3.19, respectively. The corresponding tables listing the detailed individual class accuracies are shown in Table 3.14 and Table 3.15, respectively. For the multiclass classification scenario between healthy Soldiers, Soldiers diagnosed with just PTSD and those with both PCS and PTSD, the results are shown in Figure 3.20 and Figure 3.21 with detailed accuracy performance presented in Table 3.16 and Table 3.17, respectively. The results

indicate a large difference in accuracy between the biased cross-validation accuracy and the hold-out test accuracy across all the classifiers. As with ABIDE dataset, some of this difference can be attributed to age variability due to t-test filtering performed on just the training/validation data. Most classifiers are extremely unreliable in the case where the training/validation and the hold-out test data were from different age groups. This indicates that the classifiers were overfitting the data in the training/validation dataset and were not learning the connectivity modulations effected by PTSD. The classification performance on the training/validation data did not translate to good performance on the hold-out test data for a majority of the classifiers (although a couple of them did perform well with hold-out test data), a problem that might plague single site classification studies with subjects belonging to narrow age ranges.

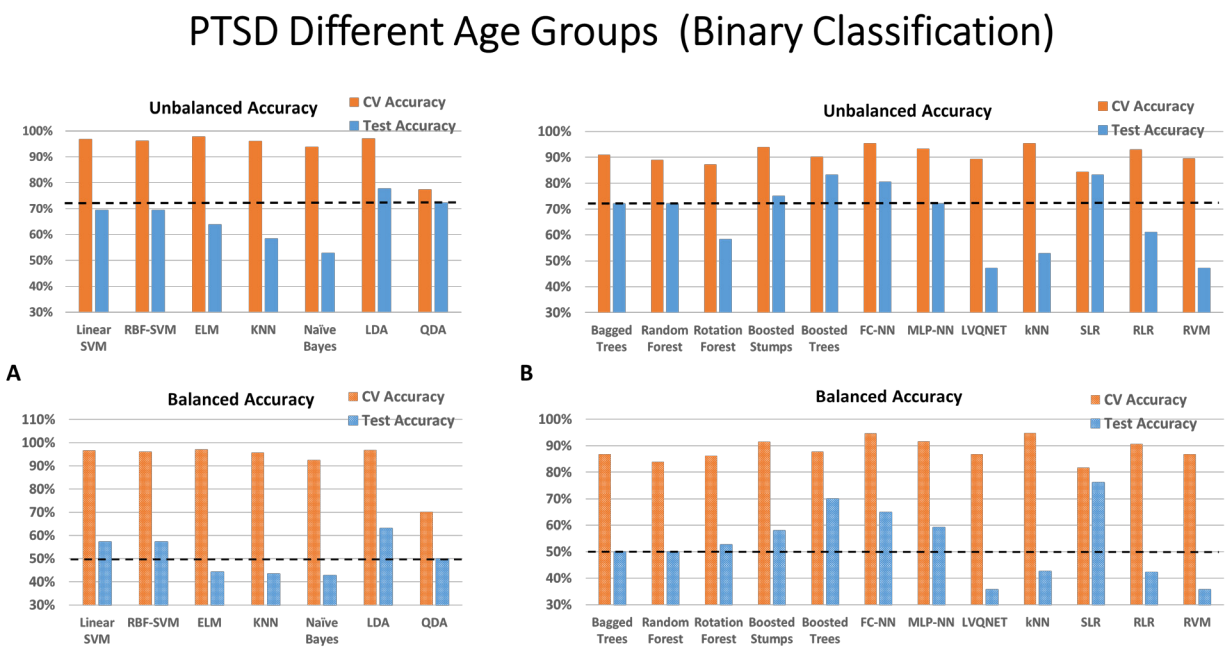


Figure 3.18. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for PTSD data when the training/validation data and the hold-out test data are from different age groups in the range for the multiclass classification between healthy controls and subjects with PTSD. The training/validation data is from an age range of 23-37 years while the data from the age range of 38-53 years was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The

orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 72.2% since subjects with PTSD formed 72.2% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 83.3% while the best balanced hold-out test accuracy obtained was 76.2% for Sparse Logistic Regression.

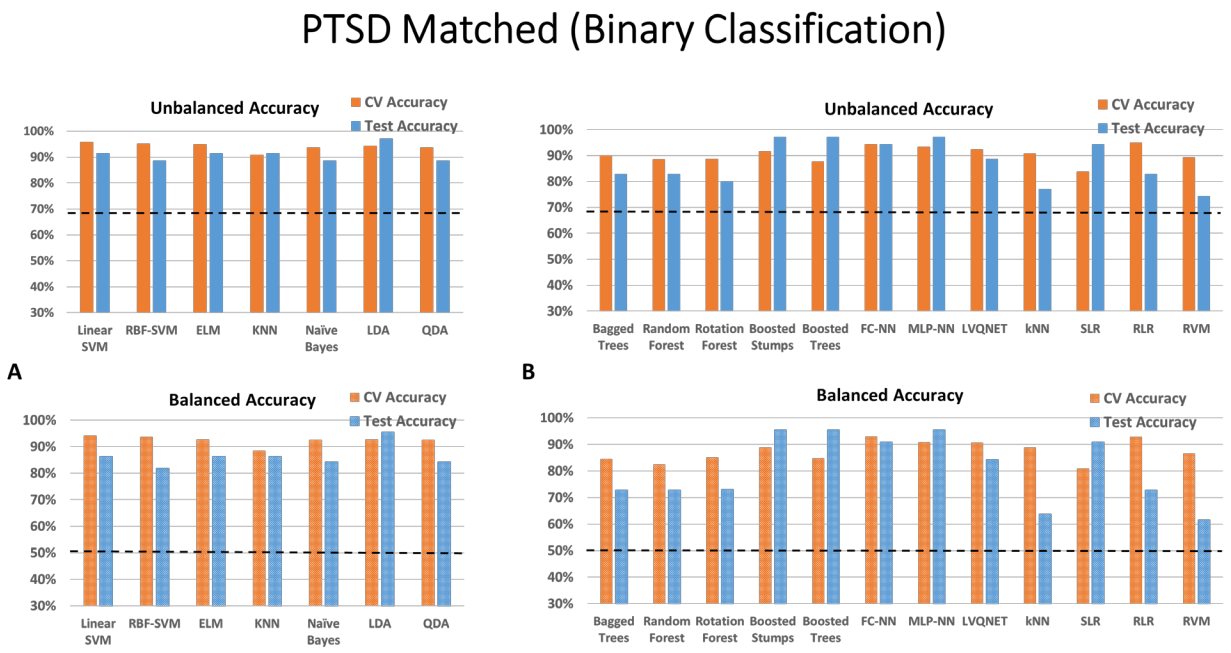


Figure 3.19. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for PTSD data when the training/validation data and the hold-out test data are from same age groups in the range for the multiclass classification between healthy controls and subjects with PTSD. The training/validation data and the hold-out test data are matched in age with subjects from age range of 23-53 years. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 68.6% since subjects with PTSD formed 68.6% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The best hold-out test accuracy was 97.1%, whereas the best balanced hold-out test accuracy obtained was 95.5%, obtained for Boosted Stumps, MLP-NN and LDA implemented within the RCE framework.

PTSD Different Age Groups (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	PTSD	w/o Voting	Voting	Balanced	Controls	PTSD
Linear SVM	96.8% (4.1%)	96.7%	96.2%	97.1%	66.2% (7.3%)	69.4%	57.3%	30%	84.6%
RBF-SVM	96.2% (4.4%)	96.1%	95.8%	96.4%	66.4% (7.1%)	69.4%	57.3%	30%	84.6%
ELM	97.8% (3.6%)	97.1%	95.2%	99%	61% (6.8%)	63.9%	44.3%	0%	88.5%
KNN	96.1% (4.9%)	95.7%	94.3%	97%	56.2% (8.8%)	58.3%	43.5%	10%	76.9%
Naïve Bayes	93.8% (4.9%)	92.6%	88.8%	96.3%	60.2%(7.2%)	52.8%	42.7%	20%	65.4%
LDA	97.1% (4.5%)	96.8%	96%	97.6%	69.4% (8.2%)	77.8%	63.1%	30%	96.2%
QDA	77.4% (8%)	70.2%	48.6%	91.8%	69.1% (8.1%)	72.2%	50%	0%	100%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	PTSD	w/o Voting	Voting	Balanced	Controls	PTSD
Bagged Trees	90.9% (5.7%)	86.8%	74.7%	99%	72.6%(2.9%)	72.2%	50%	0%	100%
Random Forest	88.9% (6%)	83.9%	68.9%	98.9%	71.8% (3%)	72.2%	50%	0%	100%
Rotation Forest	87.2% (6.8%)	86.2%	83%	89.3%	57.6% (7.5%)	58.3%	52.7%	40%	65.4%
Boosted Stumps	93.8% (5.1%)	91.6%	85.1%	98.1%	75.6% (5%)	75%	58.1%	20%	96.2%
Boosted Trees	90.2% (6.3%)	87.8%	80.4%	95.1%	72.9% (6.9%)	83.3%	70%	40%	100%
FC-NN	95.3% (4.8%)	94.6%	92.8%	96.5%	71.5% (5.8%)	80.5%	65%	30%	100%
MLP-NN	93.2% (8.3%)	91.7%	87.0%	96.4%	70% (7.3%)	72.2%	59.2%	30%	88.5%
LVQNET	89.3% (6.1%)	86.8%	79.2%	94.4%	50.1% (5.8%)	47.2%	35.8%	10%	61.5%
KNN	95.3% (5.1%)	94.7%	92.7%	96.6%	53.6% (5%)	52.8%	42.7%	20%	65.4%
SLR	84.4% (7.3%)	81.7%	73.7%	89.7%	77.3% (5.7%)	83.3%	76.2%	60%	92.3%
RLR	93% (4.9%)	90.7%	83.6%	97.8%	58.3% (2.8%)	61.1%	42.3%	0%	84.6%
RVM	89.5% (6%)	86.7%	78.3%	95.1%	53.4% (5.6%)	47.2%	35.8%	10%	61.5%

Table 3.14. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the PTSD data we collected, for the binary classification problem between healthy controls and subjects with PTSD. The training/validation data is from an age range of 23-37 years while the data from the age range of 38-53 years was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE

framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 83.3% while the best balanced hold-out test accuracy obtained was 76.2% for Sparse Logistic Regression.

PTSD Matched (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	PTSD	w/o Voting	Voting	Balanced	Controls	PTSD
Linear SVM	95.8% (4%)	94.1%	89.2%	99.0%	88.7% (6.2%)	91.4%	86.4%	72.7%	100%
RBF-SVM	95.2% (4.2%)	93.6%	88.7%	98.4%	85.1% (6.7%)	88.6%	81.8%	63.6%	100%
ELM	94.9% (4.4%)	92.8%	86.9%	98.8%	88.5% (4.8%)	91.4%	86.4%	72.7%	100%
KNN	90.8% (6%)	88.5%	82%	95%	83.2% (8.3%)	91.4%	86.4%	72.7%	100%
Naïve Bayes	93.7% (5.1%)	92.7%	89.7%	95.6%	84.4% (4.6%)	88.6%	84.3%	72.7%	95.8%
LDA	94.3% (4.7%)	92.8%	88.6%	97%	89.4% (5.8%)	97.1%	95.5%	90.9%	100%
QDA	93.6% (4.6%)	92.7%	89.8%	95.5%	84.3% (4.6%)	88.6%	84.3%	72.7%	95.8%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	PTSD	w/o Voting	Voting	Balanced	Controls	PTSD
Bagged Trees	89.9% (5.7%)	84.6%	69.4%	99.7%	80.9% (2.6%)	82.9%	72.8%	45.5%	100%
Random Forest	88.5% (5.9%)	82.4%	65.2%	99.6%	79.3% (2.7%)	82.9%	72.8%	45.5%	100%
Rotation Forest	88.7% (6.2%)	85.1%	74.9%	95.3%	78% (4.8%)	80%	73.1%	55%	91.7%
Boosted Stumps	91.6% (5.4%)	88.9%	81.3%	96.5%	93.7% (3.5%)	97.1%	95.5%	90.9%	100%
Boosted Trees	87.6% (7.2%)	84.8%	76.8%	92.8%	88.6% (5.2%)	97.1%	95.5%	90.9%	100%
FC-NN	94.3% (4.6%)	92.9%	88.6%	97.1%	89.2% (4.2%)	94.3%	90.9%	81.8%	100%
MLP-NN	93.3% (7.8%)	90.8%	83.7%	97.9%	88.7% (6.8%)	97.1%	95.5%	90.9%	100%
LVQNET	92.3% (5.1%)	90.7%	86.1%	95.2%	82.8% (3.8%)	89%	84.3%	72.7%	95.8%
KNN	90.7% (5.7%)	88.8%	83.2%	94.3%	76.9% (6.5%)	77.1%	63.7%	27.3%	100%
SLR	83.7% (8.3%)	80.9%	73.1%	88.7%	85.7% (6.4%)	94.3%	90.9%	81.8%	100%
RLR	95% (4%)	92.8%	86.4%	99.1%	81.8% (2.6%)	82.9%	72.8%	45.5%	100%
RVM	89.3% (5.9%)	86.5%	78.6%	94.4%	76.1% (3.3%)	74.3%	61.6%	27.3%	95.8%

Table 3.15. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the PTSD data we collected, for the binary classification problem between healthy controls and subjects with PTSD. The training/validation data and the hold-out test data are matched in age with subjects from age range of 23-53 years. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates

the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 97.1%, whereas the best balanced hold-out test accuracy obtained was 95.5%, obtained for Boosted Stumps, MLP-NN and LDA implemented within the RCE framework.

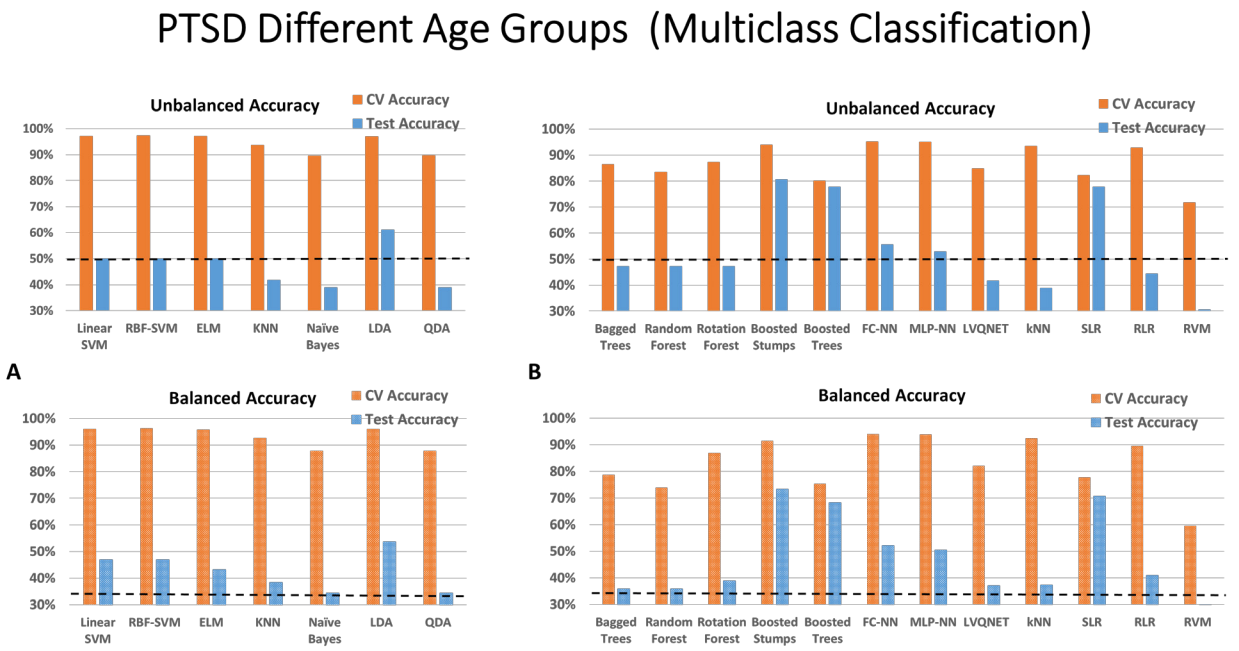


Figure 3.20. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for PTSD data when the training/validation data and the hold-out test data are from different age groups in the range for the multiclass classification between healthy controls and subjects with just PTSD and those with both PCS and PTSD. The training/validation data is from an age range of 23-37 years while the data from the age range of 38-53 years was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 50% since subjects with PCS+PTSD formed 50% of the total size of the hold-out test data. For balanced accuracy, this is 33.3%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 80.6% while the best balanced hold-out test accuracy obtained was 73.3% for Boosted Stumps.

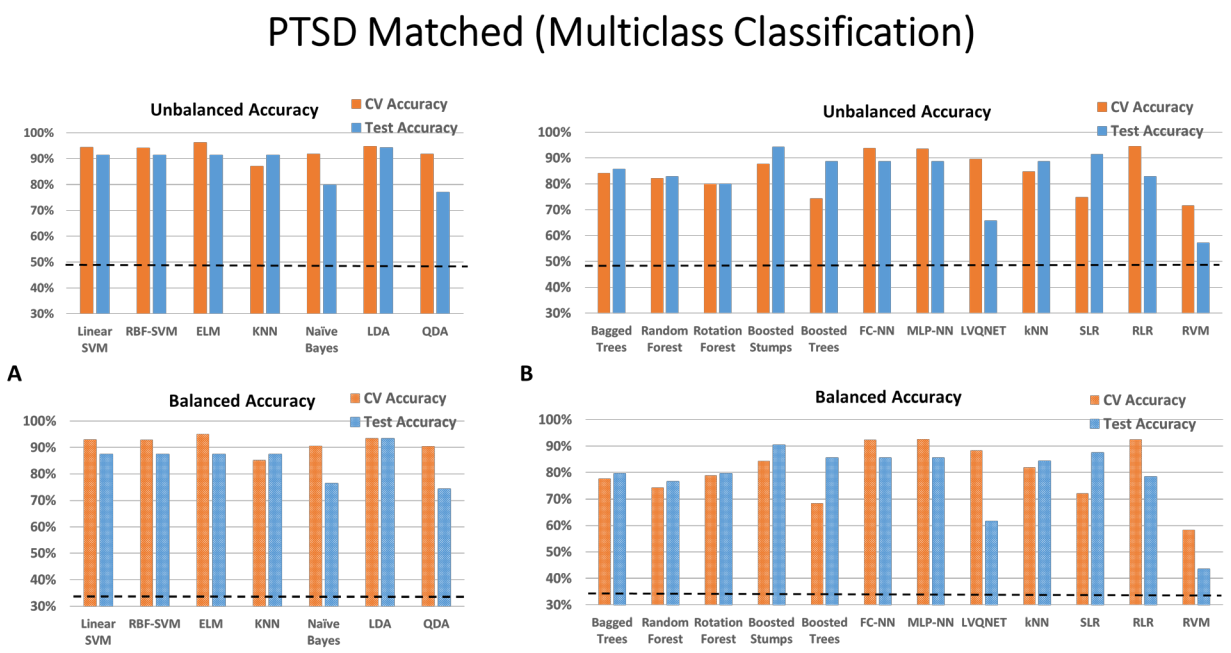


Figure 3.21. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for PTSD data when the training/validation data and the hold-out test data are from same age groups in the range for the multiclass classification between healthy controls and subjects with just PTSD and those with both PCS and PTSD. The training/validation data and the hold-out test data are matched in age with subjects from age ranges of 23-53 years. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 48.6% since subjects with PCS+PTSD formed 48.6% of the total size of the hold-out test data. For balanced accuracy, this is 33.3%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The best hold-out test accuracy was 94.3% for Boosted Stumps, and LDA implemented within RCE framework, while the best balanced hold-out test accuracy obtained was 93.3% for LDA implemented within RCE framework.

PTSD Different Age Groups (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	PTSD	PCS+PTSD	w/o Voting	Voting	Balanced	Controls	PTSD	PCS+PTSD
Linear SVM	97.1% (4.1%)	96.1%	97.9%	91.5%	98.8%	45.4% (6.5%)	50%	47.0%	30%	50%	61.1%
RBF-SVM	97.3% (3.8%)	96.3%	98.2%	91.8%	98.8%	47.2% (6.1%)	50%	47.0%	30%	50%	61.1%
ELM	97% (4.1%)	95.8%	98.2%	90.5%	98.7%	43.5% (8.4%)	50%	43.2%	20%	37.5%	72.2%
KNN	93.7% (6%)	92.7%	96.3%	87.6%	94.3%	39.7% (9.3%)	41.7%	38.3%	40%	25%	50%
Naïve Bayes	89.6% (6.9%)	87.8%	91.5%	79.5%	92.3%	42% (6.1%)	38.9%	34.4%	10%	37.5%	55.6%
LDA	96.9% (4.3%)	96.1%	95.3%	94.0%	99.1%	49.6% (8.6%)	61.1%	53.6%	40%	37.5%	83.3%
QDA	89.8% (6.5%)	87.9%	91.8%	79.8%	92.2%	41.5% (6%)	38.9%	34.4%	10%	37.5%	55.6%

B

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	PTSD	PCS+PTSD	w/o Voting	Voting	Balanced	Controls	PTSD	PCS+PTSD
Bagged Trees	86.4% (6.4%)	78.8%	91.1%	46.6%	98.8%	47.6% (4.9%)	47.2%	35.9%	30%	0%	77.8%
Random Forest	83.5% (6.4%)	74%	89.5%	33.6%	99%	46.5% (4.7%)	47.2%	35.9%	30%	0%	77.8%
Rotation Forest	87.3% (7.5%)	86.9%	89.7%	84.2%	86.8%	46% (2.7%)	47.2%	38.9%	50%	0%	66.7%
Boosted Stumps	93.9% (5.6%)	91.4%	95.3%	80.5%	98.3%	65.3% (7.3%)	80.6%	73.3%	70%	50%	100%
Boosted Trees	80.1% (8.1%)	75.3%	78.2%	57.5%	90.2%	61.6% (7.8%)	77.8%	68.3%	80%	25%	100%
FC-NN	95.2% (5.1%)	94%	94.5%	89.5%	98%	51.1% (6.2%)	55.6%	52.2%	40%	50%	66.7%
MLP-NN	95.1% (5.1%)	93.9%	96.3%	88.5%	96.8%	49.3% (6%)	52.8%	50.4%	40%	50%	61.1%
LVQNET	84.8% (7.1%)	82.1%	83.9%	72%	90.5%	39.5% (4.9%)	41.7%	37%	0%	50%	61.1%
KNN	93.5% (6.4%)	92.4%	96%	87.1%	94.2%	41.1% (5.1%)	38.9%	37.3%	30%	37.5%	44.4%
SLR	82.2% (7.7%)	77.8%	81.4%	60.8%	91.2%	57% (8.6%)	77.8%	70.6%	80%	37.5%	94.4%
RLR	92.8% (5.6%)	89.6%	95.2%	76%	97.7%	42.6% (3.6%)	44.4%	41%	30%	37.5%	55.6%
RVM	71.8% (7.1%)	59.5%	83.9%	5.1%	89.5%	33.5% (4.8%)	30.6%	24.8%	30%	0%	44.4%

Table 3.16. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the PTSD data we collected, for the 3-way classification problem between healthy controls and subjects with PTSD and subjects who experienced both PCS and diagnosed with PTSD. The training/validation data is from an age range of 23-37 years while the data from the age range of 38-53 years was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 80.6% while the best balanced hold-out test accuracy obtained was 73.3% for Boosted Stumps.

PTSD Matched (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	PTSD	PCS+PTSD	w/o Voting	Voting	Balanced	Controls	PTSD	PCS+PTSD
Linear SVM	94.5% (4.6%)	92.9%	93.5%	87.3%	98%	84.2% (4.9%)	91.4%	87.4%	90.9%	71.4%	100%
RBF-SVM	94.2% (5%)	92.8%	93.2%	87.6%	97.5%	82.8% (5.6%)	91.4%	87.4%	90.9%	71.4%	100%
ELM	96.2% (3.8%)	95.1%	94.6%	91.2%	99.4%	85.4% (4.4%)	91.4%	87.4%	90.9%	71.4%	100%
KNN	87.1% (7.6%)	85.2%	88.5%	76.9%	90.1%	83% (5.6%)	91.4%	87.4%	90.9%	71.4%	100%
Naïve Bayes	91.8% (5.6%)	90.6%	92.4%	85.3%	94.2%	75.5% (4.9%)	80%	76.4%	63.6%	71.4%	94.1%
LDA	94.8% (4.4%)	93.5%	94.4%	88.3%	97.8%	86.8% (5.4%)	94.3%	93.3%	100%	85.7%	94.1%
QDA	91.8% (5.5%)	90.5%	92.3%	85.3%	94%	75.4% (5.2%)	77.1%	74.4%	63.6%	71.4%	88.2%

B

Classifiers	Cross-validation Accuracy					Hold-out Test Accuracy					
	Unbalanced	Balanced	Controls	PTSD	PCS+PTSD	w/o Voting	Voting	Balanced	Controls	PTSD	PCS+PTSD
Bagged Trees	84.1% (7%)	77.6%	81.5%	52.7%	98.5%	82% (4.8%)	85.7%	79.6%	81.8%	57.1%	100%
Random Forest	82.1% (6.7%)	74.3%	80.1%	43.7%	99%	76.8% (4.9%)	82.9%	76.6%	72.7%	57.1%	100%
Rotation Forest	79.9% (8.6%)	78.9%	79.3%	75.2%	82.2%	73.2% (5.6%)	80%	79.6%	90.9%	71.4%	76.5%
Boosted Stumps	87.7% (6.6%)	84.4%	88.1%	71%	94.1%	87.6% (4%)	94.3%	90.5%	100%	71.4%	100%
Boosted Trees	74.3% (8.9%)	68.4%	72%	46.1%	87.1%	74.2% (6.7%)	88.6%	85.5%	90.9%	71.4%	94.1%
FC-NN	93.8% (4.7%)	92.3%	93.2%	86.7%	97.1%	83% (5.5%)	88.6%	85.5%	90.9%	71.4%	94.1%
MLP-NN	93.6% (5.1%)	92.5%	92.9%	88.3%	96.3%	82.7% (5.8%)	88.6%	85.5%	90.9%	71.4%	94.1%
LVQNET	89.6% (6.2%)	88.3%	87.4%	84.4%	93.2%	65.7% (4.1%)	65.7%	61.7%	45.5%	57.1%	82.4%
KNN	84.8% (7.6%)	82.0%	87.6%	69.6%	88.9%	79.1% (9.5%)	88.6%	84.4%	81.8%	71.4%	100%
SLR	74.8% (9.1%)	72.1%	74.9%	61%	80.3%	79.8% (6.9%)	91.4%	87.4%	90.9%	71.4%	100%
RLR	94.6% (4.6%)	92.4%	93%	84.5%	99.7%	80.1% (3.5%)	82.9%	78.3%	63.6%	71.4%	100%
RVM	71.6% (5.7%)	58.3%	80.4%	0%	94.6%	57.8% (4.3%)	57.1%	43.5%	36.4%	0%	94.1%

Table 3.17. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the PTSD data we collected, for the 3-way classification problem between healthy controls and subjects with PTSD and subjects who experienced both PCS and diagnosed with PTSD. The training/validation data and the hold-out test data are matched in age with subjects from age range of 23-53 years. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best holdout test accuracy was 94.3% for Boosted Stumps, and LDA implemented within RCE framework, while the best-balanced hold-out test accuracy obtained was 93.3% for LDA implemented within RCE framework.

In the binary classification scenario for the split in which the training/validation and the hold-out test data belonged to different age ranges, the best hold-out test accuracy was 83.3% while the best balanced hold-out test accuracy obtained was 76.2% for Sparse Logistic Regression. In fact, SLR along with Boosted Trees were the only two classifiers which gave good performances followed by FCC-NN and MLP-NN. In the multiclass classification for the same split, the best hold-out test accuracy was 80.6% while the best balanced hold-out test accuracy obtained was 73.3% for Boosted Stumps. In the split wherein the training/validation and the hold-out test data were matched for age, the binary classification results were higher compared to the unmatched case with accuracies close to 90% on the hold-out test data. The best hold-out test accuracy for the binary case was 97.1%, whereas the best balanced hold-out test accuracy obtained was 95.5%, using Boosted Stumps, MLP-NN and LDA implemented within the RCE framework. For the 3-way classification, the best hold-out test accuracy was 94.3% for Boosted Stumps, and LDA implemented within RCE framework, while the best balanced hold-out test accuracy obtained was 93.3% for LDA implemented within RCE framework.

Feature importance: The combined feature importance scores (CFIS) for the two splits were plotted in a scatter plot as shown in Figure 3.22 for the binary and multiclass scenarios. The plot illustrates variability and the negative slope particularly in the multiclass classification case which can be attributed to the age ranges in each split. This means that the CFIS for multiclass classification which are higher in one split were lower in the other. So age has a significant impact in altering the feature importance. For binary classification, however, the slope was still positive. Using the CFIS we identified the top connectivity paths (shown in Figure 3.23) whose means were significantly different between the groups ($p<0.05$, corrected) as well as have high

combined feature importance scores. The top 20 regions in the brain whose connectivity paths were altered in the disease are listed in the Table 3.18.

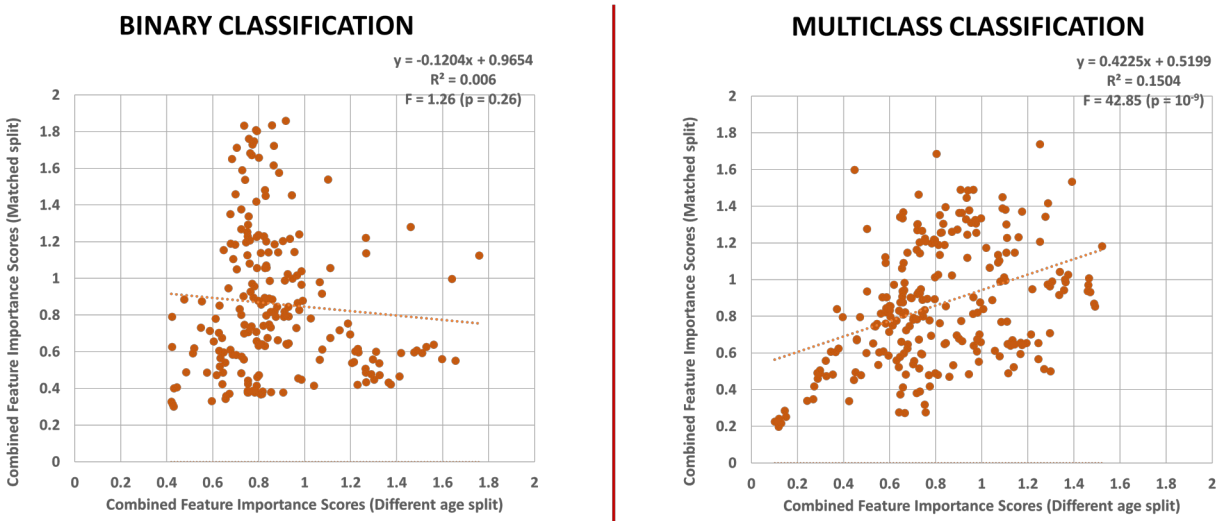
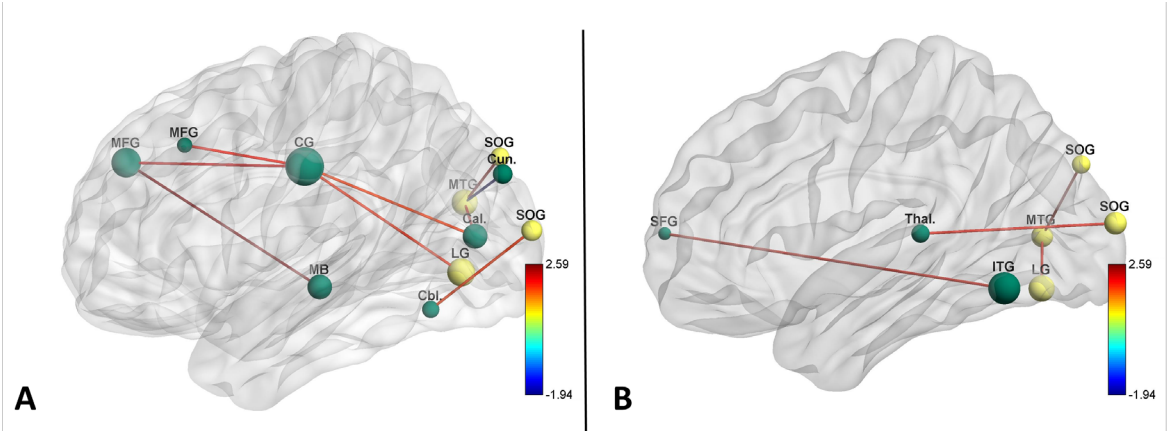


Figure 3.22. Scatter plots of combined feature importance scores (CFIS) for the two splits performed on the PTSD data. We selected the common features in obtained and plotted them as scatter plots for both binary and multiclass classification scenarios. The plot illustrates a lot of variability in the CFIS, particularly in the binary classification scenario which can be attributed to the age ranges of the training/validation data in each split. There is however significant agreement in the CFIS across the age split for the multiclass classification scenario.



Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(0.05,-23.79,-11.82)	Midbrain	MB	(32.92,47.43,25.66)	Middle frontal gyrus	MFG	2.60
(-9.60,-100.93,12.25)	Superior occipital gyrus	SOG	(0.42,-63.99,-19.92)	Cerebellum	Cbl.	1.80
(0.35,-18.33,32.96)	Cingulate gyrus	CG	(3.01,-79.95,6.81)	Calcarine	Cal.	1.83
(0.35,-18.33,32.96)	Cingulate gyrus	CG	(-0.97,25.10,40.84)	Middle frontal gyrus	MFG	1.98
(21.29,-74.31,-11.23)	Lingual gyrus	LG	(0.35,-18.33,32.96)	Cingulate gyrus	CG	1.85
(50.35,-74.53,7.56)	Middle temporal gyrus	MTG	(-9.42,-90.39,33.16)	Cuneus	Cun.	-1.95
(50.35,-74.53,7.56)	Middle temporal gyrus	MTG	(3.01,-79.95,6.81)	Calcarine	Cal.	1.98
(15.33,-88.55,33.41)	Superior occipital gyrus	SOG	(50.35,-74.53,7.56)	Middle temporal gyrus	MTG	2.10
(32.92,47.43,25.66)	Middle frontal gyrus	MFG	(0.35,-18.33,32.96)	Cingulate gyrus	CG	1.99

Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(-9.60,-100.93,12.25)	Superior occipital gyrus	SOG	(-31.00,-3.00,-33.00)	Thalamus	Thal.	2.30
(27.43,61.62,8.53)	Superior frontal gyrus	SFG	(-52.85,-60.85,-11.17)	Inferior temporal gyrus	ITG	2.46
(50.35,-74.53,7.56)	Middle temporal gyrus	MTG	(21.29,-74.31,-11.23)	Lingual gyrus	LG	1.99
(15.33,-88.55,33.41)	Superior occipital gyrus	SOG	(50.35,-74.53,7.56)	Middle temporal gyrus	MTG	2.01

Figure 3.23. The figure illustrates the connectivity paths which have significantly different means between the groups ($p<0.05$, corrected for multiple comparisons using permutation test) as well as are among the top hundred most discriminative paths for PCS and PTSD for (A) binary classification between combat controls and PTSD. (B) 3-way classification between healthy combat controls, PTSD and PCS+PTSD. The size of the nodes indicates the relative importance of the region (Table 3.18). Common nodes between binary and multiclass classification are indicated by yellow while other nodes are indicated by green. The sign of the paths indicates over-connectivity (positive) or under-connectivity (negative) in healthy controls compared to clinical populations. So, red represents a higher connectivity between controls compared to the diseased populations and blue represents a lower connectivity. The numerical values in the color bar denote the combined feature importance score of the path (CFIS) obtained from classification. A higher absolute number indicates more discriminative ability for the functional connectivity path. The table below the figure tabulates the brain regions involved in the paths visualized above along with the abbreviations of the two regions and the CFIS (combined feature importance score) for the connectivity paths.

A

Rank	X	Y	Z	Brain Regions	Abbr.
1	0	-18	33	Cingulate gyrus	CG
2	33	47	26	Middle frontal gyrus	MFG
3	21	-74	-11	Lingual gyrus	LG
4	-30	-87	23	Middle occipital gyrus	MOG
5	50	-75	8	Middle temporal gyrus	MTG
6	-19	-76	-9	Lingual gyrus	LG
7	3	-80	7	Calcarine	Cal.
8	51	-64	-11	Inferior temporal gyrus	ITG
9	0	-24	-12	Midbrain	MB
10	-11	12	4	Caudate	Caud.
11	39	-82	30	Middle occipital gyrus	MOG
12	35	-89	8	Middle occipital gyrus	MOG
13	-28	-96	-2	Middle occipital gyrus	MOG
14	-46	-80	4	Middle occipital gyrus	MOG
15	22	-58	0	Lingual gyrus	LG
16	-16	-59	1	Lingual gyrus	LG
17	2	6	35	Cingulate gyrus	CG
18	16	-7	20	Caudate	Caud.
19	-10	-101	12	Superior occipital gyrus	SOG
20	-57	-7	28	Postcentral gyrus	PCG

B

Rank	X	Y	Z	Brain Regions	Abbr.
1	33	47	26	Middle frontal gyrus	MFG
2	0	-18	33	Cingulate gyrus	CG
3	-53	-61	-11	Inferior temporal gyrus	ITG
4	35	-89	8	Middle occipital gyrus	MOG
5	-30	-87	23	Middle occipital gyrus	MOG
6	21	-74	-11	Lingual gyrus	LG
7	22	-58	0	Lingual gyrus	LG
8	3	-80	7	Calcarine	Cal.
9	2	6	35	Cingulate gyrus	CG
10	-19	-76	-9	Lingual gyrus	LG
11	-46	-80	4	Middle occipital gyrus	MOG
12	16	-7	20	Caudate	Caud.
13	39	-82	30	Middle occipital gyrus	MOG
14	-10	-101	12	Superior occipital gyrus	SOG
15	-21	-26	-14	Inferior temporal gyrus	ITG
16	51	-64	-11	Hippocampus	HC
17	-16	-59	1	Lingual gyrus	LG
18	50	-75	8	Middle temporal gyrus	MTG
19	0	-24	-12	Midbrain	MB
20	40	9	29	Inferior frontal gyrus	IFG

Table 3.18. This table illustrates the top 20 regions for PCS and PTSD as identified by (A) Binary classification between Healthy Controls and PTSD (B) Multiclass classification between Controls, PCS, and PTSD, using PSTD data.

3.3.5 ADNI

Classification results: The classification performance for the binary classification scenario between healthy adults and adults diagnosed with Alzheimer's disease for the age split as well as

for the age-matched data are shown in Figure 3.24 and Figure 3.25, respectively. The tables corresponding to the above results with detailed individual class accuracies are shown in Table 3.19 and Table 3.20, respectively. Results for the multiclass classification scenario (Controls, EMCI, LMCI and AD) are shown in Figure 3.26 and Figure 3.27 with detailed accuracy performance presented in Table 3.21 and Table 3.22. The biased cross-validation accuracy was very high with most classifiers, but it did not translate to high hold-out test accuracy. In fact, for many classifiers especially in the multiclass scenario, hold-out test accuracy was better than using a majority classifier, though not by much. All classifiers were generally unreliable in the cases where the training/validation and the test data were from different age group, clearly indicating a case of overfitting.

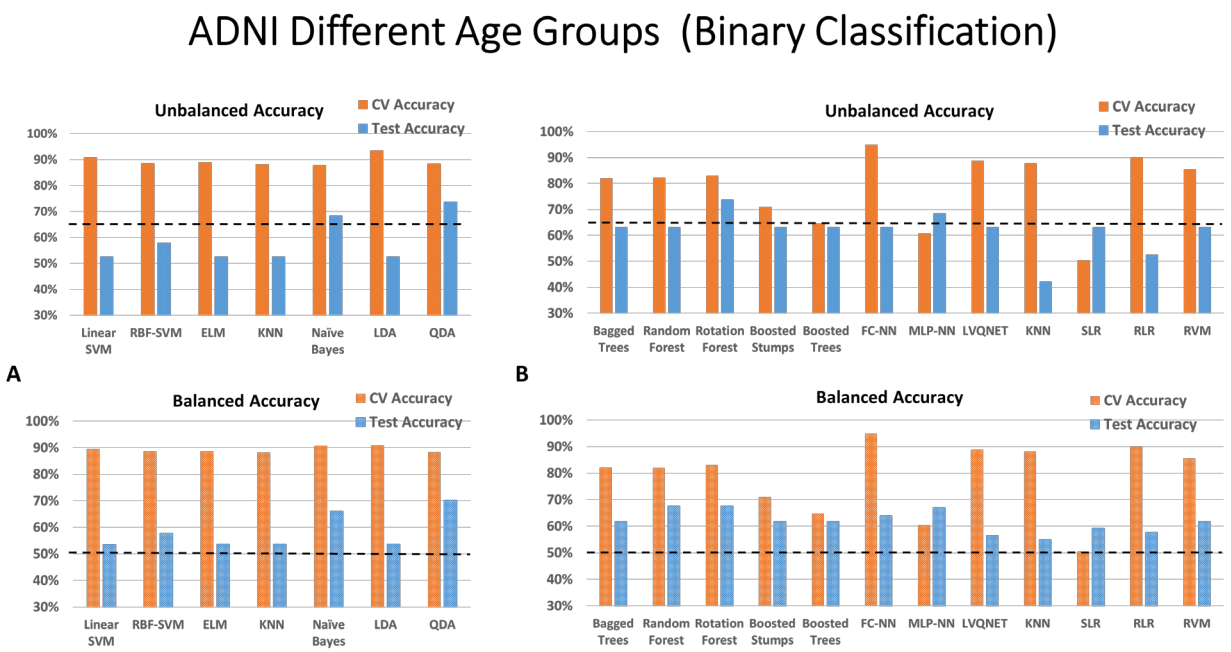


Figure 3.24. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ADNI data when the training/validation data and the hold-out test data are from different age groups in the range for the binary classification between healthy controls and subjects with Alzheimer's disease (AD). The training/validation data is from an age range of 56-76 years while the data from the age range of 77-88 years was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class

accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 63.2% since healthy controls formed 63.2% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 83.3% while the best balanced hold-out test accuracy obtained was 76.2% for Sparse Logistic Regression.

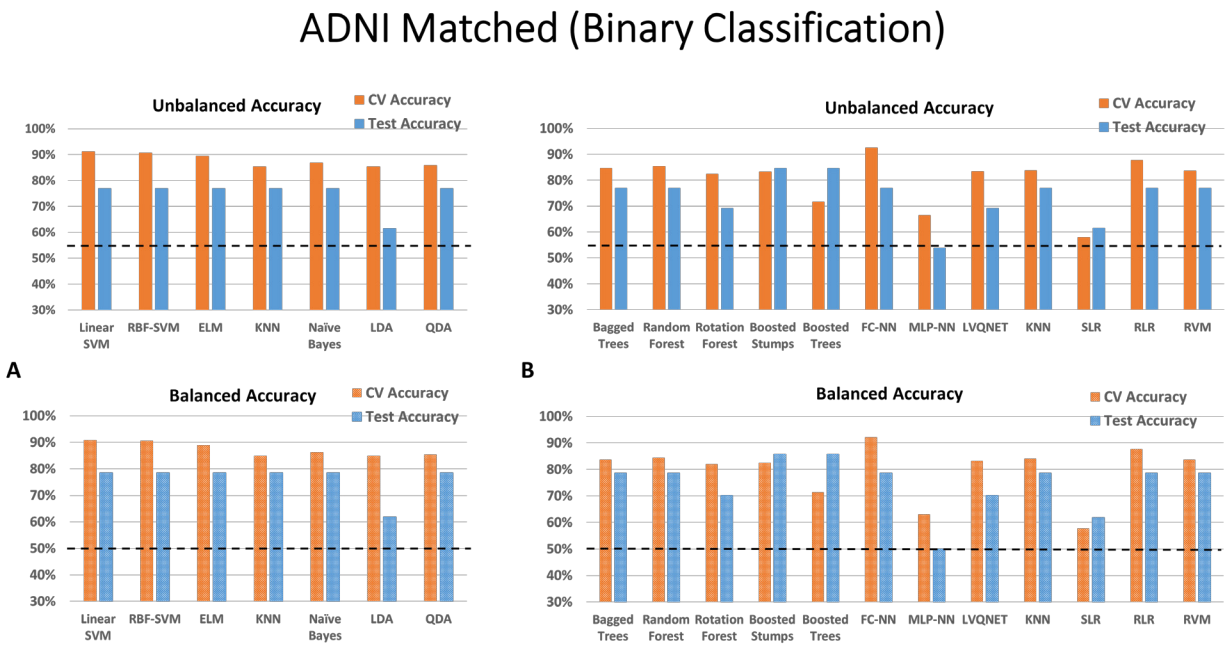


Figure 3.25. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ADNI data when the training/validation data and the hold-out test data are from the same age groups in the range for the binary classification between healthy controls and subjects with Alzheimer's disease. The training/validation data and the hold-out test data are matched in age with subjects from age range of 56-88 years. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 53.8% since healthy controls formed 53.8% of the total size of the hold-out test data. For balanced accuracy, this is exactly 50%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy

was 84.6% while the best balanced hold-out test accuracy obtained was 85.7% for Boosted Trees and Stumps.

ADNI Different Age Groups (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	AD	w/o Voting	Voting	Balanced	Controls	AD
Linear SVM	90.8% (10.4%)	89.5%	89.8%	91.5%	54.9% (9.5%)	52.6%	53.5%	50%	57.1%
RBF-SVM	88.5% (11.2%)	88.6%	87.5%	89.4%	57.5% (9.3%)	57.9%	57.7%	58.3%	57.1%
ELM	88.9% (12.1%)	88.6%	86.4%	91.2%	57.3% (8.2%)	52.6%	53.6%	50%	57.1%
KNN	88.2% (11.4%)	88.1%	82.3%	94.5%	57.2% (9.7%)	52.6%	53.6%	50%	57.1%
Naïve Bayes	87.8 %(14.2%)	90.7%	93%	82.4%	64.3% (9.9%)	68.4%	66.1%	75%	57.1%
LDA	93.4% (10.5%)	90.9%	92.5%	94.8%	57.1% (8.5%)	52.6%	53.6%	50%	57.1%
QDA	88.4% (13.8%)	88.2%	92.7%	83.6%	65% (9.9%)	73.7%	70.2%	83.3%	57.1%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	AD	w/o Voting	Voting	Balanced	Controls	AD
Bagged Trees	82% (13.4%)	82%	87%	77%	62.1% (6.4%)	63.2%	61.9%	66.7%	57.1%
Random Forest	82.2% (13.4%)	81.9%	85.5%	78.4%	63.2% (6.7%)	63.2%	67.6%	66.7%	57.1%
Rotation Forest	83% (12.6%)	83%	84.1%	81.8%	66.1% (7.4%)	73.7%	67.6%	75.0%	71.4%
Boosted Stumps	71% (13.5%)	70.9%	72.2%	69.6%	61.7% (7.1%)	63.2%	61.9%	66.7%	57.1%
Boosted Trees	64.5% (14.8%)	64.7%	61.4%	67.9%	58.5% (8.2%)	63.2%	61.9%	66.7%	57.1%
FC-NN	94.9% (8.3%)	94.9%	94.8%	95.1%	59.1% (6%)	63.2%	64%	66.7%	57.1%
MLP-NN	60.6% (18.3%)	60.4%	67.5%	53.3%	54.4% (11.3%)	68.4%	67%	75%	57.1%
LVQNET	88.7% (11.3%)	88.8%	85.2%	92.4%	59.4% (6.5%)	63.2%	56.5%	50%	85.7%
KNN	87.8% (11.4%)	88.1%	78.5%	97.6%	52.6% (7.5%)	42.1%	55%	33.3%	57.1%
SLR	50.3% (15.3%)	50.5%	51.9%	49.1%	53.4% (10.9%)	63.2%	59.2%	58.3%	71.4%
RLR	90% (12.3%)	90.1%	89%	91.2%	54.3% (2.6%)	52.6%	57.7%	50%	57.1%
RVM	85.3% (13.2%)	85.6%	89.6%	81.5%	63.3% (5.8%)	63.2%	61.9%	66.7%	57.1%

Table 3.19. The table shows cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ADNI data we collected, for the binary classification problem between healthy controls and subjects with Alzheimer's disease. The training/validation data is from an age range of 56-76 years while the data from the age range of 77-88 years was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained

by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 73.7% obtained for Random Forest, and QDA implemented within RCE framework while the best balanced hold-out test accuracy obtained was 70.2% for QDA implemented within RCE framework.

ADNI Matched (Binary Classification)

A

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	AD	w/o Voting	Voting	Balanced	Controls	AD
Linear SVM	91.2% (9.7%)	90.8%	95.2%	86.4%	68.3% (11.6%)	76.9%	78.6%	57.1%	100%
RBF-SVM	90.7% (11.2%)	90.6%	91.9%	89.3%	73.6% (6.3%)	76.9%	78.6%	57.1%	100%
ELM	89.4% (9.9%)	88.9%	94.3%	83.5%	73% (5.2%)	76.9%	78.6%	57.1%	100%
KNN	85.3% (12.2%)	84.9%	88.1%	81.7%	70.3% (10.6%)	76.9%	78.6%	57.1%	100%
Naïve Bayes	86.8% (9.8%)	86.3%	91.7%	80.9%	68.5% (12.2%)	76.9%	78.6%	57.1%	100%
LDA	85.3% (12.2%)	84.9%	88.1%	81.7%	59.9% (12.5%)	61.5%	61.9%	57.1%	66.7%
QDA	85.9% (9.3%)	85.4%	91.7%	79.1%	70.3% (10.3%)	76.9%	78.6%	57.1%	100%

B

Classifiers	Cross-validation Accuracy				Hold-out Test Accuracy				
	Unbalanced	Balanced	Controls	AD	w/o Voting	Voting	Balanced	Controls	AD
Bagged Trees	84.6% (11.6%)	83.7%	92.9%	74.5%	76.7% (5%)	76.9%	78.6%	57.1%	100%
Random Forest	85.3% (12.5%)	84.4%	92.6%	76.2%	78.5% (5.1%)	76.9%	78.6%	57.1%	100%
Rotation Forest	82.3% (14.2%)	82.1%	85%	79.1%	71.1% (5.1%)	69.2%	70.2%	57.1%	83.3%
Boosted Stumps	83.2% (13.7%)	82.5%	89%	75.9%	79.7% (8.4%)	84.6%	85.7%	71.4%	100%
Boosted Trees	71.6% (16.2%)	71.3%	73.8%	68.7%	71.2% (12.2%)	84.6%	85.7%	71.4%	100%
FC-NN	92.5% (8.6%)	92.2%	96%	88.4%	70.8%(5.6%)	76.9%	78.6%	57.1%	100%
MLP-NN	66.4% (17.5%)	63.1%	94.5%	31.6%	59.7% (9.7%)	53.8%	50%	100%	0%
LVQNET	83.3% (12.5%)	83.1%	86.4%	79.7%	69.1% (8.4%)	69.2%	70.2%	57.1%	83.3%
KNN	83.7% (14.3%)	84%	83.3%	84.6%	74% (5.1%)	76.9%	78.6%	57.1%	100%
SLR	57.9% (14.9%)	57.8%	60.2%	55.4%	56.5%(13.8%)	61.5%	61.9%	57.1%	66.7%
RLR	87.8% (10.5%)	87.7%	89.0%	86.4%	76.9% (1%)	76.9%	78.6%	57.1%	100%
RVM	83.6% (12.7%)	83.6%	87.1%	80.0%	77.1% (1.1%)	76.9%	78.6%	57.1%	100%

Table 3.20. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE

framework for the ADNI data we collected, for the binary classification problem between healthy controls and subjects with Alzheimer's disease. The training/validation data and the hold-out test data are matched in age with subjects from age range of 56-88 years. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 84.6% while the best balanced hold-out test accuracy obtained was 85.7% for Boosted Trees and Stumps.

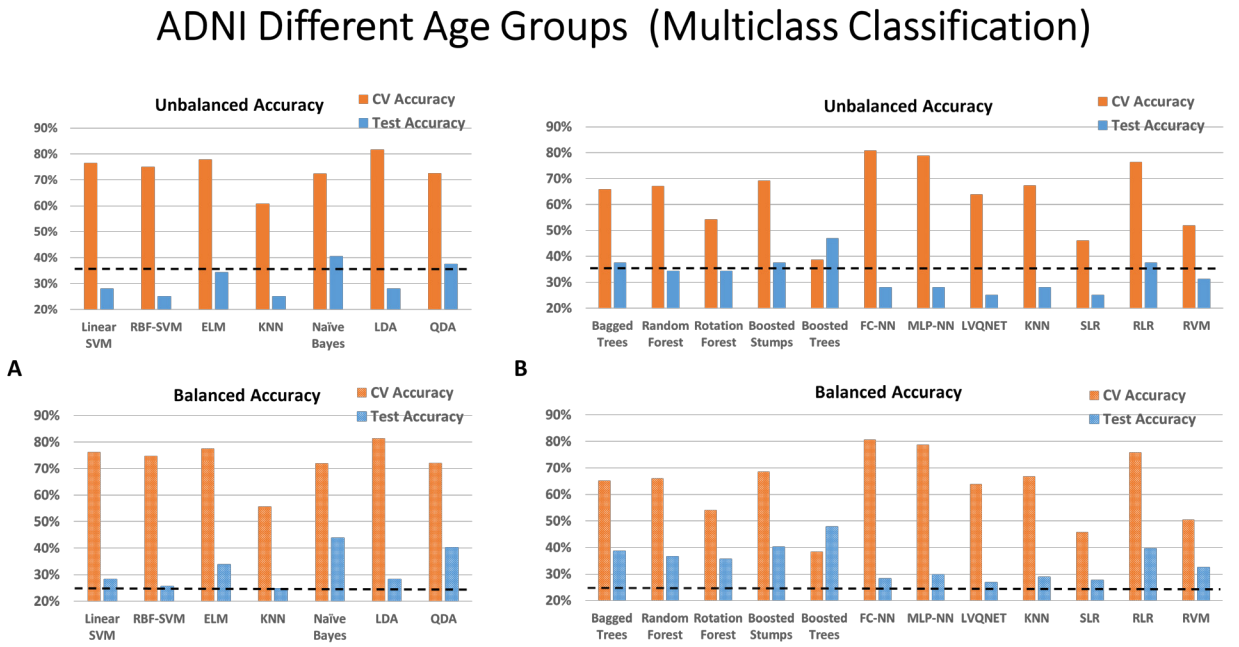


Figure 3.26. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ADNI data when the training/validation data and the hold-out test data are from different age groups in the range for the multiclass classification between healthy controls and subjects with early and late mild cognitive impairment and Alzheimer's disease at the end of spectrum. The training/validation data is from an age range of 56-76 years while the data from the age range of 77-88 years was used as a hold-out test data. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 37.5% since healthy controls formed 37.5% of the total size of the hold-out test data. For balanced accuracy, this is exactly 25%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best

hold-out test accuracy was 46.9% while the best balanced hold-out test accuracy was 47.9% both obtained for Boosted Trees.

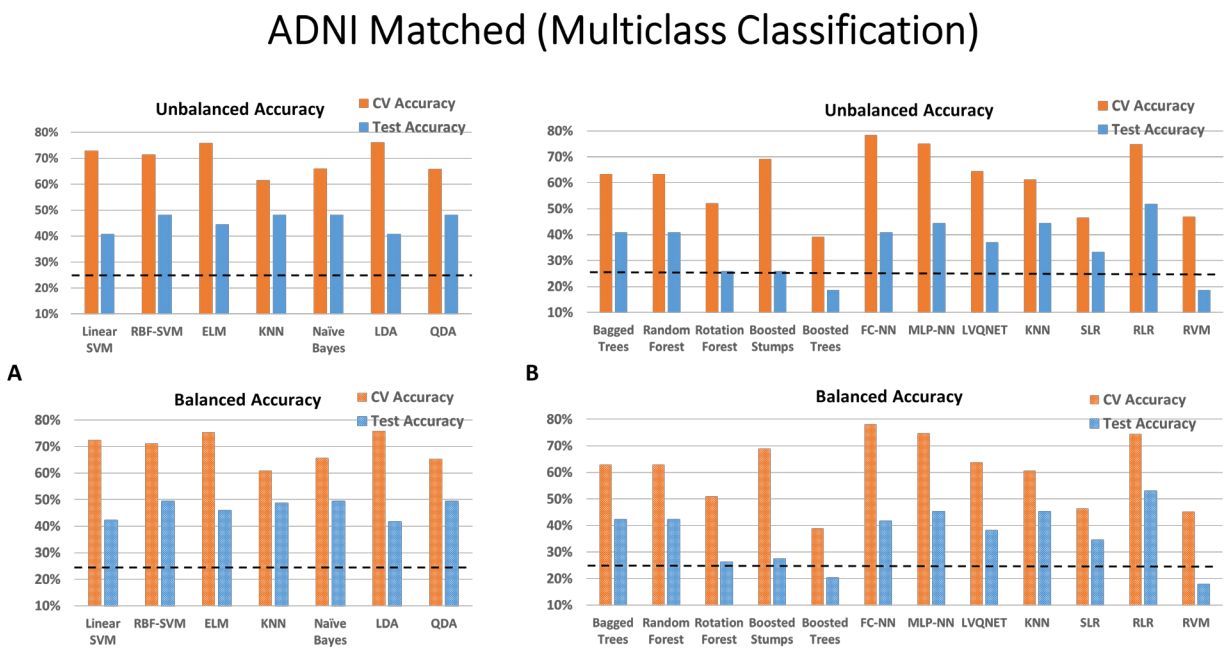


Figure 3.27. Unbalanced and balanced accuracy estimates for various classifiers (A) with RCE framework (B) outside RCE framework for ADNI data when the training/validation data and the hold-out test data are from different age groups in the range for the multiclass classification between healthy controls and subjects with early and late mild cognitive impairment and Alzheimer's disease at the end of spectrum. The training/validation data and the hold-out test data are matched in age with subjects from age range of 56-88 years. The balanced accuracy was obtained by averaging the individual class accuracies. The orange bars indicate the cross-validation (CV) accuracy while the blue bars indicate the accuracy for the hold-out test data obtained by the voting procedure. The dotted line indicates the accuracy obtained when the classifier assigns the majority class to all subjects in the test data. For unbalanced accuracy, this happens to be 25.9% since healthy controls formed 25.9% of the total size of the hold-out test data. For balanced accuracy, this is exactly 25%. We chose the majority classifier as the benchmark since the accuracy obtained must be greater than that if it learns anything from the training data. The discrepancy between the biased estimates of the cross-validation accuracy and the unbiased estimates of the hold-out accuracy is noteworthy. The best hold-out test accuracy was 51.8% while the best balanced hold-out test accuracy was 47.9% both obtained for Regularized Logistic Regression (RLR).

ADNI Different Age Groups (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy						Hold-out Test Accuracy						
	Unbalanced	Balanced	Controls	EMCI	LMCI	AD	w/o Voting	Voting	Balanced	Controls	EMCI	LMCI	AD
Linear SVM	76.5% (9.8%)	76.2%	74.5%	78.8%	79.3%	72.3%	29.7% (6%)	28.1%	28.3%	25%	16.7%	28.6%	42.9%
RBF-SVM	75% (10%)	74.7%	74.5%	79.1%	74.8%	70.3%	30.1% (6.5%)	25%	25.6%	16.7%	0%	28.6%	57.1%
ELM	77.8% (9.5%)	77.5%	73.5%	81.5%	78.5%	76.6%	31.1% (5.7%)	34.4%	33.9%	33.3%	16.7%	28.6%	57.1%
KNN	60.7% (10.8%)	55.7%	60.9%	42.5%	55.0%	64.4%	30.1% (6.1%)	25%	24.7%	25%	16.7%	14.3%	42.9%
Naïve Bayes	72.4% (10.5%)	72%	72.4%	82.2%	67.7%	65.6%	33.9% (5.8%)	40.6%	43.8%	25%	50%	42.9%	57.1%
LDA	81.6% (8.7%)	81.4%	79.1%	81.4%	85.7%	79.2%	29.2% (5.6%)	28.1%	28.3%	25%	16.7%	28.6%	42.9%
QDA	72.6% (10.1%)	72.2%	72.3%	82.5%	67.8%	66.2%	33% (5.8%)	37.5%	40.2%	25%	50%	28.6%	57.1%

B

Classifiers	Cross-validation Accuracy						Hold-out Test Accuracy						
	Unbalanced	Balanced	Controls	EMCI	LMCI	AD	w/o Voting	Voting	Balanced	Controls	EMCI	LMCI	AD
Bagged Trees	65.9% (11.9%)	65.2%	55.8%	78.0%	67.8%	59.1%	34.5% (5.3%)	37.5%	38.7%	33.3%	50%	14.3%	57.1%
Random Forest	67% (11%)	66.1%	55.5%	79.7%	69.8%	59.4%	33.4% (4.8%)	34.4%	36.6%	25%	50%	14.3%	57.1%
Rotation Forest	54.2% (11.6%)	54.1%	54.2%	55.3%	55.1%	51.9%	35.1% (5%)	34.4%	35.7%	33.3%	66.7%	28.6%	14.3%
Boosted Stumps	69.1% (11%)	68.6%	60.4%	73.4%	74.5%	65.9%	37.6 %(5.9%)	37.5%	40.2%	25%	50%	42.9%	42.9%
Boosted Trees	38.7% (10.4%)	38.4%	34.1%	43.9%	39.9%	35.5%	30.3% (7.6%)	46.9%	47.9%	41.7%	50%	42.9%	57.1%
FC-NN	80.8% (9%)	80.6%	75.6%	82.6%	84.6%	79.4%	29% (6.3%)	28.1%	28.3%	25%	16.7%	28.6%	42.9%
MLP-NN	78.7% (9.8%)	78.7%	71.5%	77.4%	82.2%	83.7%	27.6% (5.8%)	28.1%	29.8%	16.7%	16.7%	28.6%	57.1%
LVQNET	63.9% (11.3%)	63.9%	61.3%	68.7%	59.1%	66.4%	31.4% (4.7%)	25%	26.8%	16.7%	33.3%	0%	57.1%
KNN	67.3% (10.3%)	66.7%	58.6%	78%	65.7%	64.5%	31.5% (4.2%)	28.1%	28.9%	25%	33.3%	14.3%	42.9%
SLR	46.1% (11.5%)	45.8%	35.7%	50.5%	48.4%	48.4%	28.5% (7.2%)	25%	27.7%	8.3%	16.7%	42.9%	42.9%
RLR	76.3% (9.7%)	75.8%	65.9%	84.8%	78.7%	73.6%	31.9% (3.8%)	37.5%	39.6%	25%	33.3%	42.9%	57.1%
RVM	51.9% (13.1%)	50.5%	40%	71.9%	56.1%	33.8%	28.8% (7.2%)	31.2%	32.5%	25%	33.3%	42.9%	28.6%

Table 3.21. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ADNI data we collected, for the multiclass classification between healthy controls and subjects with early and late mild cognitive impairment and Alzheimer's disease at the end of spectrum. The training/validation data is from an age range of 56-76 years while the data from the age range of 77-88 years was used as a hold-out test data. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 46.9% while the best balanced hold-out test accuracy was 47.9% both obtained for Boosted Trees.

ADNI Matched (Multiclass Classification)

A

Classifiers	Cross-validation Accuracy						Hold-out Test Accuracy						
	Unbalanced	Balanced	Controls	EMCI	LMCI	AD	w/o Voting	Voting	Balanced	Controls	EMCI	LMCI	AD
Linear SVM	72.8% (10.2%)	72.5%	76.5%	75.3%	74.4%	63.6%	38.7% (7.3%)	40.7%	42.3%	42.9%	28.6%	14.3%	83.3%
RBF-SVM	71.3% (10.6%)	71.1%	74.6%	75.9%	69.6%	64.2%	41.1% (7.2%)	48.1%	49.4%	57.1%	28.6%	28.6%	83.3%
ELM	75.8% (9.8%)	75.4%	80.1%	79.5%	75.2%	66.9%	42.7% (6.5%)	44.4%	45.9%	42.9%	42.9%	14.3%	83.3%
KNN	61.5% (10.8%)	61%	78.9%	57.8%	55.7%	51.6%	35.9% (7.8%)	48.1%	48.8%	57.1%	42.9%	28.6%	66.7%
Naïve Bayes	65.9% (10.5%)	65.8%	72.9%	80.1%	48.3%	61.7%	42.7% (8.6%)	48.1%	49.4%	42.9%	42.9%	28.6%	83.3%
LDA	76.1% (8.6%)	75.9%	76%	81%	77.6%	68.9%	36.6% (7.5%)	40.7%	41.7%	57.1%	28.6%	14.3%	66.7%
QDA	65.7% (10.7%)	65.4%	72.5%	80%	47.7%	61.5%	41.7% (9.3%)	48.1%	49.4%	42.9%	42.9%	28.6%	83.3%

B

Classifiers	Cross-validation Accuracy						Hold-out Test Accuracy						
	Unbalanced	Balanced	Controls	EMCI	LMCI	AD	w/o Voting	Voting	Balanced	Controls	EMCI	LMCI	AD
Bagged Trees	63.3% (11.2%)	62.9%	74.4%	75.1%	48.6%	53.3%	40.5% (5.4%)	40.7%	42.3%	42.9%	42.9%	0%	83.3%
Random Forest	63.3% (10.6%)	62.9%	74.2%	76%	48.4%	52.8%	40.2% (5.4%)	40.7%	42.3%	42.9%	42.9%	0%	83.3%
Rotation Forest	52% (10%)	50.9%	79.8%	62.7%	32.3%	28.9%	23.6% (5.2%)	25.9%	26.2%	42.9%	28.6%	0%	33.3%
Boosted Stumps	69.1% (10.3%)	68.9%	74.6%	74.5%	63.7%	62.6%	30.1% (6.3%)	25.9%	27.4%	14.3%	28.6%	0%	66.7%
Boosted Trees	39.1% (10.7%)	38.9%	39.1%	36.8%	46.7%	33%	27.4% (8.4%)	18.5%	20.3%	14.3%	0%	0%	66.7%
FC-NN	78.3% (9.1%)	78.1%	78.2%	84.2%	77.5%	72.6%	39.2% (6.3%)	40.7%	41.7%	57.1%	28.6%	14.3%	66.7%
MLP-NN	75% (9.8%)	74.7%	77%	76%	77.9%	67.9%	36.9% (7.7%)	44.4%	45.3%	57.1%	28.6%	28.6%	66.7%
LVQNET	64.3% (10.5%)	63.8%	77.5%	69.1%	52.6%	56.1%	36.7% (6.2%)	37%	38.1%	42.9%	14.3%	28.6%	66.7%
KNN	61.1% (10.7%)	60.6%	77.8%	56.3%	56.1%	52.1%	40% (5.5%)	44.4%	45.3%	57.1%	28.6%	28.6%	66.7%
SLR	46.4% (10.4%)	46.4%	50.3%	41.8%	48.4%	44.9%	30.4% (7.3%)	33.3%	34.6%	42.9%	28.6%	0%	66.7%
RLR	74.8% (10.6%)	74.5%	81.6%	77.7%	70.6%	68%	48.3% (4.8%)	51.8%	53%	57.1%	42.9%	28.6%	83.3%
RVM	46.8% (10.3%)	45.2%	72%	69.2%	32%	7.4%	23.7% (6.5%)	18.5%	17.9%	28.6%	14.3%	28.6%	0%

Table 3.22. The table shows the cross-validation and the test accuracy as well the individual class accuracies for the classifiers implemented (A) within the RCE and (B) outside the RCE framework for the ADNI data we collected, for the multiclass classification between healthy controls and subjects with early and late mild cognitive impairment and Alzheimer's disease at the end of spectrum. The training/validation data and the hold-out test data are matched in age with subjects from age range of 56-88 years. The values in the parenthesis indicate the standard deviation for the accuracy metrics. The test accuracy with voting indicates the accuracy obtained when all classifier models obtained by the different partitionings during cross-validation, vote on the observations in the hold-out test data. The test accuracy without voting indicates mean accuracy when individual classifier models are used to classify the test observations. The top 3 classifiers both within and outside the RCE framework which had the highest hold-out test accuracies are highlighted. The best hold-out test accuracy was 51.8% while the best balanced hold-out test accuracy was 47.9% both obtained for Regularized Logistic Regression (RLR).

The accuracies were much better when the extreme ends of the spectrum were considered between healthy adults and adults diagnosed with Alzheimer's disease. In the binary classification scenario between Controls and AD, for the split in which the training/validation and the hold-out test data belonged to different age ranges, the best hold-out test accuracy was 83.3% while the best balanced hold-out test accuracy obtained was 76.2% for Sparse Logistic Regression. In the multiclass classification for the same split, the best hold-out test accuracy was 46.9% while the best balanced hold-out test accuracy was 47.9% both obtained for Boosted Trees. In the split wherein the training/validation and the hold-out test data were matched for age, accuracies were higher compared to the unmatched case. The best hold-out test accuracy was 84.6% while the best balanced hold-out test accuracy obtained was 85.7% for Boosted Trees and Stumps. For the 4-way classification across the spectrum, the best hold-out test accuracy was 51.8% while the best balanced hold-out test accuracy was 47.9% both obtained for Regularized Logistic Regression (RLR).

Feature importance: The combined feature importance scores (CFIS) for the two splits were plotted in a scatter plot as shown in Figure 3.28 for the binary and multiclass cases. The CFIS have higher variability and a smaller slope in binary compared to the multiclass classification scenario. The top connectivity paths whose means were significantly different between the groups ($p<0.05$, corrected) as well as have high combined feature importance scores are shown in Figure 3.29, while the top 20 regions in the brain whose connectivity paths were altered in the disease are shown in the Table 3.23.

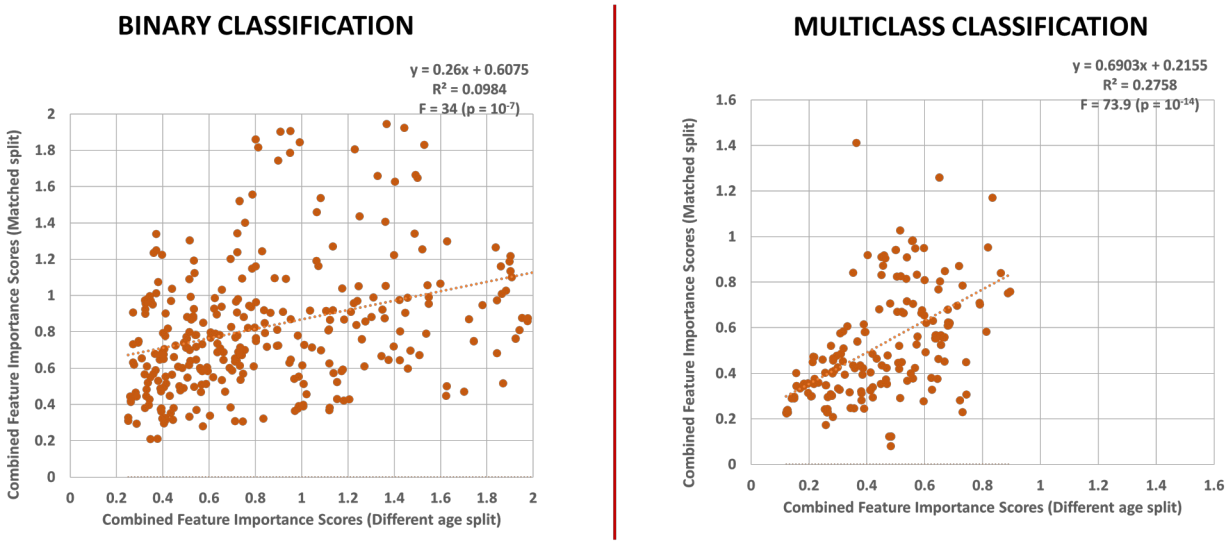
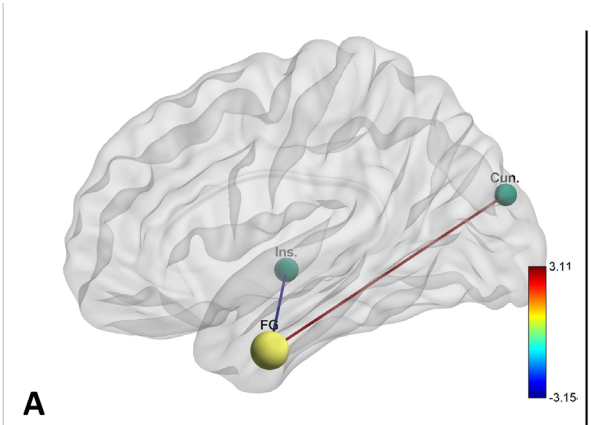
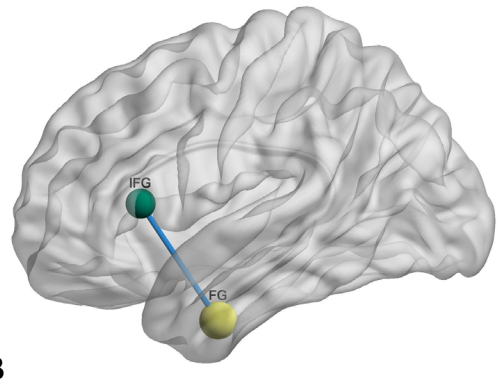


Figure 3.28. Scatter plots of combined feature importance scores (CFIS) for the two splits performed on the ADNI data. We selected the common features in obtained and plotted them as scatter plots for both binary and multiclass classification scenarios. The plot illustrates that there is significant agreement in the CFIS across splits. However, a lot of variability is present as well and this can be attributed to age ranges of the training/validation data obtained from the two splits.





Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(40.00,-9.00,-5.00)	Insula	Ins.	(-31.00,-3.00,-33.00)	Fusiform gyrus	FG	-3.15
(17.00,-88.00,21.00)	Cuneus	Cun.	(-31.00,-3.00,-33.00)	Fusiform gyrus	FG	3.12

Brain Region 1			Brain Region 2			CFIS
Coordinates	Brain Regions	Abbr.	Coordinates	Brain Regions	Abbr.	
(-50.00,25.00,8.00)	Inferior frontal gyrus	IFG	(-31.00,-3.00,-33.00)	Fusiform gyrus	FG	-1.31

Figure 3.29. The figure illustrates the connectivity paths which have significantly different means between the groups ($p<0.05$, corrected for multiple comparisons using permutation test) as well as are among the top hundred most discriminative paths in ADNI data for (A) binary classification between controls and Alzheimer's disease (AD). (B) 4-way classification between healthy controls, Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and AD. The size of the nodes indicates the relative importance of the region (Table 3.23). Common nodes between binary and multiclass classification are indicated by yellow while other nodes are indicated by green. The sign of the paths indicates over-connectivity (positive) or under-connectivity (negative) in healthy controls compared to clinical populations. So, red

represents a higher connectivity between controls compared to the diseased populations and blue represents a lower connectivity. The numerical values in the color bar denote the combined feature importance score of the path (CFIS) obtained from classification. A higher absolute number indicates more discriminative ability for the functional connectivity path. The table below the figure tabulates the brain regions involved in the paths visualized above along with the abbreviations of the two regions and the CFIS (combined feature importance score) for the connectivity paths.

A

Rank	X	Y	Z	Brain Regions	Abbr.
1	-31	-3	-33	Fusiform gyrus	FG
2	29	-32	62	Postcentral gyrus	PCG
3	-41	-51	-35	Cerebellum	Cbl.
4	-41	-11	12	Insula	Ins.
5	43	-72	-12	Inferior occipital gyrus	IOG
6	36	20	3	Insula	Ins.
7	-50	25	8	Inferior frontal gyrus	IFG
8	-15	-64	55	Precuneus	Prec.
9	51	23	20	Inferior frontal gyrus	IFG
10	1	-55	-13	Cerebellum	Cbl.
11	-27	-72	40	Middle occipital gyrus	MOG
12	-19	-79	-32	Cerebellum	Cbl.
13	11	-18	8	Thalamus	Thal.
14	42	12	-38	Middle temporal gyrus	MTG
15	40	-9	-5	Insula	Ins.
16	-29	-36	-12	ParaHippocampal gyrus	PHG
17	8	-34	46	Cingulate gyrus	CG
18	45	11	31	Middle frontal gyrus	MFG
19	35	-85	9	Middle occipital gyrus	MOG
20	17	-88	21	Cuneus	Cun.

B

Rank	X	Y	Z	Brain Regions	Abbr.
1	-41	-11	12	Insula	Ins.
2	0	54	25	Middle frontal gyrus	MFG
3	-31	-3	-33	Fusiform gyrus	FG
4	-40	22	41	Middle frontal gyrus	MFG
5	43	-72	-12	Inferior occipital gyrus	IOG
6	51	30	4	Inferior frontal gyrus	IFG
7	-41	-51	-35	Cerebellum	Cbl.
8	-5	-51	57	Precuneus	Prec.
9	52	-60	0	Middle temporal gyrus	MTG
10	-57	-9	4	Superior temporal gyrus	STG
11	-50	25	8	Inferior frontal gyrus	IFG
12	57	-6	-9	Superior temporal gyrus	STG
13	14	58	28	Superior frontal gyrus	SFG
14	1	-55	-13	Cerebellum	Cbl.
15	17	-54	2	Lingual gyrus	LG
16	-32	-19	-19	Hippocampus	HC
17	29	-32	62	Postcentral gyrus	PCG
18	42	12	-38	Superior temporal gyrus	STG
19	-19	-79	-32	Cerebellum	Cbl.
20	-52	-12	39	Postcentral gyrus	PCG

Table 3.23. This table illustrates the top 20 regions for MCI and AD as identified by (A) Binary classification between Healthy Controls and AD (B) Multiclass classification between Controls, MCI, and AD, using ADNI data.

3.3.6 Performance metrics from the consensus classifier

The hold-out test accuracies obtained from the consensus classifier (when all the classifiers are combined) for each of the four datasets are shown in Table 3.24. We list voting hold-out test accuracy, balanced hold-out test accuracy and individual class accuracies obtained by the consensus classifier. It is clear from the results of the various splits that the classifier which has the best hold-out test accuracy in one split may not perform as well on the other splits. Similarly, the classifiers which have high cross-validation accuracy does not always have a high hold-out test accuracy. Though the performance from the consensus classifier is less than the performance

of the best classifier for the split, it consistently gives excellent performance across all splits by leveraging the predictive power of individual classifiers.

	Binary Classification				Multiclass Classification					
A	Data Splits	Voting	Balanced	Controls	ASD	Data Splits	Voting	Balanced	Controls	Asperger's	Autism
	Different Age Ranges	66.2%	60.1%	72.4%	47.8%	Different Age Ranges	58.3%	41.2%	78.7%	0.0%	45.0%
	Different Acquisition Sites	64.3%	65.3%	52.8%	77.8%	Different Acquisition Sites	58.2%	40.8%	77.4%	0.0%	45.1%
	Matched	67.2%	66.0%	75.7%	56.3%	Matched	62.3%	43.7%	81.1%	0.0%	50.0%
B	Data Splits	Voting	Balanced	Controls	ADHD	Data Splits	Voting	Balanced	Controls	ADHD-C	ADHD-I
	ADHD-200 Competition	57.2%	59.6%	87.2%	27.2%	ADHD-200 Competition	54.1%	35.5%	90.4%	12.2%	3.9%
C	Data Splits	Voting	Balanced	Controls	PTSD	Data Splits	Voting	Balanced	Controls	PTSD	PCS+PTSD
	Different Age Ranges	72.2%	50%	0%	100%	Different Age Ranges	58.3%	51.8%	40%	37.5%	77.8%
	Matched	91.4%	86.4%	72.7%	100%	Matched	91.4%	87.4%	90.9%	71.4%	100%
D	Data Splits	Voting	Balanced	Controls	AD	Data Splits	Voting	Balanced	Controls	EMCI	LMCI	AD
	Different Age Ranges	57.9%	57.7%	58.3%	57.1%	Different Age Ranges	37.5%	40.2%	25.0%	50.0%	28.6%	57.1%
	Matched	76.9%	78.6%	57.1%	100%	Matched	48.1%	49.4%	42.9%	42.9%	28.6%	83.3%

Table 3.24. Accuracy obtained by the consensus classifier for the various splits (A) ABIDE dataset (B) ADHD-200 dataset (C) PTSD dataset (D) ADNI dataset.

3.3.7 Effect of age and site variability

To understand the effects of age and site variability on the accuracy obtained from the hold-out test data, we used the consensus classifier to compare and contrast. This way we can draw generalized inferences about the predictive capability of the classifiers without reference to any specific classifier. We compared the overall accuracies as well as the individual class accuracies when the training/validation data and the hold-out test data were matched as well as a case in which there was age or site differences between the two. The corresponding consensus accuracies for ABIDE, PSTD and ADNI datasets are shown in Figures 3.30-3.32, respectively. As expected, the accuracy with matched data was higher than in unmatched case. The difference in consensus accuracies due to age was particularly sharp in small datasets such as PTSD and

ADNI. These figures illustrate that smaller datasets with high homogeneity overestimate the actual predictive capability of the classifiers and could give optimistic accuracy results that do not generalize well to the larger population.

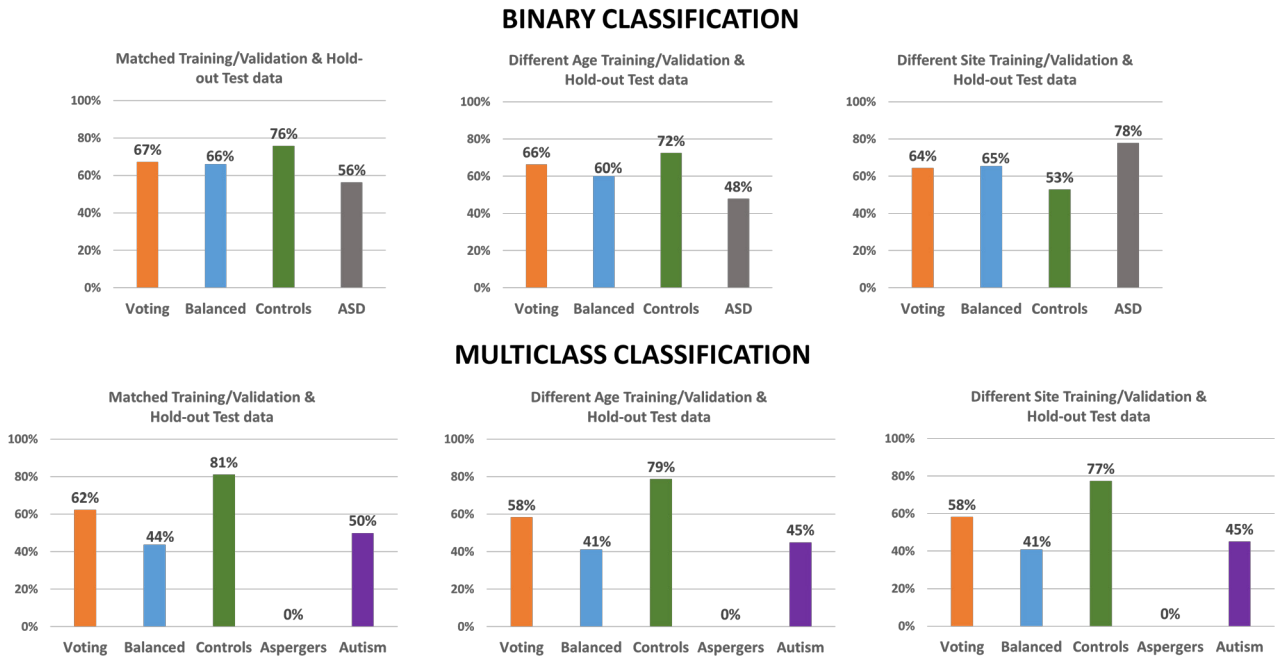


Figure 3.30. The figure shows the differences in overall accuracy as well as individual class accuracies in the ABIDE dataset which can be attributed to age and site variability. These consensus accuracies were obtained by combining the predictions of all the 19 classifiers in a probabilistic way to vote on the hold-out test dataset. As expected, the split wherein the training and hold-out test data were matched for age and acquisition site had the best performance, though it was more pronounced in the multiclass classification scenario. In fact, a three-way classification between healthy controls, subjects with Asperger's syndrome and Autism reduced the overall accuracy due to the relatively fewer subjects with Asperger's syndrome in the dataset. Overall the classifiers were reasonably successful in classifying the test observations.

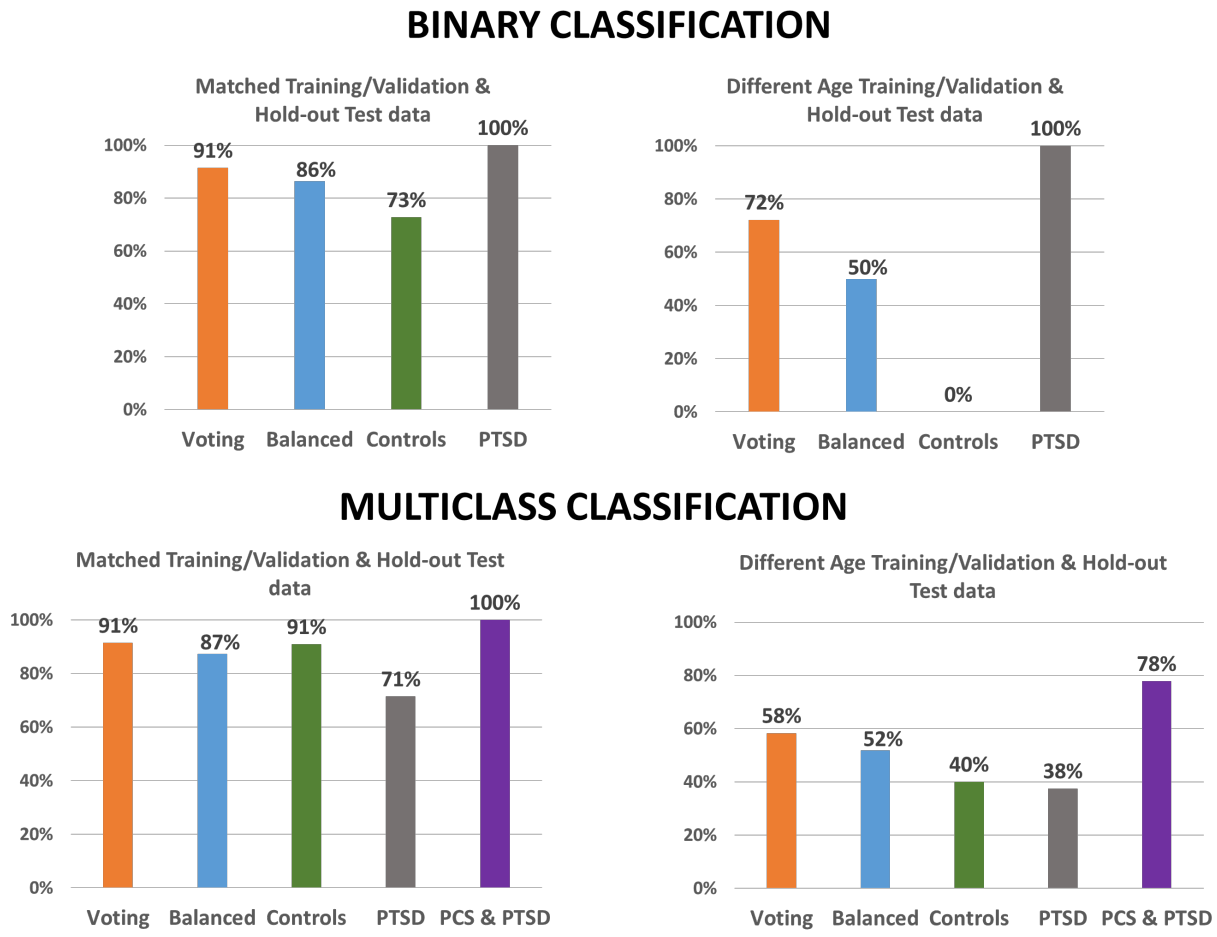


Figure 3.31. The figure shows the differences in overall accuracy as well as individual class accuracies in the PTSD dataset which can be attributed to age range differences in training/validation and hold-out test data. These consensus accuracies were obtained by combining the predictions of all the 19 classifiers in a probabilistic way to vote on the hold-out test dataset. As expected, the split wherein the training and hold-out test data were matched for age had the best performance. The accuracy in the split where the age range was different for training/validation and hold-out test data was terrible with all observations being classified as PTSD in the binary classification scenario. This shows that smaller datasets with homogeneity overestimate the actual predictive capability of the classifiers and do not generalize well to the overall population.

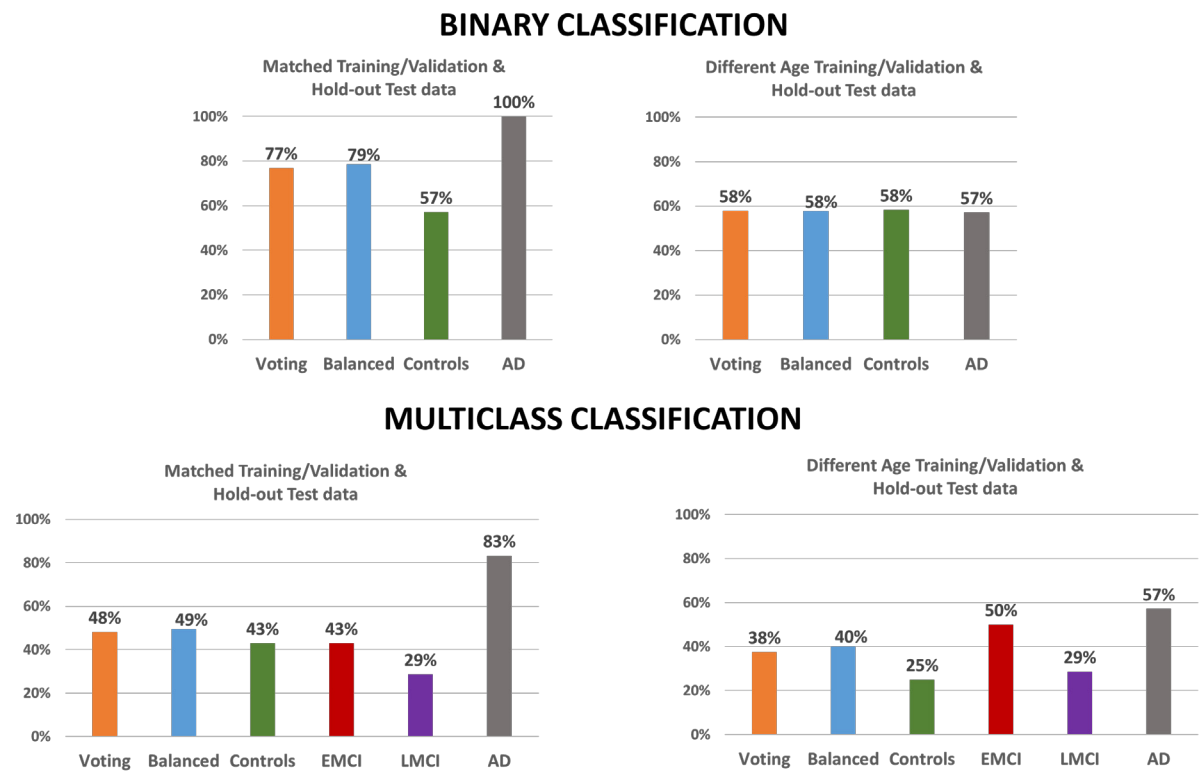


Figure 3.32. The figure shows the differences in overall accuracy as well as individual class accuracies in the ADNI dataset which can be attributed to age range differences in training/validation and hold-out test data. These consensus accuracies were obtained by combining the predictions of all the 19 classifiers in a probabilistic way to vote on the hold-out test dataset. Similar to the other two datasets, the split where the training and hold-out test data were matched on age had the best performance, though it was more pronounced in the binary classification scenario. The binary classification performed way better than multiclass classification as expected due to the difficulty in modeling the four classes with relatively small sample size. Similar to PSTD, smaller datasets with homogeneity overestimate the actual predictive capability of the classifiers and do not generalize well to the overall population.

3.3.8 Reliability of feature selection and parameter optimization

To investigate the wide discrepancies between cross-validation and hold-out test accuracies in smaller datasets, we compared the average accuracy per cluster plots as a function of features for every recursive cluster elimination step. This was done for both ADNI and ABIDE datasets. The figure for the Linear SVM classifier comparing both the datasets is shown in Figure 3.33. As non-discriminative features are eliminated in the RCE-framework, training accuracy increases.

For the ADNI dataset, unlike with the ABIDE dataset, removal of features did not translate to improvement in accuracy in the validation dataset. Similarly, model selection via parameter optimization for Support Vector machine within the RCE-framework does not particularly seem to improve the accuracy significantly for the ADNI dataset beyond that obtained by using a default value for the tuning parameter C equal to 0.1 as shown in Figure 3.34. In fact the accuracy was significantly less by using model selection than just using the default parameter. Whereas, for ABIDE dataset, hyperparameter optimization by grid search improved the accuracy compared to using the default parameter. The significance in the differences in accuracy with and without parameter optimization in ABIDE data becomes more appreciable as recursive cluster elimination progresses. The reason for the unreliability in feature selection and parameter optimization can be attributed to the lack of enough data to choose the optimal models, a problem, unfortunately, more pronounced in high dimensional datasets with smaller sample sizes.

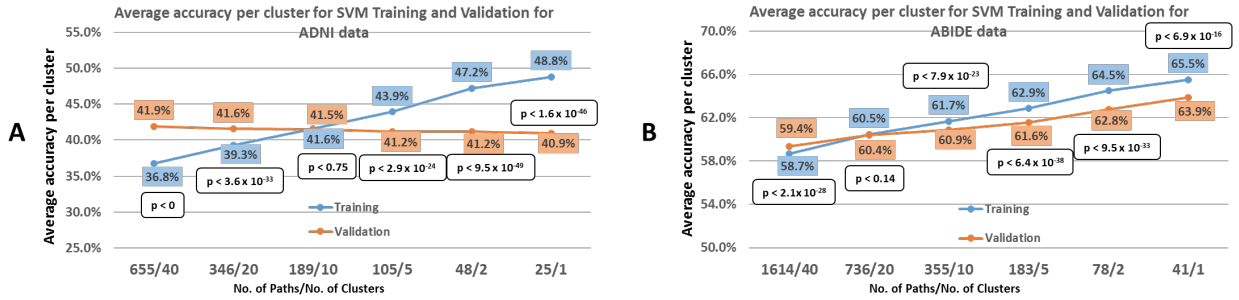


Figure 3.33. Changes in classification accuracy with feature elimination during training compared to validation. Results are shown in smaller datasets such as ADNI (A) as well as in larger datasets such as ABIDE (B). The RCE framework seemed to improve the accuracy as unnecessary features were eliminated in the training data. In ADNI dataset, unlike with the ABIDE dataset, removal of features did not translate to improvement in accuracy in the validation dataset. This demonstrates the unreliability of feature selection in smaller datasets.

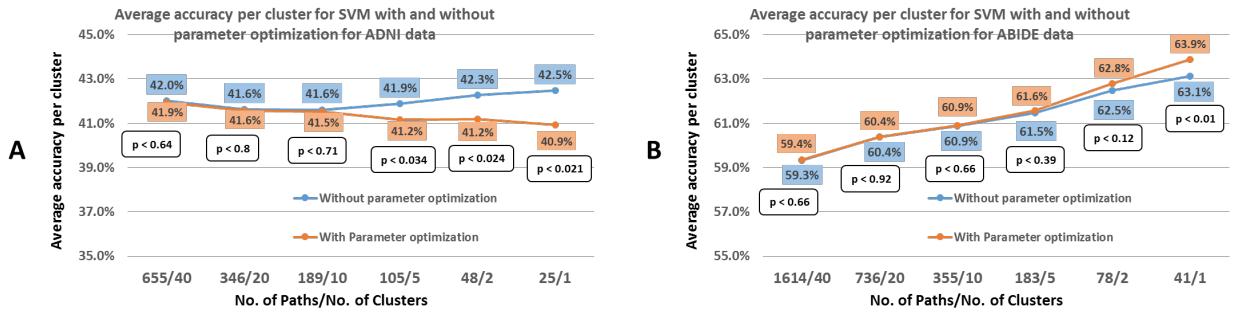


Figure 3.34. Changes in classification accuracy with feature elimination, both with and without model selection/hyperparameter optimization. Results are shown in smaller datasets such as ADNI (A) as well as with large datasets such as ABIDE (B). For the ADNI dataset, parameter optimization did not lead to an increase in the accuracy than the default parameters. In fact, it is less than what is observed without parameter optimization. Whereas, for the ABIDE dataset hyperparameter optimization by grid search improved the accuracy compared to using the default parameters as identifying the optimum hyperparameters is more reliable. This figure raise important questions about the unreliability of model selection/hyperparameter optimization in smaller datasets.

3.4 Discussion

We demonstrate that cross validation accuracy could provide overoptimistic estimates of classifier performances in homogeneous datasets and show how the hold-out test performance could actually be much lower than the cross-validation performance. This is an important conclusion given that cross-validation is a generally accepted standard in neuroimaging based classification, and yet is something that is completely unacceptable in other fields including industry. A simple example is Kaggle competitions (https://www.kaggle.com/competitions), which hosts several machine learning problems. In Kaggle competitions, data science enthusiasts submit their predictions on a hold-out test data and their performance is evaluated in real-time. This ensures that different researchers use the same hold-out test data which gives an objective estimation of performance and allows for comparison of performance across classifiers, which is especially critical for clinical diagnosis. Cross-validation is incapable of an objective estimation of performance given the differences in data samples used by individual researchers which can

affect the classification performance. Another pertinent example is the CALO (Cognitive Assistant that Learns and Organizes) project (https://en.wikipedia.org/wiki/CALO) which gave rise to many software products including the voice recognition feature "SIRI" in Apple's iPhones. While cross-validation was used to build models, they were typically not used as a performance metric while developing products such as SIRI within the CALO project. Our results suggest that neuroimaging must adopt industry-standards while employing machine learning for diagnostic classification, wherein the classification performance is always assessed using a completely independent hold-out test data.

Second, we sought to understand how overfitting can occur in the context of machine learning applied to neuroimaging-based diagnostic classification. We implemented 18 classifiers covering a spectrum of popular machine learning classifiers based on diverse principles. Our results show that during both feature selection and performance estimation, smaller datasets might give unreliable estimates of classifier performance which could lead to improper model selection leading to poor generalization across the larger population. To address issues with classifier variance and improve predictive ability, we also propose a consensus classifier. The consensus classifier exploits the predictive abilities of individual classifiers to build a single classifier so that inferences drawn are not driven by overfitting by any individual classifier.

Finally, we wanted to identify functional connectivity features that are insensitive to different sources of variability identified above as well as have good statistical separation between groups along with good predictive ability. In fact, the proposed combined feature importance scores we assigned to connectivity features were aggregated from multiple classifiers implemented in the RCE framework. Connectivity features thus identified are likely to be robust and genuine markers of underlying brain network disruptions caused by the disorders rather than an artifact of

other extraneous factors. We make our code publicly available for replication of our results and to encourage better practices in research. To the best of our knowledge, this is the most comprehensive exploration of state-of-the-art machine learning algorithms for neuroimaging based diagnostic classification.

The discussion section is organized as follows. We start with a discussion of the methodological aspects associated with our study which is then followed by a discussion of specific insights we obtained using each of the four clinical datasets. We first discuss the strengths of our study with regard to the use of a data driven feature selection strategy. We then examine in detail the issues encountered during classification. Specifically, we discuss how data heterogeneity in age and acquisition site can affect the classifier performance. As our results indicate, data heterogeneity can reduce classification accuracy and the effect is much more pronounced in relatively smaller datasets such as PTSD and ADNI. We also expand on how model selection and performance estimation can be unreliable in these small datasets. We then discuss some issues in the acquisition and processing of data that can effect classification accuracy and speculate on how multimodal imaging might improve classification performance. We end this section by discussing the classification results, connectivity paths and the regions associated with the 4 disorders we examined – ASD, ADHD, PCS & PTSD and MCI & AD, in order – with special emphasis on disrupted functional networks identified by us.

3.4.1 Use of a data driven approach for feature selection

We used a data-driven model in this study where we made no prior assumptions about the brain regions or connectivity paths involved in the underlying disorders. Prior assumptions about the effects of the disease on the brain regions have be used before to reduce the number of features to a more reasonable number [55, 56, 57]. However, in this paper, we used a data-driven

approach to reduce the number of features and improve the classification accuracy of the classifiers for three reasons: (i) The feature selection methods we used can potentially provide insights into the neurophysiological aspects of the disease and help validate current hypotheses about the underlying connectivity disruptions in these disorders [58], (ii) There is a growing body of evidence that neurological diseases target large-scale distributed brain networks. Hence by limiting ourselves to a few regions in the brain, we might not even consider features with potentially useful information and limit our understanding of the underlying neuropathology of the disease, (iii) Lack of specificity of brain networks in disease identification. For example, DMN dysregulation is implicated in several disorders. Therefore, using a data driven feature selection algorithm can help us identify networks which are likely to be specific to the disorders.

3.4.2 Issues with performance estimation and feature selection for small datasets

Cross-validation accuracy is an unreliable estimate of the true generalization accuracy in small datasets with few hundreds of samples [59, 49]. When we use cross-validation to select a large number of models, we risk overfitting and choosing the less optimal model. In machine learning literature, it is widely accepted that cross-validation performance is an ineffective measure of the true generalization performance due to the large variance associated with its estimates [60], although this fact is not widely appreciated in the neuroimaging community [58]. Also in smaller datasets, the hyperparameter values selected by minimizing the validation error might be tailored to the sample used for training and validation. This leads to overfitting in model selection and hence provides a biased estimate of classification performance in such small samples [60]. As we have shown in Figure 3.34 model selection by parameter optimization does not improve the performance in ADNI data in contrast to ABIDE data. In fact, model selection by parameter optimization in ADNI data performs significantly worse compared to the case without parameter

optimization. In ABIDE data, the optimal model can be selected based on cross-validation, as the cross-validation estimate is more reliable due to the size of the dataset. Not just for performance estimation, this fact also rings true in wrapper methods for feature selection or model selection when we use cross-validation with small datasets with higher dimensional features as is the case for typical neuroimaging datasets. It is noteworthy that datasets with hundreds of subjects are considered large in neuroimaging (which may be true for detecting activation and even for characterizing resting state networks). However, they are small given the dimensionality of the data and what we are trying to achieve with machine-learning based supervised diagnostic classification.

Given the problems associated with the "curse of dimensionality", reducing the number of features is important [41, 61]. Though useful for reducing features to more manageable numbers, t-test filtering might not be the best initial feature selection method as T-scores of features can, in principle, vary drastically across different folds of training data and consequently have poor predictive power [62, 6]. In t-test filtering, we are using statistical separation as a proxy for discriminative power and this may be true in some instances and may not be so in other instances, especially if groups of features, when combined, may provide discriminative ability in comparison to when used alone. Therefore, a filtering strategy by univariate tests might remove features with discriminative ability since those features are not selected based on a metric which directly asseses their discriminative power [62]. Our results do indicate such dangers posed by the use of filter methods such as t-test filtering, as observed in substantial variance in the selected feature importance scores, and possibly overfitting in small datasets such as PSTD (Figure 3.23) and ADNI (Figure 3.28). Therefore, instead of feature selection by t-test filtering, quick and

reliable methods such as GINI index might be useful as they may provide a better estimate of feature importance [62].

In our study, many classifiers implemented within the Recursive Cluster Elimination (RCE) framework gave better performance compared to the classifiers not implemented within RCE. The performance of the classifiers became better as the sample size increased. However, wrapper methods such as recursive cluster elimination are not perfect either. Specifically, the inner cross-validation we use in RCE for model selection does not reliably select the true model of the mapping between inputs and outputs in smaller datasets such as ADNI (Figure 3.33). This results in no significant change in accuracy per cluster in the validation data when such features are removed from the training data. In fact, models are prone to overfitting when a large number of models are tested against small samples of data [49], which holds true when RCE and/or parameter optimization is used in small datasets. Along with its feature reduction capabilities, RCE framework (and wrapper methods in general) has significant downsides as well, such as the difficulty in optimizing its tunable parameters. For obtaining the best results from RCE framework, we have to consider the dimensionality of each cluster, the computation time, number of clusters/models that we choose from and the number of features which needed to be eliminated at each step of the algorithm.

In our study, many features which had significant group differences were not useful in classification while some features with good classification scores did not necessarily show significant group differences. Therefore, we investigated features which were significantly different between the groups as well as had high discriminative/predictive ability (high FIS scores). Since different classifiers are sensitive to different patterns of features [61], the FIS scores obtained for the same features from different classifiers can, in principle be, different.

Therefore, we combined the FIS scores obtained from multiple classifiers for a given feature to provide a single combined feature importance score (CFIS). CFIS is our novel contribution and has not been done in previous studies to the best of our knowledge.

Finally, it is important that features are not selected using the entire dataset as it could lead to overoptimistic results which could generalize poorly to unseen data. Unfortunately, this practise is quite common in neuroimaging and it could lead to the leakage of information from training data to test data. This is sometimes referred to as "double dipping" [63] and could result in extremely optimistic accuracies in smaller datasets with large number of features [5]. In our results, some of the difference in CV accuracy and the test accuracy, even in the case where the training/validation data and the hold-out test data are matched for age and imaging site can be attributed to the t-test filtering performed on just the training/validation data. It should also be noted that if the features selected by the t-test indeed do have predictive power and are reproducible, then feature selection by t-test filtering should have a minimal impact on the classification accuracy. Compared to other classifiers, the difference between cross-validation and the hold-out test accuracy was the smallest for Boosted trees across all classifiers, indicating that it may not be overfitting the data compared to other classifiers. This may be because boosting is generally considered relatively robust to overfitting [64, 65]. However, dividing the data into training/validation and hold-out test data and performing feature selection only using the training/validation data may not be feasible in smaller datasets which are typical in neuroimaging. Probably the use of classifiers with built-in feature selection such as SLR or feature ranking such as Random Forest might be the way forward with feature reduction in noisy and relatively small datasets. It is also important to remember that the hold-out test accuracy is a conservative estimate of the actual predictive performance and hence is a better indicator of the

performance on unseen data especially in disorders with high heterogeneity and/or features which are not highly reproducible.

3.4.3 Effect of the data heterogeneity on the classification performance

Our results, taken together with previous reports, indicate that generalizing a classifier across different age groups and acquisition sites is difficult. We observed differences in accuracies when the model trained on data acquired at particular sites or age ranges were tested on data from a different age range or acquisition site (Table 3.24, Figures 3.30-3.32). In many studies, matching for age, sex, motion, scanning protocol, acquisition site, and IQ may not be feasible between the training/validation and the hold-out test data as well as between the Controls and the diseased group. This is truer for datasets which are pooled from many acquisition sites prospectively given the prohibitive costs involved in acquiring such large data homogeneously and retrospectively (although large retrospective studies have gotten underway recently. E.g. UK biobank study [40]. It is also possible that disease populations in a particular age range might exhibit over-connectivity compared to age-matched Controls while subjects with the same disease diagnosis from a different age range exhibit under-connectivity compared to age-matched Controls, as in ASD.

Data from different scanning sites are associated with variability in scanning equipment, scanning parameters, demographic, genetic or other experimental factors [7]. Datasets we encounter might not sample the entire population distribution. Cross-validation (CV) accuracy in our results suffer due to two primary factors. The first factor affecting the CV accuracy is the difficulty in generalizing the classifier to variations in the disease populations. The second factor affecting the CV accuracy is the bias introduced by feature selection on the cross-validation data. Because of this bias, choosing the optimum model from a large number of models with limited

validation data is difficult, as the data samples we collect might not adequately sample the population distribution space. With low disease prevalence and the high heterogeneity in disease populations, identifying reliable biomarkers with high sensitivity to the disease as well as good generalization in the population can only be achieved with the help of large collaborative multisite neuroimaging efforts even if they are put together retrospectively [58, 6]. Examples of such efforts include ADNI for Alzheimer's [66], ABIDE for Autism [29], ADHD-200 for ADHD [31], 1000 functional connectomes project for healthy subjects [67] and International Neuroimaging Data-sharing Initiative (INDI) [68]. The classification performances tested on such large datasets help in reproducabilty and generalizability of classification results. Three of the four datasets we used – ADNI, ABIDE and ADHD-200 – are from collaborative multisite acquisitons. Hence even if our accuracy appear to be lower compared to that reported by single-site studies with relatively smaller sample sizes, we expect our classification accuracy and consequently the disease encoding neuroimaging features to generalize well to the general population. Another factor limiting the utility of automated diagnostic tools is low disease prevalence in the general population. For developing diagnostic classification tools for the general population, low disease prevalence could lead to large false positives despite high specificity, thereby limiting its usefulness.

The diagnostic label associated with spectrum disorders such as Autism spectrum disorder might not be entirely informative given that these diseases correspond to multiple etiologies under the same term and thus finding biomarkers and getting good and reliable classification performance in classifying them are tough [6]. Many disorders are highly heterogeneous, and categorization of subgroups within many disorders is yet to be thoroughly established. Also, some of them are characterized by behavioral disabilities that form continuous or nearly continuous spectrum

spanning from relatively mild symptoms to very pronounced behavioral difficulties as can be the case with cognitive impairment and Alzheimer's disease. Another huge issue which is gaining attention is the misdiagnosis or over-diagnosis of subjects especially in children which make it difficult to know the actual disease state of the subject and make predictive learning models extremely difficult to optimize.

The above factors can make classification of sub-classes within spectrum disorders extremely difficult. All classifiers we tested struggled with the multiclass classification for almost all datasets compared to binary classification when using the hold-out test dataset. In a three-way classification between Controls, Asperger's syndrome and Autism, most classifiers failed to classify even a single subject with Asperger's syndrome accurately in the hold-out test data (Tables 3.7-3.9). The 3-way classification between Controls, ADHD-I, and ADHD-C, (Table 3.11) resulted in decreased accuracy compared to binary classification between Controls and ADHD (Table 3.12). As discussed previously, over-diagnosis or misdiagnosis hinders supervised classification. One way to overcome this issue will be to use unsupervised classification to drive subject labeling [69, 70]. In the case of ADNI dataset (Figures 3.24-3.27), the high accuracy obtained from the four-way classification using cross-validation did not generalize well to the holdout test dataset. Several factors could contribute to these results. Primarily, the lack of large training data available for disease subtypes, which is particularly the case for the ADNI dataset and the Asperger's sample in the ABIDE dataset.

3.4.4 Issues with the use of machine learning classifiers

The choice of the classifier and the features extracted are extremely crucial in providing insights about the neurobiological origins of the disease. There is no universally best learning algorithm that gives excellent performance for all datasets and features. So it is extremely difficult to know

beforehand which classifier might give the best performance. High prediction accuracy and interpretability of the classifier model are somewhat conflicting goals in neuroimaging [1]. In using complex classifiers using RBF kernels or neural networks might give excellent performance, but utility in translating the models to understanding the disrupted neural circuits in neurological disease is limited given the "black box" nature of such classifiers. Also, non-linear methods might not give optimal performance compared to linear methods when available training data is limited to model the complex relationships between the features and the disease status of the subject. In fact, the relative success of linear classifiers in neuroimaging is not due to the absence of complex relationships between features and subjects' diagnostic status, but rather due to the unavailability of large datasets required to model such relationships [43]. In our study we get the best of both the worlds in our use of machine learning classifiers as we achieve high prediction accuracy as well as interpretability for our results. Due to the use of multiple classifiers within the RCE framework, we were able to leverage the strengths of multiple types of machine learning classifiers not only to improve the classification performance, but also to provide us with combined feature importance scores (CFIS). CFIS scores were then used to identify the connectivity paths and regions encoding the disease states for the various discords studied. This greatly aids in interpretability of our results and provides us valuable information about connectivity dysregulation in the clinical populations.

Another observation from our results pertains to models with built-in regularization. Models with regularization to control model complexity performed well consistently across all datasets. In fact, sparse models such as RVM, SLR, RLR, and regularized neural networks gave consistently good performance across most datasets and the multiple splits we performed on each dataset.

Therefore, we believe that quality and the quantity of the data available must guide the best feature extraction methods and the choice of the classifier for each particular study.

In some cases, classifiers which performed best on the cross-validation dataset did not perform as well on the hold-out test data. It is possible that by reporting only the results of the classifier having the best performance, we are prone to using optimistic estimates of classification performance which might not even generalize well to subjects from the same population [53]. It is one of the reasons as to why we combined predictions from multiple classifiers to build a consensus classifier rather than reporting and emphasizing accuracy obtained by the best classifier. Also, combining multiple predictions from different classifiers can usually improve the overall classification performance as different types of classifiers rarely make the same kinds of mistakes on unseen data.

3.4.5 Issues with disease classification using RSFC metrics

Some issues we have not considered in this study might influence reported classification performance. The confounding effects of head motion and the correction strategies applied to ameliorate head motion artifacts, inclusion/exclusion of global signal regression (GSR) in the preprocessing pipeline, spatial variation in the Hemodynamic Response Function (HRF) across brain regions and subjects, as well as the duration of the scans can affect the reliability and reproducibility of RSFC metrics, which might ultimately affects classification performance. There is no general consensus in neuroimaging community on how to address these issues.

Scan duration is of particular importance in clinical populations when motion corrupted volumes are removed from the data (censoring), and it is crucial to have the necessary amount of data to reliably estimate RSFC metrics. Also, proper motion correction might improve the accuracy in classifying Controls from clinical subjects [71]. Probably the easiest way to increase the

accuracy of classification is to have longer scans times to make the features derived from RSFC more reliable [14, 15].

3.4.6 Multimodal Imaging

Resting state functional connectivity measures may not necessarily contain discriminative information for all mental disorders. With the rise of multimodal imaging, using multiple imaging metrics as features can capture different aspects of neuropathology. Structural connectivity measures obtained from DTI, morphological features from anatomical images, network theoretic measures derived from graph theory such as local connectivity, global efficiency, clustering coefficient, network modularity, characteristic path length etc., RSFC derived measures such as Amplitude of low-frequency fluctuations (ALFF)/ and fraction of amplitude of low-frequency fluctuations (fALFF), Regional Homogeneity (ReHo), Degree Centrality (DC), seed-based connectivity, causal directional relationships between brain regions (effective connectivity) [72] or use dynamic measures of synchronization between brain regions (Dynamic Functional and Effective Connectivity) [73], task-based activation, as well as measures derived from magnetic resonance spectroscopy could potentially be used as features to train classifiers. Multimodal measures have been used for classification in Autism with good results [74, 75]. In fact, to build a better model, along with multimodal imaging we can move beyond imaging metrics and incorporate prior information about the disease prevalence and its distribution in the population based on demographic and phenotype data into the classification algorithm. If the results reported by several research groups that participated in the ADHD-200 global competition is an indication [31], then much better accuracies can be achieved by combining neuroimaging data with the phenotypic data than by using neuroimaging based data alone.

3.4.7 ASD

In the binary classification of controls from ASD, we achieved accuracy as high as 67.2% (balanced accuracy of 66%) on the separate hold-out test dataset (Table 3.24). However, when the training/validation and the hold-out datasets are from different age ranges, the accuracy was reduced to 66.2% (balanced accuracy of 60.1%). In fact, the impact of age on the classification performance in the ABIDE data has been previously documented by Vigneshwaran et al. (2015). They report higher accuracies on the hold-out test dataset when adult males and adolescent males (age<18) were considered separately in classification, than when all male subjects were considered [76]. This study also reports higher hold-out test accuracy for adult males compared to adolescents, indicating the difficulty in classifying ASD in adolescents compared to young adults using RSFC metrics. The results of this study contradict an earlier study which obtained better classification performance for adolescents (89% with LOOCV, 91% with replication dataset) compared to young adults (79% with LOOCV, 71% with replication dataset) with 80 subjects for training/validation and 21 subjects in the replication dataset [50]. Age dependence is to be expected since ASD is a developmental disorder with atypical developmental trajectories including compensatory mechanisms in adulthood [77, 27]. Previous studies have reported increased resting state functional connectivity in ASD subjects under the age of 12 years, while studies involving adolescents and adults have reported reduced functional connectivity compared to healthy controls [78]. Also, behavioral measures have been shown to outperform fMRI-based measures for supervised classification of Autism [79]. Previous studies report accuracies in the mid to high 70s for single site studies with 40-80 subjects [50, 55]. The classification accuracy drops as the size of the dataset increases, with 79% LOOCV accuracy reported with 240 subjects [80] and dropping to as low as 60% LOOCV accuracy with 964 subjects [3] in multisite

studies using the ABIDE dataset. Motion does seem to play a signficant role in reducing classification performance as several studies using low-motion subjects achived much higher accuracies. Using 252 low-motion, age and motion matched cohorts from ABIDE, Chen et al (2013) achieved accuracies an of 91% (1- out of bag error (OOB)) with Random Forests [4] and using 640 subjects with age<20, Iidaka (2015) achieved cross-validation accuracy of 90% with probabilistic neural network (PNN) [81]. It is noteworthy that these high accuracies have been obtained using cross-validation. Differences between the training/validation and the hold-out test data in several factors such as imaging site, head motion, age, sex, IQ, and imaging protocol can cause an overestimation of classification accuracy in cross-validation. In fact, compared to 91% accuracy reported by using OOB error for Random Forests, considerably lower accuracies of 62% was obtained from hold-out test data in the same study [4].

ASD involves disruptions of interacting large-scale brain networks distributed across the brain [28]. We observed both under-connectivity and over-connectivity in subjects with ASD compared to the controls (Figure 3.9) as reported in previous studies [55, 27]. In fact several of the regions (Table 3.10) and connectivity paths (Figure 3.14) obtained in this study were also shown to be implicated in Autism [82]. Many regions associated with the default mode network such as posterior cingulate cortex (PCC), precuneus, medial prefrontal cortex, angular gyrus were found to be disrupted in subjects with ASD as several previous studies have indicated [83, 29, 84, 85]. Medial prefrontal cortex (MPFC) and anterior cingulate are involved in social processing [86] and hence are likely to be altered in subjects with ASD. Using ABIDE dataset, Iidaka found that the superior frontal gyrus (SFG), anterior and posterior cingulate (ACC & PCC) as well as the thalamus were most disrupted in Autism [81]. Connectivity between fusiform gyrus (FG) and middle occipital gyrus (MOG) [29] was reported to be lower in children

with Autism compared to controls and this might explain the difficulty in facial information processing for subjects with Autism. Other regions involved in Autism include caudate and thalamus. Middle temporal gyrus (MTG) is implicated in speech processing, theory of mind and memory encoding and has also been shown to be affected in ASD [82, 87]. The features we identified not only had predictive power, but also had significant group differences across all the 3 splits accounting for variations in age and acquisition site. Therefore, it is likely that these features are robust to age changes and variations in acquisition sites. This factor is especially crucial given the atypical developmental trajectories in ASD. Therefore, unlike results reported by other studies which may have considered narrow age ranges, the connectivity paths we identified are reliable across age variations, though further study is necessary to confirm our findings about the age invariance of disease encoding paths and regions involved in ASD.

3.4.8 ADHD

For ADHD, we report accuracies of 57.2% and 54.1% for binary and multiclass classification, respectively (Table 3.24). Although we did not perform the classification strictly according to the ADHD-200 competition guidelines [31], it is still crucial to examine the results obtained from the competition because it elicited a response from several research institutions to work on a common dataset. The winning team for the competition from Johns Hopkins University reported classification results on the hold-out test dataset release by the competition with a specificity of 94% and a sensitivity of 21% using a weighted combination of several algorithms [88]. Most teams reported hold-out test accuracies in the range of 37.4-60.5%, which are similar to those obtained by us. In fact using just phenotypic data allowed a team from University of Alberta to achieve a higher accuracy (62.5%) than using neuroimaging based metrics [53]. Combining phenotypic data with imaging data helped several groups to achieve higher accuracies than using

imaging data alone [89, 90]. Using ADHD-200 data, Colby et al. reported using site-specific classifiers and suggested that the top features varied across sites, and classifiers trained with data across imaging sites performed worse than classifiers trained using data from the same imaging site [90]. Similar to our results, none of their classifiers performed well for the 3-way classification between Controls, ADHD-I, and ADHD-C. Though these accuracies were above chance levels, they still highlight the challenges encountered in neuroimaging based metrics from multisite acquisitions [31]. By combining structural, functional and demographic information, an accuracy of 55% with 33% sensitivity and 80% specificity was achieved [90]. Many studies reported higher accuracies classifying the disease subtypes ADHD i.e. ADHD-I from ADHD-C than the between Controls and ADHD [46, 90, 88]. This result is surprising given that we expect children with ADHD subtypes to be more similar to each other than with healthy controls. It is not clear at this stage whether it is due to overdiagnosis of ADHD, or if it has some neurological basis, or if it is just an artifact of the peculiarities of the ADHD-200 data. Some studies achieved higher performance of 80-85% using LOOCV and Regional Homogeneity (ReHo) features in a relatively small sample (20-46 subjects) of age-matched populations [91, 92]. Using the entire dataset and Artificial Neural Networks (ANN) based on deep learning architectures, LOOCV accuracies of 80% have been reported in classifying Controls from ADHD-I and Controls from ADHD-C, and 95% in classifying the ADHD subtypes [46, 90]. Since the ADHD-200 competition closely resembles real world classification scenarios, the challenges in classification encountered in this dataset, will apply to future studies utilizing multisite acquisitions.

From our results as well as from those reported previously, it is apparent that ADHD is characterized by large-scale disruptions in connectivity in the frontal and the temporal lobes. We

did not find a lot of overlap between the connectivity paths for the two-way and the multiclass classification though roughly the same brain regions appear to be involved in both classification schemes (Figure 3.20, Table 3.13). In fact, one of the top regions associated with changes in functional connectivity is the dorsal region of the anterior cingulate cortex (d-ACC). It is one of the most critical nodes involved in ADHD, playing a key role in attention [93, 94]. Anterior cingulate cortex (ACC) and insula are part of the salience network and have been previously implicated in ADHD [95]. This result is not surprising as these regions are involved in attention and control [96]. Dorsolateral prefrontal cortex (DLPFC), anterior prefrontal cortex (aPFC) and caudate are part of the executive control network and these regions along with the supplementary motor area (SMA) are involved in attentional control [97, 98]. Along with these networks, DMN also plays a crucial role in ADHD [99, 100, 98, 101, 102]. Several studies have demonstrated the role of the frontal cortex, caudate, basal ganglia, insula, and cingulate gyrus in ADHD [103, 104, 105, 106, 107, 108, 109]. The connections between the nodes in the frontal cortex and basal ganglia form a part of the frontal–striatal network which is involved in response inhibition [110, 111, 112] with inferior frontal gyrus (IFG) playing an especially important role in salience processing and initiation of the response inhibition signal. Though several networks such as salience network, executive control network and default mode network are implicated in ADHD, only a subset of connections between the regions seemed to have predictive power as well as statistical separation as our results indicate (Figure 3.17). In accordance with our results connectivity between IFG, ACC, superior frontal gyrus (SFG), and temporal regions have been reported to be altered in ADHD [113]. There is also growing evidence of temporal lobe as a key area for ADHD [114, 115, 116], though further studies might be needed to support our findings. Our results are in general conformity with prior results discussed above.

3.4.9 PCS & PTSD

We achieved excellent performance in classifying subjects with PTSD from Controls in age-matched training/validation and hold-out test data than in the unmatched scenario (Figures 3.18-3.21, Table 3.24). This result underscores the issues with overfitting the data. Unfortunately, there are not many studies which used RSFC or RSFC- derived metrics for classification of PTSD. However, the few studies which looked at classification of PTSD using RSFC indicate that by integrating multiple features, higher accuracies can be achieved. Using features derived from both RSFC and Amplitude of Low-Frequency Fluctuations (ALFF), Liu et al obtained cross-validation accuracies of 92.5%, an increase of 17.5% in the cross-validation accuracy compared to using just ALFF in a sample containing 40 subjects [117]. Using gray matter volume from structural MRI, as well as ALFF and regional homogeneity from Rs-fMRI, a LOOCV accuracy of 90% was obtained in classifying controls from PTSD using a multi-kernel SVM classifier in a sample containing 37 trauma exposed subjects [118].

Some of the most important regions associated with PTSD classification which we obtained (shown in Figure 3.23, Table 3.18), such as right superior frontal gyrus, cingulate gyrus, right middle temporal gyrus, calcarine fissure and lingual gyrus, have been reported to have alterations in PTSD before [119, 118, 117]. Several of our top classification paths involved regions such as middle occipital gyrus (MOG), angular gyrus, cuneus, middle temporal gyrus (MTG), [120] cingulate gyrus (CG), calcarine fissure, and occipital cortex [117]. Many functional connectivity paths in the visual areas were observed in our study, and is in agreement with previous reports of such alterations in PTSD [121, 122, 123, 117, 124]. These alterations may be associated with visual imagery in PTSD [125]. Increased activity in the superior frontal gyrus and middle temporal gyrus might be linked to anxiety and have been shown to be affected in PTSD [126].

Regions identified in our study such as middle cingulate cortex, thalamus are some of the regions reported to be affected by PTSD along with some other regions not identied as important such as hippocampus, putamen, amygdala, insula, orbitofrontal cortex (OFC) and anterior cingulate cortex (ACC) [127, 128, 129, 121, 130, 131, 132]. Our results are in general conformity with prior results discussed above.

3.4.10 MCI and AD

In previous classification studies of MCI and AD, integration of imaging modalities such as Diffusion Tensor Imaging (DTI) and Rs-fMRI achieved a much higher cross-validation accuracy of 96.3% than Rs-fMRI alone, which achieved only 70.37% in cross-validation accuracy in a dataset of 27 subjects [133]. Even when classifying healthy controls from patients with AD, a relatively lower cross-validation accuracy of 74% for Rs-fMRI was achieved using a dataset containing 43 subjects [134]. Employing the same dataset by integrating multiple imaging modalities such as DTI, Rs-fMRI and Gray Matter (GM) volume, a much higher cross-validation accuracy of 85% was reported [134]. This result is similar to our results in the age-matched split in which we achieved a hold-out test accuracy of 76.9% and a balanced hold-out test accuracy of 78.6% (Table 3.24). In an age-matched sample of 40 subjects, using graph theory-based metrics derived from Rs-fMRI data, Khazaee et al. achieved a LOOCV accuracy of 100% in classifying patients with Alzheimer's disease using Linear-SVM [135]. In a sample containing 27 subjects with AD, 50 with MCI and 30 controls, using Bayesian Gaussian process logistic regression (GP-LR) model, Challis et al. achieved an accuracy of 75% in separating healthy controls from MCI and 97% in separating MCI from AD on a hold-out test data [136]. Using network-based measures several studies obtained a LOOCV accuracy in the range of 86% to 92% in separating in controls from MCI [137, 138, 139] on a dataset with 12 subjects with MCI and 25 healthy

controls . Similar to our results, a study using structural MRI, Positron Emission tomography (PET) and cerebrospinal fluid (CSF) data from ADNI does indicate the relative ease in separating healthy controls from Alzheimer's than from MCI and healthy controls [140].

Our results (Figure 3.29) indicate that the connectivity paths between fusiform gyrus and insula, cuneus and inferior frontal gyrus seem to be most important in binary and multiclass classification of early and late MCI and Alzheimer's disease. Since the data size was small for this dataset, very few paths crossed significance for both age-unmatched and the age-matched split. Since the features we observe are a subset of features which satisfy 3 criteria: (i) Robust to effects of age (ii) High predictive ability (iii) Significant group difference, very few features are reported for this dataset. So fusiform gyrus is associated with visual cognition and plays a key role in MCI and AD [141]. Insula, on the other hand, is associated with perception, cognition, emotion and self-awareness [142, 143, 144] and has been implicated in Alzheimer's disease as well [145, 146, 147]. We found several brain regions in the temporal lobe (Table 3.23) to be affected in Alzheimer's disease, including the hippocampus, temporal pole, parahippocampal gyrus. These regions are involved in memory related processes [147] and have been implicated in AD before [148, 149, 150, 151, 152]. Along with regions in the temporal gyrus, other regions with discriminative ability in MCI as reported in other studies include, insula, precuneus and inferior frontal gyrus (IFG) [153, 154, 133]. Given that the regions involved in functional connectivity paths are in general conformity with the existing results, it is likely that the few connectivity paths we identified might have large discriminative ability and is robust to variations in age.

3.5 Acknowledgements

We would like to acknowledge the contributions of International Neuroimaging Data –Sharing Initiative (INDI), the organizers of the International ADHD-200 competition and Neurobureau for providing us with access to the ADHD neuroimaging data. We also used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. We would also like to acknowledge the researchers and agencies that contributed to the ABIDE database. Finally, the authors acknowledge financial support for PTSD/PCS data acquisition from the U.S. Army Medical Research and Material Command (MRMC) (Grant # 00007218). The views, opinions, and/or findings from PTSD/PCS data contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the U.S. Army or the Department of Defense (DoD) or the United States Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank the personnel at the TBI clinic and behavioral health clinic, Fort Benning, GA, USA and the US Army Aeromedical Research Laboratory, Fort Rucker, AL, USA, and most of all, the Soldiers who participated in the study. The authors thank Julie Rodiek and Wayne Duggan for facilitating PTSD data acquisition.

3.6 Bibliography

[1] C. Kelly, B. B. Biswal, R. C. Craddock, F. X. Castellanos and M. P. Milham, "Characterizing variation in the functional connectome: promise and pitfalls," *Trends in Cognitive Sciences,* vol. 16, no. 3, p. 181–188, 2012.

[2] W. Huf, K. Kalcher, R. N. Boubela, G. Rath, A. Vecsei, P. Filzmoser and E. Moser, "On the generalizability of resting-state fMRI machine learning classifiers," *Frontiers in Human Neuroscience,* vol. 8, p. 502, 2014.

[3] J. A. Nielsen, B. A. Zielinski, P. T. Fletcher, A. L. Alexander, N. Lange, E. D. Bigler, J. E. Lainhart and J. S. Anderson, "Multisite functional connectivity MRI classification of autism: ABIDE results," *Frontiers in Human Neuroscience,* vol. 7, p. 599, 2013.

[4] C. P. Chen, C. L. Keown, A. Jahedi, A. Nair, M. E. Pflieger, B. A. Bailey and R.-A. Müller, "Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism," *NeuroImage: Clinical,* vol. 8, pp. 238-245, 2015.

[5] K. R. Foster, R. Koprowski and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research - commentary," *BioMedical Engineering OnLine,* vol. 13, no. 1, p. 94, 2014.

[6] M. R. Arbabshirani, S. Plis, J. Sui and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage,* p. inpress, 2016.

[7] O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl, G. D. Pearlson and V. D. Calhoun, "A Review of Challenges in the Use of fMRI for Disease Classification / Characterization and A Projection Pursuit Application from A Multi-site fMRI Schizophrenia Study," *Brain Imaging and Behavior,* vol. 2, no. 3, p. 207–226, 2008.

[8] B. Horwitz and J. B. Rowe, "Functional biomarkers for neurodegenerative disorders based on the network paradigm," *Progress in Neurobiology,* vol. 95, no. 4, pp. 505-509, 2011.

[9] Y.-h. Chou, L. Panych, C. Dickey, J. Petrella and N.-k. Chen, "Investigation of Long-Term Reproducibility of Intrinsic Connectivity Network Mapping: A Resting-State fMRI Study," *Am J Neuroradiol,* vol. 33, pp. 833-838, 2012.

[10] Z. Shehzad, A. M. C. Kelly, . P. T. Reiss, D. G. Gee, K. Gotimer, L. Q. Uddin, S. H. Lee, D. S. Margulies, A. K. Roy, B. B. Biswal, E. Petkova, F. X. Castellanos and M. P. Milham, "The Resting Brain: Unconstrained yet Reliable," *Cerebral Cortex,* vol. 19, no. 10, pp. 2209-2229, 2009.

[11] A. S. Choe, C. K. Jones, S. E. Joel, J. Muschelli, V. Belegu, B. S. Caffo, M. A. Lindquist, P. C. M. van Zijl and J. J. Pekar, "Reproducibility and Temporal Structure in Weekly Resting-State fMRI over a Period of 3.5 Years," *PLoS ONE,* vol. 10, no. 10, p. e0140134, 2015.

[12] C. C. Guo, F. Kurth, J. Zhou, E. A. Mayer, S. B. Eickhoff, J. H. Kramer and W. W. Seeley, "One-year test–retest reliability of intrinsic connectivity network fMRI in older adults," *NeuroImage,* vol. 61, no. 4, pp. 1471-1483, 2012.

[13] J.-H. Wang, X.-N. Zuo, S. Gohel, M. P. Milham, B. B. Biswal and Y. He, "Graph Theoretical Analysis of Functional Brain Networks: Test-Retest Evaluation on Short- and Long-Term Resting-State Functional MRI Data," *PLoS ONE,* vol. 6, no. 7, p. e21976, 2011.

[14] J. Anderson, M. Ferguson, M. Lopez-Larson and D. Yurgelun-Todd, "Reproducibility of Single-Subject Functional Connectivity Measurements," *AJNR,* vol. 32, pp. 548-555, 2011.

[15] R. M. Birn, E. K. Molloy, R. Patriat, T. Parker, T. B. Meier, G. R. Kirk, V. A. Nair, M. E. Meyerand and V. Prabhakaran, "The effect of scan length on the reliability of resting-state fMRI connectivity estimates," *NeuroImage,* vol. 83, pp. 550-558, 2013.

[16] U. Braun, M. M. Plichta, C. Esslinger, C. Sauer, L. Haddad, O. Grimm, D. Mier, S. Mohnke, A. Heinz, S. Erk, H. Walter, N. Seiferth, P. Kirsch and A. Meyer-Lindenberg, "Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures," *NeuroImage,* vol. 59, no. 2, pp. 1404-1412, 2012.

[17] T. Meindl, S. Teipel, R. Elmouden, S. Mueller, W. Koch, O. Dietrich, U. Coates, M. Reiser and C. Glaser, "Test–retest reproducibility of the default-mode network in healthy individuals," *Human Brain Mapping,* vol. 31, no. 2, p. 237–246, 2009.

[18] D. Pinter, C. Beckmann, M. Koini, E. Pirker, N. Filippini, A. Pichler, S. Fuchs, F. Fazekas and C. Enzinger, "Reproducibility of Resting State Connectivity in Patients with Stable Multiple Sclerosis," *PLoS ONE,* vol. 11, no. 3, p. e0152158, 2016.

[19] K. Somandepalli, C. Kelly, P. T. Reiss, X.-N. Zuo, R. Craddock, C.-G. Yan, E. Petkova, F. Castellanos, M. P. Milham and A. Di Martino, "Short-term test–retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder," *Developmental Cognitive Neuroscience,* vol. 15, pp. 83-93, 2015.

[20] R. Marchitelli, L. Minati, M. Marizzoni, B. Bosch, D. Bartrés-Faz, B. W. Müller, J. Wiltfang, U. Fiedler, L. Roccatagliata, A. Picco, F. Nobili, O. Blin, S. Bombois, R. Lopes, R. Bordet, J. Sein, J.-P. Ranjeva, M. Didic, H. Gros-Dagnac, P. Payoux, G. Zoccatelli, F. Alessandrini, A. Beltramello, N. Bargalló, A. Ferretti, M. C. Caulo, M. Aello, C. Cavaliere, A. Soricelli, L. Parnetti, R. Tarducci, P. Floridi, M. Tsolaki, M. Constantinidis, A. Drevelegas, P. M. Rossini, C. Marra, P. Schönknecht, T. Hensch, K.-T. Hoffmann, J. P. Kuijer, P. J. Visser, F. Barkhof, G. B. Frisoni and J. Jovicich, "Test-retest reliability of the default mode network in a multi-centric fMRI study of healthy elderly: Effects of data-driven physiological noise correction techniques," *Human Brain Mapping,* vol. 37, no. 6, p. 2114–2132, 2016.

[21] P. Orban, C. Madjar, M. Savard, C. Dansereau, A. Tam, S. Das, A. C. Evans, P. Rosa-Neto, J. C. Breitner, P. Bellec and The PREVENT-AD Research Group, "Test-retest resting-state fMRI in healthy elderly persons with a family history of Alzheimer's

disease," *Scientific Data,* vol. 2, p. 150043, 2015.

[22] M. Fiecas, H. Ombao, D. v. Lunen, R. Baumgartner, A. Coimbra and D. Feng, "Quantifying temporal correlations: A test–retest evaluation of functional connectivity in resting-state fMRI," *NeuroImage,* vol. 65, pp. 231-241, 2013.

[23] X. Liang, J. Wang, C. Yan, N. Shu, K. Xu, G. Gong and Y. He, "Effects of Different Correlation Metrics and Preprocessing Factors on Small-World Brain Functional Networks: A Resting-State Functional MRI Study," *PLoS ONE,* vol. 7, no. 3, p. e32766, 2012.

[24] L. M. Shah, J. A. Cramer, M. A. Ferguson, R. M. Birn and J. S. Anderson, "Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state," *Brain and Behavior,* vol. 6, no. 5, pp. 2162-3279, 2016.

[25] Centers for Disease Control and Prevention, "Prevalence of autism spectrum disorder among children aged 8 years -autism and developmental disabilities monitoring network, 11 Sites, united states, 2010," *MMWR Surveillance Summaries,* vol. 63, no. 2, pp. 1-21, 2014.

[26] American Psychiatric Association, Diagnostic and statistical manual of mental disorders (5th ed.), Arlington, VA: American Psychiatric Publishing, 2013.

[27] J. O. Maximo, E. J. Cadena and R. K. Kana, "The Implications of Brain Connectivity in the Neuropsychology of Autism," *Neuropsychol. Rev.,* vol. 24, no. 1, p. 16–31, 2014.

[28] S. J. Gotts, W. K. Simmons, L. A. Milbury, G. L. Wallace, R. W. Cox and A. Martin, "Fractionation of social brain circuits in autism spectrum disorders," *Brain,* vol. 135, no. 9, pp. 2711-2725, 2012.

[29] A. Di Martino, . C.-G. Yan, Q. Li, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky and M. P. Milham, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry,* vol. 19, pp. 659-667, 2014.

[30] S. N. Visser, M. L. Danielson, R. H. Bitsko, J. R. Holbrook, M. D. Kogan, R. M. Ghandour, R. Perou and S. J. Blumberg, "Trends in the Parent-Report of Health Care Provider-Diagnosed and Medicated Attention-Deficit/Hyperactivity Disorder: United States, 2003–2011," *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 53, no. 1, pp. 34-46, 2014.

[31] The ADHD-200 Consortium, "The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience," *Front. Syst. Neurosci.,* vol. 6, no. 62, 2012.

[32] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas and E. E. Walters, "Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication," *Arch Gen Psychiatry,* vol. 62, no. 6, pp. 593-602, 2005.

[33] H. K. Kang, B. H. Natelson, C. M. Mahan, K. Y. Lee and F. M. Murphy, "Post-Traumatic Stress Disorder and Chronic Fatigue Syndrome-like Illness among Gulf War Veterans: A Population-based Survey of 30,000 Veterans," *Am. J. Epidemiol.,* vol. 157, no. 2, pp. 141-148, 2003.

[34] T. Tanielian and L. (. Jaycox, "Invisible Wounds of War: Psychological and Cognitive Injuries, Their Consequences, and Services to Assist Recovery," RAND Corporation, Santa Monica, CA:, 2008.

[35] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, J. L. Cummings, M. d. Leon, H. Feldman, M. Ganguli, H. Hampel, P. Scheltens, M. C. Tierney, P. Whitehouse and B. Winblad, "Mild cognitive impairment," *The Lancet,* vol. 67, no. 9518, pp. 1262 - 1270, 2006.

[36] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies and C. H. Phelps, "The diagnosis of mild cognitive impairment due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & Dementia,* vol. 7, no. 3, pp. 270-279, 2011.

[37] C.-G. Yan and Y.-F. Zang, "DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI," *Frontiers in Systems Neuroscience,* vol. 4, p. 13, 2010.

[38] G.-R. Wu, W. Liao, S. Stramaglia, J.-R. Ding, H. Chen and D. Marinazzo, "A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data," *Medical Image Analysis,* vol. 17, no. 3, pp. 365-374, 2013.

[39] R. C. Craddock, G. James, P. E. Holtzheimer, X. P. Hu and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Human Brain Mapping,* vol. 33, no. 8, pp. 1914-1928, 2012.

[40] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. R. Andersson, L. Griffanti, G. Douaud, T. W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. M. Matthews and S. M. Smith, "Multimodal population brain imaging in the UK Biobank

prospective epidemiological study," *Nature Neuroscience,* vol. 19, p. 1523–1536, 2016.

[41] B. Mwangi, T. S. Tian and J. C. Soares, "A Review of Feature Reduction Techniques in Neuroimaging," *Neuroinformatics,* vol. 12, no. 2, p. 229–244, 2014.

[42] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning,* vol. 7, no. 8, pp. 1157-1182, 2003.

[43] F. Pereira, T. Mitchell and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage,* vol. 45, no. 1, p. S199–S209, 2009.

[44] R. C. Craddock, P. E. Holtzheimer, X. P. Hu and H. S. Mayberg, "Disease state prediction from resting state functional connectivity," *Magn. Reson. Med.,* vol. 62, p. 1619–1628, 2009.

[45] G. Deshpande, Z. Li, P. Santhanam, C. D. Coles, M. E. Lynch, S. Hamann and X. Hu, "Recursive Cluster Elimination Based Support Vector Machine for Disease State Prediction Using Resting State Functional and Effective Brain Connectivity," *PLoS ONE,* vol. 5, no. 12, p. e14277, 2010.

[46] G. Deshpande, P. Wang, D. Rangaprakash and B. Wilamowski, "Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data," *IEEE Transactions on Cybernetics,* vol. 45, no. 12, pp. 2668 - 2679, 2015.

[47] G. Deshpande, L. E. Libero, K. R. Sreenivasan, H. D. Deshpande and R. K. Kana, "Identification of neural connectivity signatures of autism using machine learning," *Front. Hum. Neurosci.,* vol. 7, p. 670, 2013.

[48] M. Yousef, S. Jung, L. C. Showe and M. K. Showe, "Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data," *BMC Bioinformatics,* vol. 8, no. 1, p. 144, 2007.

[49] R. Rao, G. Fung and R. Rosales, "On the Dangers of Cross-Validation. An Experimental Evaluation," in *Proceedings of the 2008 SIAM International Conference on Data Mining*, 2008.

[50] J. S. Anderson, J. A. Nielsen, A. L. Froehlich, M. B. DuBray, T. J. Druzgal, A. N. Cariello, J. R. Cooperrider, B. A. Zielinski, C. Ravichandran, P. T. Fletcher, A. L. Alexander, E. D. Bigler, N. Lange and J. E. Lainhart, "Functional connectivity magnetic resonance imaging classification of autism," *Brain,* vol. 134, pp. 3742-3754, 2011.

[51] E. Edgington, Randomization Tests, New York: Marcel Dekker, Inc., 1980.

[52] M. Xia, J. Wang and Y. He, "BrainNet Viewer: A Network Visualization Tool for Human

Brain Connectomics," *PLOS ONE,* vol. 8, no. 7, p. e68910, 2013.

[53] M. R. G. Brown, G. S. Sidhu, R. Greiner, N. Asgarian, M. Bastani, P. H. Silverstone, A. J. Greenshaw and S. M. Dursun, "ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements," *Front. Syst. Neurosci.,* vol. 6, p. 69, 2012.

[54] J. Sato, M. Hoexter, A. Fujita and R. Luis, "Evaluation of pattern recognition and feature extraction methods in ADHD prediction," *Frontiers in Systems Neuroscience,* vol. 6, p. 68, 2012.

[55] L. Q. Uddin, K. Supekar, C. J. Lynch, A. Khouzam, J. Phillips, C. Feinstein, S. Ryali and V. Menon, "Salience Network–Based Classification and Prediction of Symptom Severity in Children With Autism," *JAMA Psychiatry,* vol. 70, no. 8, pp. 869-879, 2013.

[56] J. Zhou, M. D. Greicius, E. D. Gennatas, M. E. Growdon, J. Y. Jang, G. D. Rabinovici, J. H. Kramer, M. Weiner, B. L. Miller and W. W. Seeley, "Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease," *Brain,* vol. 133, no. 5, pp. 1352-1367, 2010.

[57] W. Koch, S. Teipel, S. Mueller, J. Benninghoff, M. Wagner, A. L. Bokde, H. Hampel, U. Coates, M. Reiser and T. Meindl, "Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease," *Neurobiology of Aging,* vol. 33, no. 3, pp. 466-478, 2012.

[58] F. X. Castellanos, A. Di Martino, R. C. Craddock, A. D. Mehta and M. P. Milham, "Clinical applications of the functional connectome," *NeuroImage,* vol. 80, pp. 527-540, 2013.

[59] A. Isaksson, M. Wallman, H. Göransson and M. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognition Letters,* vol. 29, no. 14, pp. 1960-1965, 2008.

[60] G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Journal of Machine Learning Research,* vol. 11, pp. 2079-2107, 2010.

[61] C. J. Brown and G. Hamarneh, "Machine Learning on Human Connectome Data from MRI," 2016.

[62] A. Venkataraman, M. Kubicki, C. F. Westin and . P. Golland, "Robust feature selection in resting-state fMRI connectivity based on population studies," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, 2010.

[63] N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nature Neuroscience,* vol. 12, pp. 535-540, 2009.

[64] A. Vezhnevets and O. Barinova, "Avoiding Boosting Overfitting by Removing Confusing Samples," in *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings*, J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič and A. Skowron, Eds., Berlin, Heidelberg, Springer Berlin Heidelberg, 2007, pp. 430-441.

[65] A. J. Grove and D. Schuurmans, "Boosting in the limit: Maximizing the margin of learned ensembles," in *In Proc. of the Fifteenth National Conference on Artifical Intelligence*, 1998.

[66] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga and . L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)," *Alzheimer's & Dementia,* vol. 1, no. 1, pp. 55-66, 2005.

[67] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, A.-M. Dogonowski, M. Ernst, D. Fair, M. Hampson, M. J. Hoptman, J. S. Hyde, V. J. Kiviniemi, R. Kötter, S.-J. Li, C.-P. Lin, M. J. Lowe, C. Mackay, D. J. Madden, K. H. Madsen, D. S. Margulies, H. S. Mayberg, K. McMahon, C. S. Monk, S. H. Mostofsky, B. J. Nagel, J. J. Pekar, S. J. Peltier, S. E. Petersen, V. Riedl, S. A. R. B. Rombouts, B. Rypma, B. L. Schlaggar, S. Schmidt, R. D. Seidler, G. J. Siegle , C. Sorg, G.-J. Teng, J. Veijola, A. Villringer, M. Walter, L. Wang, X.-C. Weng, S. Whitfield-Gabrieli , P. Williamson, C. Windischberger, Y.-F. Zang, H.-Y. Zhang, F. X. Castellanos and M. P. Milham, "Toward discovery science of human brain function," *PNAS,* vol. 107, no. 10, p. 4734–4739, 2010.

[68] M. Mennes, B. B. Biswal, F. X. Castellanos and M. P. Milham, "Making data sharing work: The FCP/INDI experience," *NeuroImage,* vol. 82, pp. 683-691, 2013.

[69] X. Zhao, D. Rangaprakash, D. N. Dutt and G. Deshpande, "Investigating the correspondence of clinical diagnostic grouping with underlying neurobiological and phenotypic clusters using unsupervised learning: An application to the Alzheimer's spectrum," in *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM*, Singapore, 2016.

[70] D. Gamberger, B. Ženko, A. Mitelpunkt, N. Shachar and N. Lavrač, "Clusters of male and female Alzheimer's disease patients in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database," *Brain Informatics,* vol. 3, no. 3, p. 169–179, 2016.

[71] D. A. Fair, J. T. Nigg, S. Iyer, D. Bathula, K. L. Mills, N. U. F. Dosenbach, B. L. Schlaggar, M. Mennes, D. Gutman, S. Bangaru, J. K. Buitelaar, D. P. Dickstein, A. Di

Martino, D. N. Kennedy, C. Kelly, B. Luna, J. B. Schweitzer, K. Velanova, Y.-F. Wang, S. Mostofsky, F. X. Castellanos and M. P. Milham, "Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data," *Front. Syst. Neurosci.,* vol. 6, p. 80, 2013.

[72] G. Deshpande, S. LaConte, G. A. James, S. Peltier and X. Hu, "Multivariate Granger Causality Analysis of fMRI Data," *Human Brain Mapping,* vol. 30, p. 1361–1373, 2009.

[73] Y. Wang, S. Katwal, B. Rogers, J. Gore and G. Deshpande, "Experimental Validation of Dynamic Granger Causality for Inferring Stimulus-evoked Sub-100ms Timing Differences from fMRI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* p. (in press), 2016.

[74] T. Price, C.-Y. Wee, W. Gao and D. Shen, "Multiple-network classification of childhood autism using functional connectivity dynamics," in *Med Image Comput Comput Assist Interv.-MICCAI 2014*, Boston, MA, 2014.

[75] L. E. Libero, T. P. DeRamus, A. C. Lahti, G. Deshpande and R. K. Kana, "Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates," *Cortex,* vol. 66, pp. 46-59, 2015.

[76] S. Vigneshwaran, B. S. Mahanand, S. Suresh and N. Sundararajan, "Using regional homogeneity from functional MRI for diagnosis of ASD among males," in *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, 2015.

[77] J. P. Gentile, R. Atiq and P. M. Gillig, "Adult ADHD: Diagnosis, Differential Diagnosis, and Medication Management," *Psychiatry (Edgmont),* vol. 3, no. 8, p. 25–30, 2006.

[78] L. Q. Uddin, K. Supekar and V. Menon, "Reconceptualizing functional brain connectivity in autism from a developmental perspective," *Front. Hum. Neurosci,* vol. 7, p. 458 , 2013.

[79] M. Plitt, K. A. Barnes and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage: Clinical,* vol. 7, pp. 359-366, 2015.

[80] H. Chen, X. Duan, F. Liu, F. Lu, X. Ma, Y. Zhang, L. Q. Uddin and H. Chen, "Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivity—A multi-center study," *Progress in Neuro-Psychopharmacology and Biological Psychiatry,* vol. 64, p. 1–9, 2016.

[81] T. Iidaka, "Resting state functional magnetic resonance imaging and neural network classified autism and control," *Cortex,* vol. 63, pp. 55-67, 2015.

[82] W. Cheng, E. T. Rolls, H. Gu, J. Zhang and J. Feng, "Autism: reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self,"

Brain, vol. 138, pp. 1382-1393, 2015.

[83] M. Assaf, K. Jagannathan, V. D. Calhoun, L. Miller, M. C. Stevens, R. Sahl, J. G. O'Boyle, R. T. Schultz and G. D. Pearlson, "Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients," *NeuroImage,* vol. 53, no. 1, pp. 247-256, 2010.

[84] C. S. Monk, S. J. Peltier, J. L. Wiggin, S.-J. Weng, M. Carrasco, S. Risi and C. Lord, "Abnormalities of intrinsic functional connectivity in autism spectrum disorders," *NeuroImage,* vol. 47, no. 2, pp. 764-772, 2009.

[85] S. D. Washington, E. M. Gordon, J. Brar, S. Warburton, A. T. Sawyer, A. Wolfe, E. R. Mease-Ference, L. Girton, A. Hailu, J. Mbwana, W. D. Gaillard, M. L. Kalbfleisch and J. W. VanMeter, "Dysmaturation of the default mode network in autism," *Human Brain Mapping,* vol. 35, no. 4, pp. 1284-1296, 2014.

[86] P. Mundy, "Annotation: The neural basis of social impairments in autism: the role of the dorsal medial-frontal cortex and anterior cingulate system," *Journal of Child Psychology and Psychiatry,* vol. 44, no. 6, p. 793–809, 2003.

[87] C. H. Salmond, J. Ashburner, A. Connelly, K. J. Friston, D. G. Gadian and F. Vargha-Khadem, "The role of the medial temporal lobe in autistic spectrum disorders," *European Journal of Neuroscience,* vol. 22, no. 3, pp. 762-772, 2005.

[88] A. Eloyan, J. Muschelli, M. B. Nebel, H. Liu, F. Han, T. Zhao, A. D. Barber, S. Joel, J. J. Pekar, S. H. Mostofsky and B. Caffo, "Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging," *Front. Syst. Neurosci.,* vol. 6, p. 61, 2012.

[89] G. S. Sidhu, N. Asgarian, R. Greiner and M. R. G. Brown, "Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD," *Front. Syst. Neurosci.,* vol. 6, p. 74, 2012.

[90] J. B. Colby, J. D. Rudie, J. A. Brown, P. K. Douglas, M. S. Cohen and Z. Shehzad, "Insights into multimodal imaging classification of ADHD," *Front. Syst. Neurosci.,* vol. 6, p. 59, 2012.

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

[92] X. Wang, Y. Jiao, T. Tang, H. Wang and Z. Lu, "Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder," *European Journal of Radiology,* vol. 82, no. 9, pp. 1552-1557, 2013.

[93] B. J. Casey, R. Trainor, J. Giedd, Y. Vauss, C. K. Vaituzis, S. Hamburger, P. Kozuch and J. L. Rapoport, "The role of the anterior cingulate in automatic and controlled processes: A developmental neuroanatomical study," *Dev. Psychobiol.,* vol. 30, pp. 61-69, 1997.

[94] G. Bush, J. A. Frazier, S. L. Rauch, L. J. Seidman, P. J. Whalen, M. A. Jenike, B. R. Rosen and J. Biederman, "Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fMRI and the counting stroop," *Biological Psychiatry,* vol. 45, no. 12, pp. 1542-1552, 1999.

[95] M. P. Lopez-Larson, J. B. King, J. Terry, E. C. McGlade and D. Yurgelun-Todd, "Reduced insular volume in attention deficit hyperactivity disorder," *Psychiatry Research: Neuroimaging,* vol. 204, no. 1, pp. 32-39, 2012.

[96] V. Menon and L. Q. Uddin, "Saliency, switching, attention and control: a network model of insula function," *Brain Structure and Function,* vol. 214, no. 5, pp. 655-667, 2010.

[97] F. X. Castellanos and E. Proal, "Large-scale brain systems in ADHD: beyond the prefrontal–striatal model," *Trends in Cognitive Sciences,* vol. 16, no. 1, pp. 17-26, 2012.

[98] A. Elton, S. Alcauter and W. Gao, "Network connectivity abnormality profile supports a categorical-dimensional hybrid model of ADHD," *Human Brain Mapping,* vol. 35, no. 9, p. 4531–4543, 2014.

[99] K. Konrad and S. B. Eickhoff, "Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder," *Human Brain Mapping,* vol. 31, no. 6, p. 904–916, 2010.

[100] D. Tomasi and N. D. Volkow, "Abnormal Functional Connectivity in Children with Attention-Deficit/Hyperactivity Disorder," *Biological Psychiatry,* vol. 71, no. 5, pp. 443-450, 2012.

[101] F. X. Castellanos and Y. Aoki, "Intrinsic Functional Connectivity in Attention-Deficit/Hyperactivity Disorder: A Science in Development," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,* vol. 1, no. 3, pp. 253-261, 2016.

[102] J. C. Mostert, E. Shumskaya, M. Mennes, A. M. H. Onnink, M. Hoogman, C. C. Kan, A. A. Vasquez, J. Buitelaar, B. Franke and D. G. Norris, "Characterising resting-state functional connectivity in a large sample of adults with ADHD," *Progress in Neuro-Psychopharmacology and Biological Psychiatry,* vol. 67, pp. 82-91, 2016.

[103] E. H. Aylward, A. L. Reiss, M. J. Reader, H. S. Singer, J. E. Brown and M. B. Denckla, "Basal Ganglia Volumes in Children With Attention-Deficit Hyperactivity Disorder," *J Child Neurol.,* vol. 11, no. 2, pp. 112-115, 1996.

[104] A. Garrett, L. Penniman, J. N. Epstein, B. Casey, S. P. Hinshaw, G. Glover, S. Tonev, A. Vitolo, M. Davidson, J. Spicer, L. L. Greenhill and A. L. Reiss, "Neuroanatomical

Abnormalities in Adolescents With Attention-Deficit/Hyperactivity Disorder," *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 47, no. 11, pp. 1321-1328, 2008.

[105] K. Rubia, S. Overmeyer, E. Taylor, M. Brammer, S. C. Williams, A. Simmons and E. T. Bullmore, "Hypofrontality in Attention Deficit Hyperactivity Disorder During Higher-Order Motor Control: A Study With Functional MRI," *American Journal of Psychiatry,* vol. 156, no. 6, pp. 891-896, 1999.

[106] M.-g. Qiu, Z. Ye, Q.-y. Li, G.-j. Liu, B. Xie and J. Wang, "Changes of Brain Structure and Function in ADHD Children," *Brain Topography,* vol. 24, no. 3, p. 243–252, 2011.

[107] S. G. Dickstein, K. Bannon, F. Xavier Castellanos and M. P. Milham, "The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis," *Journal of Child Psychology and Psychiatry,* vol. 47, no. 10, pp. 1051-1062, 2006.

[108] L. Tian, T. Jiang, Y. Wang, Y. Zang, Y. He, M. Liang, M. Sui, Q. Cao, S. Hu, M. Peng and Y. Zhuo, "Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder," *Neuroscience Letters,* vol. 400, no. 1-2, pp. 39-43, 2006.

[109] S. Cortese, C. Kelly, C. Chabernaud, E. Proal, A. Di Martino, M. P. Milham and F. X. Castellanos, "Toward Systems Neuroscience of ADHD: A Meta-Analysis of 55 fMRI Studies," *American Journal of Psychiatry,* vol. 169, no. 10, pp. 1038-1055, 2012.

[110] A. R. Aron and R. A. Poldrack, "Cortical and Subcortical Contributions to Stop Signal Response Inhibition: Role of the Subthalamic Nucleus," *Journal of Neuroscience,* vol. 26, no. 9, pp. 2424-2433, 2006.

[111] A. Cubillo, R. Halari, C. Ecker, V. Giampietro, E. Taylor and K. Rubia, "Reduced activation and inter-regional functional connectivity of fronto-striatal networks in adults with childhood Attention-Deficit Hyperactivity Disorder (ADHD) and persisting symptoms during tasks of motor inhibition and cognitive switching," *Journal of Psychiatric Research,* vol. 44, no. 10, pp. 629-639, 2010.

[112] N. Makris, J. Biederman, M. C. Monuteaux and L. J. Seidman, "Towards Conceptualizing a Neural Systems-Based Anatomy of Attention-Deficit/Hyperactivity Disorder," *Dev Neurosci.,* vol. 31, p. 36–49, 2009.

[113] D. v. Rooij, C. A. Hartman, M. Mennes, J. Oosterlaan, B. Franke, N. Rommelse, D. Heslenfeld, S. V. Faraone, J. K. Buitelaar and P. J. Hoekstraa, "Altered neural connectivity during response inhibition in adolescents with attention-deficit/hyperactivity disorder and their unaffected siblings," *NeuroImage: Clinical,* vol. 7, pp. 325-335, 2016.

[114] M. Kobel, N. Bechtel, K. Specht, M. Klarhöfer, P. Weber, K. Scheffler, K. Opwis and I.-

K. Penner, "Structural and functional imaging approaches in attention deficit/hyperactivity disorder: Does the temporal lobe play a key role?," *Psychiatry Research: Neuroimaging,* vol. 183, no. 3, pp. 230-236, 2010.

[115] S. Carmona, O. Vilarroya, A. Bielsa, . V. Trèmols, J. Soliva, M. Rovira, J. Tomàs, C. Raheb, J. Gispert, S. Batlle and A. Bulbena, "Global and regional gray matter reductions in ADHD: A voxel-based morphometric study," *Neuroscience Letters,* vol. 389, no. 2, pp. 88-93, 2005.

[116] E. R. Sowel, P. M. Thompson, S. E. Welcome, A. L. Henkenius, A. W. Toga and B. S. Peterson, "Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder," *The Lancet,* vol. 362, no. 9397, pp. 1699-1707, 2003.

[117] F. Liu, B. Xie, Y. Wang, W. Guo, J.-P. Fouche, Z. Long, W. Wang, H. Chen, M. Li, X. Duan, J. Zhang, M. Qiu and H. Chen, "Characterization of Post-traumatic Stress Disorder Using Resting-State fMRI with a Multi-level Parametric Classification Approach," *Brain Topography,* vol. 28, pp. 221-237, 2015.

[118] Q. Zhang, Q. Wu, H. Zhu, L. He, H. Huang, J. Zhang and W. Zhang, "Multimodal MRI-Based Classification of Trauma Survivors with and without Post-Traumatic Stress Disorder," *Frontiers in Neuroscience,* vol. 10, p. 292, 2016.

[119] Y. Yin, C. Jin, L. T. Eyler, H. Jin, X. Hu, L. Duan, H. Zheng, B. Feng, X. Huang, B. Shan, Q. Gong and L. Li, "Altered regional homogeneity in post-traumatic stress disorder: a restingstate functional magnetic resonance imaging study," *Neuroscience Bulletin,* vol. 28, no. 5, p. 541–549, 2012.

[120] P. Christova, L. M. James, B. E. Engdahl, S. M. Lewis and A. P. Georgopoulos, "Diagnosis of posttraumatic stress disorder (PTSD) based on correlations of prewhitened fMRI data: outcomes and areas involved," *Experimental Brain Research,* vol. 233, no. 9, p. 2695–2705, 2015.

[121] H. Zhu, J. Zhang, W. Zhan, C. Qiu, R. Wu, Y. Meng, H. Cui, X. Huang, T. Li, Q. Gong and W. Zhang, "Altered spontaneous neuronal activity of visual cortex and medial anterior cingulate cortex in treatment-naïve posttraumatic stress disorder," *Comprehensive Psychiatry,* vol. 55, no. 7, pp. 1688-1695, 2014.

[122] J. D. Bremner, M. Narayan, L. H. Staib, S. M. Southwick, T. McGlashan and D. S. Charney, "Neural Correlates of Memories of Childhood Sexual Abuse in Women With and Without Posttraumatic Stress Disorder," *American Journal of Psychiatry,* vol. 156, no. 11, pp. 1787-1795, 1999.

[123] J. Bremner, E. Vermetten, M. Vythilingam, N. Afzal, C. Schmahl, B. Elzinga and D. S. Charney, "Neural correlates of the classic color and emotional stroop in women with abuse-related posttraumatic stress disorder," *Biological Psychiatry,* vol. 55, no. 6, pp. 612-

620, 2004.

[124] L. L. Chao, M. Lenoci and T. C. Neylan, "Effects of post-traumatic stress disorder on occipital lobe function and structure," *NeuroReport,* vol. 23, no. 7, pp. 412-419, 2012.

[125] I. A. Clark and C. E. Mackay, "Mental Imagery and Post-Traumatic Stress Disorder: A Neuroimaging and Experimental Psychopathology Approach to Intrusive Memories of Trauma," *Frontiers in Psychiatry,* vol. 6, p. 104, 2015.

[126] M. C. W. Kroes, M. D. Rugg, M. G. Whalley and C. R. Brewin, "Structural brain abnormalities common to posttraumatic stress disorder and depression," *J Psychiatry Neurosci,* vol. 36, no. 4, pp. 256-265, 2011.

[127] Y. Zhong, R. Zhang, K. Li, R. Qi, Z. Zhang, Q. Huang and G. Lu, "Altered cortical and subcortical local coherence in PTSD: evidence from resting-state fMRI," *Acta Radiol,* vol. 56, no. 6, pp. 746-753, 2015.

[128] R. A. Lanius, P. C. Williamson, R. L. Bluhm, M. Densmore, K. Boksman, R. W. Neufeld, J. S. Gati and R. S. Menon, "Functional connectivity of dissociative responses in posttraumatic stress disorder: A functional magnetic resonance imaging investigation," *Biological Psychiatry,* vol. 57, no. 8, pp. 873-884, 2005.

[129] L. M. Shin, S. P. Orr, M. A. Carson, S. L. Rauch, M. L. Macklin, N. B. Lasko, P. M. Peters, L. J. Metzger, D. D. Dougherty, P. A. Cannistraro, N. M. Alpert, A. J. Fischman and R. K. Pitman, "Regional Cerebral Blood Flow in the Amygdala and Medial PrefrontalCortex During Traumatic Imagery in Male and Female Vietnam Veterans With PTSD," *Arch Gen Psychiatry,* vol. 61, no. 2, pp. 168-176, 2004.

[130] B. Dunkley, S. Doesburg, P. Sedge, R. Grodecki, P. Shek, E. Pang and M. Taylor, "Resting-state hippocampal connectivity correlates with symptom severity in post-traumatic stress disorder," *NeuroImage: Clinical,* vol. 5, pp. 377-384, 2014.

[131] D. Lei, K. Li, L. Li, F. Chen, X. Huang, S. Lui, J. Li, F. Bi and Q. Gong, "Disrupted Functional Brain Connectome in Patients with Posttraumatic Stress Disorder," *Radiology,* vol. 276, no. 3, pp. 818-827, 2015.

[132] L. Li, D. Lei, L. Li, X. Huang, X. Suo, F. Xiao, W. Kuang, J. Li, F. Bi, S. Lui, G. J. Kemp, J. A. Sweeney and Q. Gong, "White Matter Abnormalities in Post-traumatic Stress Disorder Following a Specific Traumatic Event," *EBioMedicine,* vol. 4, pp. 176-183, 2016.

[133] C.-Y. Wee, P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang and D. Shen, "Identification of MCI individuals using structural and functional connectivity networks," *NeuroImage,* vol. 59, no. 3, pp. 2045-2056, 2012.

[134] M. Dyrba, M. Grothe, T. Kirste and S. J. Teipel, "Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM," *Human Brain*

Mapping, vol. 36, no. 6, p. 2118–2131, 2015.

[135] A. Khazaee, A. Ebrahimzadeh and A. Babajani-Feremi, "Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory," *Clinical Neurophysiology,* vol. 126, no. 11, pp. 2132-2141, 2015.

[136] E. Challis, P. Hurley, L. Serra, M. Bozzali, S. Oliver and M. Cercignani, "Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI," *NeuroImage,* vol. 112, pp. 232-243, 2015.

[137] B. Jie, D. Zhang, W. Gao, Q. Wang, C.-Y. Wee and D. Shen, "Integration of Network Topological and Connectivity Properties for Neuroimaging Classification," *IEEE Transactions on Biomedical Engineering,* vol. 61, no. 2, pp. 576-589, 2014.

[138] B. Jie, D. Zhang, C.-Y. Wee and D. Shen, "Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification," *Human Brain Mapping,* vol. 35, no. 7, p. 2876–2897, 2014.

[139] C.-Y. Wee, P.-T. Yap, D. Zhang, L. Wang and D. Shen, "Constrained Sparse Functional Connectivity Networks for MCI Classification," *Med Image Comput Comput Assist Interv.,* vol. 15, no. 2, pp. 212-219, 2012.

[140] D. Zhang, Y. Wang, L. Zhou, H. Yuan and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage,* vol. 55, no. 3, pp. 856-867, 2011.

[141] S. Cai, T. Chong, Y. Zhang, J. Li, K. M. von Deneen, J. Ren, M. Dong, H. Liyu and f. t. A. D. N. Initiative, "Altered functional connectivity of fusiform gyrus in subjects with amnestic mild cognitive impairment: a resting-state fMRI study," *Front. Hum. Neurosci.,* vol. 9, p. 471, 2015.

[142] A. D. Craig, "How do you feel — now? The anterior insula and human awareness," *Nature Reviews Neuroscience,* vol. 10, pp. 59-70, 2009.

[143] H.-O. Karnath, B. Baier and T. Nägele, "Awareness of the Functioning of One's Own Limbs Mediated by the Insular Cortex?," *Journal of Neuroscience,* vol. 25, no. 31, pp. 7134-7138, 2005.

[144] C. Devue, F. Collette, E. Balteau, C. Degueldre, A. Luxen, P. Maquet and S. Brédart, "Here I am: The cortical correlates of visual self-recognition," *Brain Research,* vol. 1143, pp. 169-182, 2007.

[145] A. L. Foundas, C. Leonard, S. M. Mahoney, O. F. Agee and K. M. Heilman, "Atrophy of the Hippocampus, Parietal Cortex, and Insula in Alzheimer's Disease: A Volumetric Magnetic Resonance Imaging Study," *Neuropsychiatry Neuropsychol Behav Neurol.,* vol.

10, no. 2, pp. 81-89, 1997.

[146] G. Karas, P. Scheltens, S. Rombouts, P. Visser, R. van Schijndel, N. Fox and F. Barkhof, "Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease," *NeuroImage,* vol. 23, no. 2, pp. 708-716, 2004.

[147] S. A. Rombouts, F. Barkhof, D. J. Veltman, W. C. Machielsen, M. P. Witter, M. A. Bierlaagha, R. H. Lazeron, J. Valk and P. Scheltens, "Functional MR Imaging in Alzheimer's Disease during Memory Encoding," *AJNR,* vol. 21, pp. 1869-1875, 2000.

[148] G. Allen, H. Barnard, R. McColl, A. Hester, J. Fields, M. Weiner, W. Ringe, A. Lipton, M. Brooker, E. McDonald, C. Rubin and C. Cullum, "Reduced Hippocampal Functional Connectivity in Alzheimer Disease," *Arch Neurol.,* vol. 64, no. 10, pp. 1482-1487, 2007.

[149] L. Wang, Y. Zang, Y. He, M. Liang, X. Zhang, L. Tian, T. Wu, T. Jiang and K. Li, "Changes in hippocampal connectivity in the early stages of Alzheimer's disease: Evidence from resting state fMRI," *NeuroImage,* vol. 31, no. 2, pp. 496-504, 2006.

[150] F. Bai, Z. Zhang, D. R. Watson, H. Yu, Y. Shi, Y. Yuan, Y. Zang, C. Zhu and Y. Qian, "Abnormal Functional Connectivity of Hippocampus During Episodic Memory Retrieval Processing Network in Amnestic Mild Cognitive Impairment," *Biological Psychiatry,* vol. 65, no. 11, pp. 951-958, 2009.

[151] K. A. Celone, V. D. Calhoun, B. C. Dickerson, A. Atri, E. F. Chua, S. L. Miller, K. DePeau, D. M. Rentz, D. J. Selkoe, D. Blacker, M. S. Albert and R. A. Sperling, "Alterations in Memory Networks in Mild Cognitive Impairment and Alzheimer's Disease: An Independent Component Analysis," *Journal of Neuroscience,* vol. 26 , no. 40, pp. 10222-10231, 2006.

[152] C. J. Galton, B. Gomez-Anson, N. Antounb, P. Scheltens, K. Patterson, M. Graves, B. J. Sahakiane and J. R. Hodgesa, "Temporal lobe rating scale: application to Alzheimer's disease and frontotemporal dementia," *J Neurol Neurosurg Psychiatry,* vol. 70, pp. 165-173, 2001.

[153] F.-P. Laia, J. Guàrdia-Olmos and M. Peró-Cebollero, "Mild cognitive impairment and fMRI studies of brain functional connectivity: the state of the art," *Frontiers in Psychology,* vol. 6, p. 1095, 2015.

[154] Z. Dai, C. Yan, Z. Wang, J. Wang, M. Xia, K. Li and Y. He, "Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3)," *NeuroImage,* vol. 59, no. 3, pp. 2187-2195, 2012.

Chapter 4

Conclusion

In this thesis, two current topics in resting state fMRI were studied. In the first part of this thesis after recognizing the confounding effects of head motion artifacts on resting state fMRI, we examined the advantages of prospective motion correction. We observed that using PACE-corrected EPI sequence in combination with retrospective motion correction methods was able to eliminate head motion artifacts, in particular, significant negative motion-BOLD relationships. It has been reported that these significant voxel-wise negative motion-BOLD relationships are typically associated with large signal dropouts, caused by relatively large head movements in the scanner. This finding was important because previous studies which used traditional EPI sequences were only able to eliminate significant voxel-wise negative motion-BOLD relationships after motion censoring. Another significant advantage of PACE as we observed is that PACE-corrected EPI sequence offers a good compromise between data quality and quantity. Even with a liberal censoring strategy and the loss of small amount of data, head motion artifacts were almost eliminated in high motion subjects. Finally, we identify the difficulty in separating neuronal changes in connectivity with head motion artifacts and caution against the use of deconvolution in high motion samples.

In the second part of our thesis, we investigated the effect of data heterogeneity on classification performance in four neuroimaging datasets. We specifically looked at how age and site acquisition variability impact classification accuracy. Our results indicate a significant drop in accuracy when we training a classifier with subjects from an age group/acquisition sites and tested our model against subjects from other age groups/acquisition sites. This drop was dramatic in smaller datasets, probably due to the unreliability of cross-validation for feature selection and performance estimation in the smaller datasets. We implemented a consensus classifier which combines the 18 different classifiers we implemented in a probabilistic manner. This consensus classifier improves the reliability and

robustness of the diagnostic inferences drawn from it. We also identified several connectivity paths and regions associated with the four neurological disease we studied: Autism Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD), Post-Concussion Syndrome (PCS) & Post Traumatic Stress Disorder (PTSD) and Mild Cognitive Impairment (MCI) & Alzheimer's Disease (AD). The connectivity paths and regions we identified are insensitive not only to age and acquisition site but also exhibit significant group differences and excellent discriminability. Finally, we caution against the use of cross-validation especially in smaller datasets and encourage the use of a hold-out dataset or a replication dataset for assessing the diagnostic utility of machine learning classifiers.

Appendix A

A.1 Probabilistic/Bayesian Methods

Probabilistic models introduced in this section are designed to find the Bayes optimal solution. Bayesian classification provides a framework wherein we can calculate a maximum a posteriori estimate (MAP) of the parameters by incorporating prior beliefs about the parameters. The MAP estimate is obtained by the product of the prior belief and the likelihood divided by evidence. In some probabilistic classifiers, by assuming uniform prior distribution, the MAP estimate can be reduced to just computing the maximum likelihood estimate (MLE). In the probabilistic learning framework, by computing the maximum likelihood estimate (MLE), we select parameters of the model which maximizes the probability of the observed data given the parameter.

All classifier models can be divided into two categories: generative and discriminative. Generative classifiers model the joint probability distribution $P(X, Y)$ of the input data X and the output class labels Y and predictions are made via Bayes rule. Discriminative classifier on the other hand, directly learn the mapping between inputs X and outputs Y, and hence the distribution $P(Y|X)$ [1]. Naïve Bayes, Linear and Quadratic Discriminant Analysis are examples of generative models. Logistic Regression is an example of a discriminative model.

Gaussian Naïve Bayes (GNB): In a naïve Bayes (NB) classifier, the class conditional independence of the features is assumed [2]. This implies that given the class, the features are independent of each other. When the features of each class are modeled as univariate Gaussians, we get a GNB classifier. Using the chain rule, the class-conditional features are modeled as a univariate Gaussian function as follows

$$P(x_i|\mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \, exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \quad (1)$$

where μ_i indicates the mean and σ_i indicates the standard deviation of the feature distribution x_i. The parameters of the model (the location and standard deviations of the class conditional Gaussians) are calculated with relative ease from the data. This results in a linear decision boundary which contains points that have an equal likelihood of belong to either class. The final class assignments during prediction are done by calculating the posterior class probabilities, combining the prior class probabilities with class conditional likelihoods, using the Bayes rule.

Linear Discriminant Analysis (LDA): LDA relaxes the class conditional independence of the Naïve Bayes classifiers and models data from each class as a multivariate Gaussian distribution with a covariance matrix that is shared across all the classes (homoscedasticity) [2]. So the input feature distributions *X* of the classes is written as

$$P(X|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(X-\boldsymbol{\mu})^T\Sigma^{-1}(X-\boldsymbol{\mu})\right) \quad (2)$$

where μ indicates the vector containing the feature means and Σ indicates the feature covariance matrix. The parameters of the model are estimated by maximizing the log conditional likelihood. LDA defines a hyperplane that maximizes the ratio of the variance between the classes to the variance present within the classes.

Quadratic Discriminant Analysis (QDA): QDA expands on LDA by further relaxing the assumptions of homoscedasticity and models features belonging to each class with its own covariance matrix (Σ). Unlike LDA which produces a linear decision boundary, QDA outputs a quadratic decision boundary [2]. Due to the bias-variance tradeoff, fitting complex models with more parameters require more data to estimate the parameters effectively. Therefore, the performance of methods such as GNB, LDA, QDA depends on the assumptions of normality and the ability to estimate the true mean and covariance of the feature distributions from the training data. When the assumptions of the model hold, Naïve Bayes, and LDA give good performances,

but when they are do not hold, a complex model with more parameters such as QDA would give better performance given enough data. In LDA and QDA, the number of parameters for the covariance matrix to be estimated is quadratic to the number of features. So in cases with large number of features and fewer observations, an invertible and stable estimate of the empirical covariance matrix is obtained by regularization. The empirical covariance matrix is obtained by the sum of the actual covariance matrix and the product of the scalar matrix with a shrinkage parameter denoted by γ. A case where $\gamma = 0$ corresponds to unregularized LDA and $\gamma = 1$ corresponds to a spherical covariance matrix, in which the off-diagonal values are zero [3].

Sparse Logistic Regression (SLR): Unlike the generative models discussed previously, Logistic Regression models the class distribution $P(Y|X)$ directly rather than using $P(Y)$ and $P(X|Y)$ [4]. By relaxing the assumptions of conditional independence, $P(Y = 1|X)$ is directly modeled using the logistic curve as

$$P(Y = 1|X) = \frac{1}{1 + exp(-\sum_{i=1}^{d} w_i x_i + w_0)} \quad (3)$$

where $W = [w_0, w_1, w_2, \ldots, w_d]$ indicate the parameters/weights of the model. The parameters/weights of the model W are estimated by maximizing the log-likelihood function of the observed data. Logistic Regression can be extended to multiclass classification by using the softmax function and modeling the class probability distributions as

$$P(Y = y_k|X) = \frac{\exp(-(\sum_{i=1}^{d} w_i^k x_i + w_i^k))}{\sum_{k=1}^{K} \exp(-(\sum_{i=1}^{d} w_i^k x_i + w_i^k))} \quad (4)$$

where W denotes the parameters/weights matrix of the model. We used the one-of-K target encoding [5, 6]; therefore the target vector Y for every instance becomes $Y = [y_1, y_2, \ldots y_K]$ where K is the number of classes and $y_c = 1$ if X belongs to class c and $y_c = 0$ otherwise, with each class having its own weights. Logistic Regression cannot be used when the number of

features is less than the number of data samples because it results in inversion of an ill conditioned matrix [6].

However, weights for the features are controlled by placing a prior such as a Gaussian or a Laplacian distribution over the weights to avoid overfitting as it restricts the values the weights can take. The choice of the weight priors determines the nature of the classifier model. A Gaussian prior on the weights, equivalent to an *L2* norm, leads to smoother models by controlling the weights of the model. This procedure is called regularization, and it helps in generalization. However, the Gaussian priors on the model make the weights smaller but do not drive them to zero. Driving the weights of either the features (in SLR) or kernels (in RVM) to zero helps in feature selection and better optimization as it promotes sparsity. Sparsity is promoted by either a Laplacian prior on the model weights which results in *L1* penalty, like in LASSO or by incorporating a Gamma hyperprior on the variance of the existing Gaussian weight prior (called an Automatic Relevance Determination prior) with a zero mean and a diagonal covariance matrix. Combining, the Logistic Regression with the Automatic Relevance Determination (ARD) prior on the weights, resultis in an Sparse Logistic Regression (SLR) model. The priors on the weights are written as

$$P(w|\alpha) = N(0, \alpha_d^{-1}) \text{ for all features } d \quad (5)$$

$$P(\alpha_d) = \alpha_d^{-1} \text{ for all features } d \quad (6)$$

Here α_d, called a relevance parameter, determines the range of weight parameter. As training progresses, α_d is driven to infinity for many weight vectors, thereby forcing the weight vectors to follow a Gaussian distribution with a zero mean and zero standard deviation, essentially driving their value to zeros. By using the nested priors, this model promotes sparsity as the true probability of the weight distribution is a Student t-distribution with a sharper peak at zero and

flatter tails [5]. These hyperparameters (α_d), regulate the weights with a single parameter controlling the variance of each of the weight distributions [7, 8]. This hierarchical prior, called an Automatic Relevance Determination (ARD) prior automatically drives the weights for many features to zero resulting in a compact classifier with inbuilt feature selection. For binary classification, we assume a Bernoulli distribution of the target class. For multiple classes, the likelihood can be written as

$$P(t|w) = \prod_{n=1}^{N}\prod_{k=1}^{C} \sigma\{y_k(x_n; w_k)\}^{t_{nk}} \tag{7}$$

Where *C* denotes the number of classes and N denotes the number of observations. w_k denotes the weights for the k^{th} class and t_{nk} denotes the target for the k^{th} class for the n^{th} observation. Training the SLR model requires us to maximize the log likelihood of the observed data. However, a true Bayesian model can be intractable. Using a variational Bayesian approximation to the true posterior over the weights and hyperparameters, the optimum relevance parameters and their corresponding weight vectors can be obtained by iteratively optimizing the marginal likelihood of the parameters and hyperparameters [6]. This is done by solving the weight parameters while the hyperparameters are fixed and updating the hyper-parameters while the weight parameters are fixed. This procedure is repeated until convergence, to get the final model [6].

Regularized Logistic Regression (RLR): Regularized Logistic Regression (RLR) model uses the *L2* norm to control model complexity. Instead of several hyperparameters to control the variance of the Gaussian, in RLR, there is just a single hyperparameter (α) for the Gaussian spread that is shared across the features. So the weight priors can be written as

$$P(w|\alpha) = N(0, \alpha^{-1} I_d) \tag{8}$$

where I_d denotes an identity matrix. A optimization procedure similar to that employed with SLR is followed for obtaining the optimum values of weights and the hyperparameter α. Due to the large computation time and resources associated with Sparse Multinomial Logistic Regression/Regularized Multinomial Logistic Regression, we implemented one-vs-all (ova) Sparse Logistic Regression/Regularized Logistic Regression as it gave similar results in much quicker time. We used the code provided by Yamashita et al. [6] for implementing both SLR and RLR in MATLAB.

A.2 Kernel Methods

Kernel methods use a similarity function called a kernel function, which allows them to operate in a high dimensional space without explicit computational costs of operating in the higher dimensional space. Support Vector Machine (SVM) is the most commonly used kernel method.

Support Vector Machine (SVM): SVM is considered to be one of the best out-of-box classifiers and is widely utilized by the neuroimaging community. SVM constructs a separating hyperplane between two classes that attempts to balance the dual objectives of minimizing the training error while maximizing the margin between the classes. The margin, in this case, can be defined as the smallest distance from the observations to the separating hyperplane. This ensures that the hyperplane is robust to individual observations and gives a better generalization. Only the observations that lie on the margin or the wrong side of the decision boundary (classified incorrectly) contribute to the construction of the hyperplane and are called support vectors. Support Vector Machines have built-in regularization as they tend to choose decision boundaries which have the largest margin. A margin can be formulated as an inverse of the weight norm $\|w\|$. However, when the classes are not linearly separable, we can modify the optimization problem to include a slack variable to deal with incorrectly classified instances,

which can be realized using a hinge loss function [2]. A tuning parameter *C* controls the contribution of the margin and the loss function. When the tuning parameter *C* is large, the margin is wide but a lot of training observations are misclassified, whereas with smaller *C*, the opposite is true. Since SVM requires just a product of data vectors, we can map the input vectors in a higher dimensional space by a suitable transformation when the classes are not linearly separable. Consequently, instead of the transforming the feature vectors in the higher dimensional space and calculating a dot product of the transformed data vectors, we can use a kernel to replace the product of transformed data vectors without explicitly calculating the feature mapping. This is known as the "kernel trick" [2]. A kernel can be written as

$$K(u, v) = \varphi(u).\varphi(v) \tag{9}$$

where K is the kernel, u and v are the input vectors and φ denotes the feature transformation. Besides the linear kernel, quadratic, and RBF kernels are extremely popular. The RBF kernel is denoted by

$$K(u, v) = exp(-\gamma\|u - v\|^2)) \tag{10}$$

where u and v are the input vectors and γ is the width of the radial kernel which determines the extent of the influence of the training observations on the test observations. Since SVM is an inherently binary classifier, it can be extended to multiclass classification by building several binary classifiers with either a one-vs-one or a one-vs-all strategy. In the one vs one strategy an SVM classifier is built for every pair of classes, i.e. a total of $\binom{K}{2}$ classifiers for *K* classes. The test observation is then assigned to the class which was most frequently assigned in the pair-wise classifications. In one-vs all strategy, *K* classifiers are built by comparing observations from one class with those of the rest *K-1* classes. A test observation is then assigned to a class which is farthest to the decision surface in the *K* classifiers. The difference between SVM, LDA and

logistic regression is that while SVM completely ignores the observations which are not close to the decision surface, LDA considers all observations, whereas logistic regression has lower sensitivity to observations farther from the decision surface [2]. In fact, SVM can be formulated as a regularized logistic regression with a hinge loss function. In this study, we used a linear and a radial kernel support vector machine with the hyperparameter C (in both cases) and the parameter γ (in the RBF kernel SVM) determined by cross-validation. We used LIBSVM software package [9] for the implementation of support vector machines in MATLAB.

Relevance Vector Machine (RVM): Relevance vector machine is similar in its formulation to the support vector machine in its use of the kernel functions centered on the training samples [10, 5]. However, the weights associated with the kernels are solved in a procedure similar to SLR, with weights associated with several kernel functions being driven to zero.

RVMs have a few distinct advantages over SVMs. The primary advantage is that RVMs, being probabilistic Bayesian models, give a probability distribution of the weights of the models as well as target variables rather than point estimates. Secondly, RVM provides us with a much sparser solution with fewer non-zero input kernels than SVM with similar or better accuracy [5]. Finally, the kernels in RVM do not have to satisfy the mercer conditions like in SVM, and it does not have tunable parameters such as C [5]. However, the implementation of RVM is extremely slow compared to SVM. The speed of the RVM algorithm is improved significantly with a bottom-up approach, by adding kernel functions one by one, updating their weights or removing them from the model [11]. This binary classification can easily be extended to multi-class classification by utilizing the multinomial function and one-of-K target encoding, unlike SVM. However RVM, unlike SLR, does not have automatic feature selection due to its use of the weighted kernels in the model formulation rather than weighted combination of features [12]. A

major benefit of full Bayesian formulations such as RVM is that it also takes into consideration the uncertainty in the weight estimations in the final model and the subsequent predictions. Finally, the most significant advantage of probabilistic classifiers in general is that they give posterior class probability distributions as Gaussians with means and variances estimated from the data, which is beneficial in that we can combine the posterior class probabilities with prior class probabilities and also combine multiple predictions from different sources in a probabilistic framework to make a prediction on a test observation [10].

In RVM, the most probable values of hyperparameters α (variance associated with the weights of the model) and σ^2 (variance associated with the target distribution t_n) are found by maximizing the marginal likelihood (also called Type -II likelihood) of the target distribution t, denoted by $P(t|\alpha, \sigma^2)$ using suitable approximations. Using the most probable values of hyperparameters, the probability distribution of the weight parameters is estimated. In a true Bayesian model, nuisance variables are marginalized. But given the intractability of the above computations, a variational approximation or Type -II maximum likelihood is used to approximate the marginal likelihood [13]. As mentioned previously, the computation time for RVM is significantly reduced by adopting a bottom-up greedy optimization approach. In this procedure, we started with an initial basis function and performed incremental optimization of the marginal likelihood where only one basis function is added, deleted or updated [11].Therefore, at every step, a basis function which gives the maximum reduction in the negative log marginal likelihood is added to the model, and its hyperparameters as well as the weights of the model are updated. This procedure speeds up the computations as it overcomes the difficulty in inverting a matrix for the calculation of the covariance matrix of the weights (Σ), but at the cost of slower convergence

and with a near optimal set of hyperparameters. For MATLAB implementation, we used the code provided with Thayanathan et al. [14].

A.3 Artificial Neural Networks

Artificial neural networks are computational approaches inspired from functioning of neurons in the human brain. The neural networks consist of computational units called neurons organized in multiple layers. A perceptron is a single artificial neuron which forms the building block for artificial neural networks.

Fully Connected Neural Net (FC-NN) and Multilayer Perceptron Neural Net (MLP-NN): Since a perceptron can classify only linearly separable classes, we need to use multiple layers of perceptrons to model complex relationships between the features and the target classes, so that each layer can progressively learn complex representations of the input data. Learning in neural networks is accomplished by changing the weights from the inputs to the hidden layer and from the hidden layers to the outputs, so that the mapping from the inputs to the outputs is learned. Conventionally, neural networks are trained by gradient descent using backpropagation. The training procedure by backpropagation consists of a forward pass in which errors are calculated and a backward pass in which the gradients required to change the weights are calculated. But backpropagation is extremely slow as it is a first order method and could take many iterations/epochs to reach the minimum in the error surface. For training the neural networks, we used Bayesian regularization backpropagation, which uses Levenberg-Marquard (LM) optimization to update the weights. The advantage of LM is that it is comparably fast as it is a second-order method which uses a Jacobian matrix to calculate an approximation of the Hessian, and hence converges pretty quickly to a local minimum. Since it uses Bayesian regularization, it controls the model complexity and avoids overfitting [15, 16]. With enough neurons, neural

networks can act as universal function approximators [17, 18]. However, we limit the number of hidden layer neurons and use priors on the weights between the connections to avoid overfitting as our data has far more features than the number of observations, which is ill-suited for training neural networks. Deep networks with more hidden layers can learn more complex functions. Therefore, we implemented a Fully Connected Neural Network (FC-NN) with Bayesian regularization as well. In this architecture, each layer has a single neuron and is connected to the inputs and every previous layer [19]. This architecture has more expressive power and can learn any function that could be learned by the same number of neurons in a single hidden layer, but the reverse is not true [20]. We limited the complexity of our model to just three hidden layers in FC-NN and three neurons in the hidden layer in the MLP-NN. The MLP-NN we used is essentially a Single Hidden layer Feedforward Neural Nets (SLNF). Finally, two significant issues with neural networks is that: (i) optimizing error function can get struck in local minima and the (ii) vanishing gradient problem, when the error gradient required to train the weights decreases exponentially, slowing the training process. More information on neural network architectures can be found in [21, 22].

Extreme learning machine (ELM): ELM is a class of learning algorithms for training single hidden layer feedforward neural networks (SLNF), where the hidden nodes are randomly initiated, and only the connections between the hidden layer to the output layer are tuned, resulting in extremely fast learning times and good generalization performance when the output weights are regularized [23, 24]. Unlike traditional feed forward learning procedures, in ELM, hidden nodes need not be tuned and the parameters are randomly generated, independent of the training data and each other. This unconventional approach gives extreme learning machines their speed. The hidden units may be sigmoid, like in traditional neurons or can be linear,

polynomial or RBF kernels. We used RBF/Gaussian kernels of the form $G(a,b,x) = exp(-b\|x-a\|^2)$ for generating the random feature mappings from the input data into an ELM feature space by randomly initializing the parameters for all the hidden layer neurons and optimizing the width of the RBF kernel functions by cross-validation. We then calculated the weights of the output mappings from the ELM feature space to the output feature space by minimizing a cost function. This cost function simultaneously minimizes the sum of the *L2* norm of the weights and the squared training error, using a hyper parameter *C*, which regulates the training error and the complexity of the weights.

Learning Vector Quantization Neural Network (LVQNET): LVQNET is a supervised classification algorithm introduced by Kohonen [25, 26]. It is a two-layer neural network with a competitive first layer and a linear second layer. The competitive layer learns a codebook of vectors in the input space that divides the input space into subclasses. The linear layer then combines the subclasses to the target classes. The number of neurons in the competitive layer is equal to the number of prototype vectors in the input space and is determined by cross-validation. The number of neurons in the output layer equals the number target classes. Training the neural network is accomplished by calculating the Euclidean distance between a training observation and input weight vectors. The winning neuron/prototype vector in the hidden layer which has the smallest distance as determined by its weight vector, will output a one and all other neurons output a zero in a winner takes all decision. Using the weights between the hidden and the output layer, the class of the training observations is computed. If it matches with the actual class, then the weights of the winning neuron/prototype are adjusted to move closer to the input vector, if not, the weights of the winning neuron are moved away from the input vector. By repeating this process for several epochs with all the training observations, the feature space is

divided into subclasses based on the target class. The linear layer then assigns the target class to the subclasses. This learning rule is improved by updating the weights of two closest weight vectors between the input and the hidden layers if one belongs to the correct class and the other belongs to the wrong class [25, 26].

A.4 Instance based learning

In instance based learning, no explicit model is learned during training. Rather, predictions on the test observation are made based on its similarity to the training points. The simplest and the most popular model in this category is K- Nearest Neighbors classifier.

K-Nearest Neighbors (KNN): KNN is a non-parametric lazy learning algorithm which gives excellent performance for irregular decision boundaries. To classify an unlabeled test data point, the distance of the test data point with instances in the training data are compared. The *K* closest points are determined, and the class of a test observation is assigned to the majority class of the neighbors. The hyperparameter *K* which determines the number of neighbors is determined by using cross-validation. We used the Euclidean distance measure, to measure the similarity between two data points, though it may become less discriminative if the number of features increases and only a few of them are informative [27].

A.5 Decision Tree-based ensemble methods

Multiple classifiers can give much better results than a single classifier. However, the improved performance comes at a significant cost of training multiple classifiers. Use of methods which employ multiple classifiers to solve a problem is called ensemble learning. Decision Tree is a popular choice of the base classifier in ensemble methods [28]. A decision tree tries to divide the observations space into rectangular homogeneous regions. Using a measure of purity as the splitting criteria, a greedy top-down splitting is performed to build the tree. However, individual

trees suffer from high variance and give subpar performances on many classification and regression problems [2]. So by combing multiple classifiers in an ensemble, the performance of the decision trees is improved.

Bagged Tree & Random Forest: Bagging is one of the two most common ways of training ensemble classifiers using a decision tree as the base classifier. In tree bagging, multiple classifiers are trained with randomly selected bootstrapped samples of data from the training set and the test observation is assigned to the class obtained by majority voting of all the individual classifiers, resulting in reduced variance. Bagging constructs trees that are pretty similar and hence produce correlated outputs. Further reduction in variance is possible by promoting diversity in the ensemble by de-correlating the trees. Instead of considering all features/predictors for each split, we used a subset of features (usually square root of the features). This ensemble method is called Random Forest [29]. One advantage with random forests is that it has very few user-dependent parameters and the algorithm is insensitive to even the few parameters that the users need to choose [29].

Boosted Stumps & Trees: Another class of learning algorithms which can be used to improve the performance of weak classifiers is Boosting. Boosting is a process for converting weak learners which perform slightly better than chance into strong learners. AdaBoost (Adaptive Boosting) is a boosting algorithm proposed by Yoav Freund and Robert Schapire [30, 31]. It is adaptive since it automatically adapts to the training data and constructs a single composite classifier from individual weak classifiers. AdaBoost works by sequentially building an ensemble of classifiers that iteratively down-weigh observations that are correctly classified and up-weigh observations that are incorrectly classified. This ensures that subsequent classifiers focus more on the misclassified samples. An adaptive parameter alpha which is dependent on the

classification accuracy is used to reweight the samples and is used as the weight of the current classifier. The final model is then derived by the weighted sum of individual learners, whose weights are based on the learner's performance. Any classifier which can use weighted data points for calculating performance can be boosted by AdaBoost, though the use of decision trees as a base classifier is very popular. Because of boosting's tendency to overfit in a few cases when full grown trees are used, decision stumps are sometimes preferred as a base classifier. A decision stump is a very shallow decision tree with just one split. We used AdaBoost.M1 for binary classification and AdaBoost.M2 for multiclass classification using decision stumps [32]. Along with AdaBoost, we also used Linear Programming Boosting (LPBoost) [33] for learning full grown decision trees as it gave us much better performance than AdaBoost on imbalanced datasets.

Rotation Forests: Rotation forests get their name because they rotate the original coordinate space by Principal Component Analysis (PCA) and use decision trees as the base classifier [34]. Promoting diversity without compromising the accuracy of the base classifiers is the goal of ensemble learning. In random forests, diversity is introduced through bootstrapping and splitting on a subset of features. In contrast, in rotation forests, diversity in the base classifiers is introduced by transforming subsets of features by PCA and using the transformed features to build trees. Rotation forests construct classifiers by splitting the features into K subsets, and PCA is applied for every feature subset with a subset of classes removed from the data to promote diversity. The coefficients obtained are then arranged appropriately in a transformation matrix called a rotation matrix, which is then used to transform the original features. A decision tree is constructed by using the transformed features. When classifying a test observation, features are multiplied with the rotation matrix and the mean confidence of the observation belonging to each

class is estimated. The class with the largest confidence is assigned to this test observation. Diversity in rotation forests comes from random differences in possible feature subsets and the random subset of training data and classes selected for learning each classifier. Rotation forests create classifiers that are less diverse and more accurate than random forests and boosted trees [35]. Though rotation forests give comparable or better accuracies compared to random forests, it suffers from a few limitations. It has an extra parameter to control the size of the feature subsets, and unlike random forest, rotation forests do not provide us with information about feature importance.

A.6 Bibliography

[1] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker and Z. Ghahramani, Eds., MIT Press, 2002, p. 841–848.

[2] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, New York: Springer Science+Business Media, 2013.

[3] S. Lemm, B. Blankertz, T. Dickhaus and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage,* vol. 56, no. 2, pp. 387-399, 2011.

[4] S. Ryali, K. Supekar, D. A. Abrams and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage,* vol. 51, no. 2, pp. 752-764, 2010.

[5] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research,* vol. 1, p. 211–244, 2001.

[6] O. Yamashita, M.-a. Sato, T. Yoshioka, F. Tong and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *NeuroImage,* vol. 42, no. 4, pp. 1414-1429, 2008.

[7] R. M. Neal, Bayesian Learning for Neural Networks, New York: Springer, 1996.

[8] D. J. C. MacKay, "Bayesian Methods for Backpropagation Networks," in *Models of Neural Networks III*, New York, Springer, 1996, pp. 211-254.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM*

Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 27:1-27:27, 2011.

[10] M. E. Tipping, "Bayesian Inference: An Introduction to Principles and Practice in Machine Learning," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg and G. Rätsch, Eds., Springer, 2004, pp. 41-62.

[11] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 2003.

[12] B. Krishnapuram, L. Carin, M. A. Figueiredo and A. J. Hartemink, "Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 6, pp. 957-968, 2005.

[13] C. M. Bishop and M. E. Tipping, "Variational Relevance Vector Machines," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 2000.

[14] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr and R. Cipolla, "Multivariate Relevance Vector Machines for Tracking," in *Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III*, vol. 3953, A. Leonardis, H. Bischof and A. Pinz, Eds., Berlin, Heidelberg, Springer, 2006, pp. 124-138.

[15] D. J. C. MacKay, "Bayesian Interpolation," *Neural Computation,* vol. 4, no. 3, pp. 415-447, 1992.

[16] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian learning," in *Neural Networks,1997., International Conference on*, Houston, TX, 1997.

[17] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks,* vol. 4, no. 2, pp. 251-257, 1991.

[18] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems,* vol. 2, no. 4, p. 303–314, 1989.

[19] G. Deshpande, P. Wang, D. Rangaprakash and B. Wilamowski, "Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data," *IEEE Transactions on Cybernetics,* vol. 45, no. 12, pp. 2668 - 2679, 2015.

[20] N. Kriegeskorte, "Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing," *Annual Review of Vision Science,* vol. 1, no. 1, pp. 417-446, 2015.

[21] C. M. Bishop, Neural Networks for Pattern Recognition, New York: Oxford University

Press, Inc., 1996.

[22] S. O. Haykin, Neural Networks and Learning Machines, Upper Saddle River, NJ: Prentice-Hall, Inc., 2009.

[23] G. Huang, G.-B. Huang, S. Song and K. You, "Trends in extreme learning machines: A review," *Neural Networks,* vol. 61, pp. 32-48, 2015.

[24] G.-B. Huang, D. H. Wang and Y. Lan, "Extreme learning machines: a survey," *International Journal of Machine Learning and Cybernetics,* vol. 2, no. 2, p. 107–122, 2011.

[25] T. Kohonen, "Improved versions of learning vector quantization," in *1990 IJCNN International Joint Conference on Neural Networks*, San Diego, CA, USA, 1990.

[26] T. Kohonen, "Learning Vector Quantization," in *Self-Organizing Maps*, T. Kohonen, Ed., Berlin, Heidelberg, Springer Berlin Heidelberg, 1995, pp. 175-189.

[27] F. Pereira, T. Mitchell and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage,* vol. 45, no. 1, p. S199–S209, 2009.

[28] M. Gashler, C. Giraud-Carrie and T. Martinez, "Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous," in *Seventh International Conference on Machine Learning and Applications*, San Diego, CA, 2008.

[29] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[30] Y. Freund and R. E. Schapire, "A Short Introduction to Boosting," *Journal of Japanese Society for Artificial Intelligence,* vol. 14, no. 5, pp. 771-780, 1999.

[31] R. E. Schapire, "A Brief Introduction to Boosting," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

[32] Y. Freund and R. E. Schapire, "Experiments with a new Boosting algorithm," in *International Conference on Machine Learning*, 1996.

[33] A. Demiriz, K. P. Bennett and J. Shawe-Taylor, "Linear Programming Boosting via Column Generation," *Machine Learning,* vol. 46, no. 1, p. 225–254, 2002.

[34] J. J. Rodriguez, L. I. Kuncheva and C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 10, pp. 1619-1630, 2006.

[35] L. I. Kuncheva and J. J. Rodríguez, "An Experimental Study on Rotation Forest Ensembles," in *Multiple Classifier Systems: 7th International Workshop, MCS 2007, Prague, Czech Republic, May 23-25, 2007. Proceedings*, vol. 4472, M. Haindl, J. Kittler

and F. Roli, Eds., Prague, Springer Berlin Heidelberg, 2007, pp. 459-468.