Machine Learning Approaches for Disease State Classification from Neuroimaging Data

by

Peng Wang

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
August 3, 2013

Keywords: *Neural networks, fMRI, Classification, Generalization, ADHD*

Approved by

Gopi Deshpande, Chair, Assistant Professor, Department of Electrical & Computer Engineering
Tom Denny, Director, Auburn University MRI Research Center
Jeff Katz, Alumni Professor, Department of Psychology Auburn University
Nedret Billor, Associate Professor, Department of Mathematics and Statistics

Abstract

Automated recognition and classification of brain diseases are of tremendous value to society. Attention deficit hyperactivity disorder (ADHD) is a diverse spectrum disorder whose clinical diagnosis is based on behavior. In this study, I proposed a two-step cross-validation procedure to illustrate the utility of fully connected cascade (FCC) artificial neural network (ANN) architecture, which provided excellent capability of generalization and outperformed support vector machines in terms of accuracy for both balanced and unbalanced sample sizes, irrespective of the features used. Additionally, I employed various directional and non-directional connectivity based methods to extract discriminative features. I obtained close to 90% accuracy for distinguishing ADHD from healthy subjects and 95% between the ADHD subtypes, which are better than the winning accuracy of the ADHD-200 Global Competition and those reported subsequently. Finally, the most discriminative connectivity features showed reduced and altered connectivity involving the left orbitofrontal cortex and various cerebellar regions in ADHD.

Acknowledgments

I thank the International Neuroimaging Data-sharing Initiative (INDI) for providing access to the data. I thank Dr. Gopi Deshpande for his great patience and wonderful instructiveness. I thank my friends at the MRI center of Auburn University. They have offered me numerous help completing my study. I specially thank my parents for their constant support and encouragement, without which I would never have gone this far.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

I. MRI

Magnetic resonance imaging (MRI) is a noninvasive medical imaging technique that uses magnetic fields and radio wave pulses to make pictures of organs and structures inside the body. Magnetic field gradients cause nuclei at different locations to rotate at different speeds. Magnetized nuclei generate their own magnetic fields and create signals with different frequencies and amplitudes at different locations. Therefore, Fourier analysis can be applied to recover spatial information of the measured signal and an image of the scanned area of the body can be constructed. By using gradients in different directions, 2D images and 3D volumes can be obtained in any arbitrary orientation. For many organs and structures including brain, heart, eyes, soft tissues, connective tissues, muscles, most tumors, and cartilage, MRI gives different information from what can be seen with other medical imaging techniques such as computed tomography (CT), X-rays, or ultrasound.

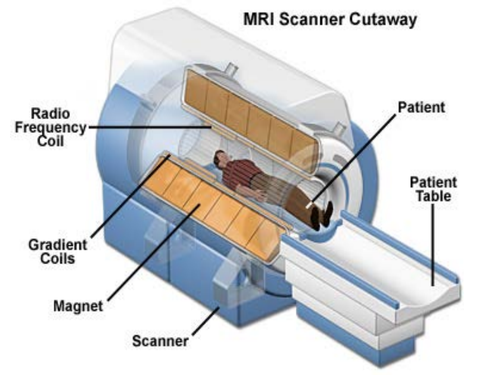

Fig.1.1 MRI scanner

By changing the settings on the scanners (a typical sectional diagram of MRI scanner is shown in Fig.1.1), contrast can be created between different types of body tissues or between other structures, as in fMRI and diffusion MRI. In some cases, contrast material and higher magnetic fields may be used during the MRI scan to show certain structures more clearly. Researchers have contributed great effort utilizing MRI technology on brain study. The development of various methods for noninvasive brain function mapping without using an exogenous contrast agent has brought revolutionary advancement in MRI [1].

II. Functional MRI

Functional magnetic resonance imaging (fMRI) is an MRI procedure that measures brain activity mainly based on the blood oxygenation level dependent contrast [2] and has developed into an essential tool for studying brain functionality in both healthy and unhealthy states [3]. fMRI relies on the fact that cerebral blood flow and neuronal activation are coupled. When an area of the brain is in use, blood flow to that region increases. The change in the magnetic resonance (MR) signal from neuronal activity is called the hemodynamic response (HDR), which lags the neuronal events triggering it by 1 to 2 seconds. From this point it typically rises to a peak at about 5 seconds after the stimulus. If the neurons keep firing, the peak spreads to a flat plateau while the neurons stay active. After activity stops, the blood oxygen level dependent (BOLD) signal falls below the original level, the baseline. Over time the signal recovers to the baseline. In short words, within the brain, changes in the local concentration of paramagnetic deoxyhemoglobin lead to alterations in the MR signal. Neuronal activation is generally believed to cause an increase in regional blood flow without a corresponding increase in the regional oxygen consumption rate [4], which should cause a decrease in the capillary and venous deoxyhemoglobin concentrations. Consequently, an increase in magnetic spin-spin relaxation times T_2* and T_2 should occur [1], thus leads to an increase of intensity in T_2*- and T_2-weighted MR images.

Spatial resolution of an fMRI image is measured by the size of voxels, as in MRI. A voxel is a three-dimensional rectangular cuboid, whose dimensions are set by the slice thickness, the area of a slice, and the grid imposed on the slice by the scanning process. Full-brain studies use larger voxels, while those focusing on specific regions of interest (RIO) typically use smaller sizes. Temporal resolution for fMRI scan is usually between 1 and 2 seconds. The scanner platform generates a 3D volume of the subject's brain every time of repetition (TR). This consists of an array of voxel intensity values, one value per voxel in the scan. The voxels are arranged one after the other, unfolding the three-dimensional structure into a single line. Several such volumes from a session are joined together to form a 4D volume corresponding to the time period when the subject stayed in the scanner without adjusting head position.

III. FMRI data preprocessing

In order to fully utilize fMRI data, preprocessing procedure has to be taken prior to brain functionality study. The first conventional step in fMRI data preprocessing is slice timing correction. The MRI scanner acquires different slices within a single brain volume at different times; therefore the slices represent brain activity at different time points. Since this complicates later analysis, a timing correction is applied to bring all slices to the same time point reference. This is done by assuming the time course of a voxel is smooth when plotted as a dotted line. Hence the voxel's intensity value at other times which is not in the sampled frames can be calculated by filling in the dots to create a continuous curve.

Head motion correction is another common preprocessing step. When the head moves, the neurons under a voxel move; therefore its current time course represents largely that of some other voxel in the past. The time course curve is effectively cut and pasted from one voxel to another. Motion correction applies a rigid-body transform to the volume, by shifting and rotating the whole volume data to account for motion. The transformed volume is compared statistically to the volume at the first time point to see how well they match, using a cost function such as correlation or mutual information. The transformation that gives the minimal cost function is chosen as the model for head motion. Since the head can move in a vastly varied number of ways, neither is it practical to search for all possible candidates; nor is there an algorithm that provides a globally optimal solution independent of the first transformations in a chain.

Distortion corrections account for field nonuniformities of the scanner. One method is to use shimming coils. Another method is to recreate a field map of the main field by acquiring two images with differing echo times. If the field were uniform, the differences between the two images also would be uniform. Bias field estimation is a preprocessing technique using mathematical models of the noise from distortion, such as Markov random fields and expectation maximization algorithms, to correct for distortion. In general, fMRI studies acquire both many functional images with fMRI and a structural image with MRI. The structural image is usually of a higher resolution and depends on the T_1 magnetic field decay after excitation. To mark regions of interest in the functional image, one needs to align it with the structural one.

Temporal filtering is the removal of frequencies of no interest from the signal. A voxel's intensity change over time can be represented as the sum of a number of different repetitive waves with differing periods and amplitudes. A plot with these periods on the x-axis and the amplitudes on the y-axis is called a power spectrum, and this plot is created with the Fourier transform technique. Temporal filtering amounts for removal of the periodic waves of least interest to us from the power spectrum, and then summing the waves back again, using the inverse Fourier transform to create a new time course for the voxel. A high-pass filter removes the lower frequencies, and the lowest frequency that can be identified with this technique is the reciprocal of twice the TR. A low-pass filter removes the higher frequencies, while a band-pass filter removes all frequencies except the particular range of interest.

Smoothing, or spatial filtering, is the idea of averaging the intensities of nearby voxels to produce a smooth spatial map of intensity change across the RIO. The averaging is often done by convolution with a Gaussian filter, which, at every spatial point, weights neighboring voxels by their distance. If the true spatial extent of activation matches the width of the filter used, this process improves the signal-to-noise ratio. It also makes the total noise for each voxel follow a bell-curve distribution. But if the presumed spatial extent of activation does not match the filter, signal is reduced.

IV. ADHD and ADHD-200 Global Competition

Attention deficit-hyperactivity disorder (ADHD) is a mental and neurobehavioral disorder characterized by either significant difficulties of inattention or hyperactivity and impulsiveness or a combination of the two. ADHD affects at least 5% of school-age children and is associated with substantial lifelong impairment, with annual direct costs exceeding $36 billion in the US. ADHD consists of three subtypes: (1) predominantly inattentive (ADHD-PI or ADHD-I), (2) predominately hyperactive-impulsive (ADHD-HI or ADHD-H), (3) combination of subtype 1 and subtype 2 (ADHD-C). The symptoms of ADHD usually emerge before age seven. Inattention, hyperactivity, disruptive behavior and impulsivity are common in ADHD. Academic difficulties are also frequently shown for ADHD patients. The symptoms are especially difficult to define because it is hard to draw a line which clearly separates normal levels from abnormal levels of inattention, hyperactivity and impulsivity. The specific causes of ADHD are not known to this day. However there are a number of factors, including genetics, diet and the social and physical environments, which may contribute to, or exacerbate ADHD.

Despite voluminous empirical literature, the scientific community is still handicapped on modeling the pathophysiology of ADHD. Further, the clinical community remains without objective biological tools capable of informing the diagnosis of ADHD for individual or guiding clinicians in their decision-making regarding treatment. The ADHD-200 Sample [5] is dedicated to accelerating the scientific community's understanding of the neural basis of ADHD through the implementation of open data-sharing and discovery-based science. The data consist of 776 resting-state fMRI and anatomical images aggregated across 8 independent imaging sites, 491 of which were obtained from typically developing individuals and 285 in children and adolescents with ADHD (ages: 7-21 years). Accompanying phenotypic information includes diagnostic status, dimensional ADHD symptom measures, age, sex, intelligence quotient (IQ) and lifetime medication status. Preliminary quality control assessments (usable vs. questionable) based upon visual time series inspection are included for all resting state fMRI scans. Winning team from Johns Hopkins University scored 119 out of 195 points, with one point awarded per correct diagnosis (typically developing, ADHD primarily inattentive type, or ADHD combined type). A half point was awarded for a diagnosis of ADHD with incorrect subtype. The winning team correctly classified 94% of Typically Developing Children (TDC), excellent specificity. Their method was not as effective in terms of sensitivity. They only identified 21% of cases; however, among those cases, they discerned the subtypes of ADHD with 89.47% accuracy [6]. The methods developed by teams from the Chinese Academy of Sciences and the University of North Carolina at Chapel Hill both scored well on the J-statistic, a joint measure of specificity and sensitivity, suggesting that tests can be developed that satisfy needs in both these crucial diagnostic areas. Participants were able to develop predictive methods that performed significantly better than chance analyzing datasets that were gathered in an uncoordinated way by multiple centers. These results suggest that effective methods can be developed in less-than-ideal and poorly controlled environments. Importantly while the intent of the competition was imaging-based classification, the team of the University of Alberta scored 124 points using all available phenotypic data while excluding imaging data – 5 more points than the winning imaging-based classification approach. Their achievement highlights both the need for carefully coordinated imaging datasets for the development of analytic tools, and that diagnostic imaging tools have not yet reached full maturity.

V. Motivation and Proposal

Metrics derived from fMRI data have been widely used for disease classification [7, 8], providing potentially important applications in clinical diagnoses that benefit society. However, not all classification efforts have been successful. For example, spectrum disorders (i.e. the ones whose clinical diagnostic criteria can be very broad), including ADHD, have been particularly difficult to classify using imaging and behavioral metrics. A recent report showed that including both phenotypic and imaging information can boost performance [9]. Another study demonstrated that careful removal of motion confounds from fMRI data improves classification [10]. However, all these previous studies used support vector machines (SVMs) for classification. In this study, I investigated the utility of deep neural network architectures for classification purposes and demonstrated its superiority over support vector machines for classification of ADHD from healthy subjects using resting state fMRI data.

In classification literature, SVMs, K Nearest Neighbor (K-NN), Logistic Regression (LogReg), Radial Basis Network (RBF) and artificial Neural Networks (NN) are the most widely used classifiers. One of the most favored one is the SVM classifier, which has been applied in many application areas including text categorization [11], handwritten character recognition [12], image based gender identification [13], and bioinformatics [14]. SVM and NN with Multi-Layer Perceptron (MLP) architecture (Fig.1.3) are both well suited for classification and regression tasks. While NN with MLP possesses better generalization ability, SVM learning algorithm is based on support vector selection and higher order optimization, and therefore, has greatly reduced training time. Hence, it is better suited for large data sets [15]. Models based on SVM with Gaussian RBF kernel generally perform better than Error Back Propagation Neural Network with MLP architecture as demonstrated by applications using financial data, blood pressure data and facial expression images [16, 17]. SVM classifier shows a significant increase (over 15%) in classification accuracy when using RBF and polynomial kernels [18] and much shorter training time than NN with MLP when diagnosing breast tumor – 1 vs. 189 seconds [19]. The main advantage of using SVM over NN is that SVM always finds a global minimum, while feed forward neural networks can be stuck in a sub-optimal solution [20]. However SVM is not always superior to NN. With a better training algorithm and more efficient architecture, NN's performance can be significantly improved. While Levenberg-Marquardt (LM) training algorithm is specifically developed for MLP, it has been recently demonstrated that Neuron-by-Neuron (NBN) training algorithm is faster and more accurate than error back propagation and LM algorithms [21]. Fully Connected Cascade (FCC) architecture (Fig.1.2) is the most powerful architecture which can apply NBN training algorithm. It allows connections across layers and therefore possesses more computing power than MLP [22]. For example, with 10 neurons, FCC is possible to solve as large a problem as Parity-1023 (1023 inputs and 2^{1023} output patterns) using 1023=1=1=1=1=1=1=1=1=1=1 topology while MLP with one hidden layer can only solve Parity-9 with 10 neurons [21, 22, 23]. Therefore in our study, we tested the hypothesis that FCC with NBN training algorithm can outperform SVM on ADHD classification using fMRI data.

In addition to the efficacy of classifier design, the classification accuracy also depends on the ability of the input features to discriminate between the classes. Therefore, it is imperative to choose the features which are biologically informed and model the underlying neural process well, such that they are most likely to discriminate between healthy subjects and ADHD patients. In this regard, it is increasingly being recognized that the connectivity (synchronization of brain activity among activated areas) among the brain regions is an important marker of brain functionality [24], in addition to activity in individual areas. Many task-based connectivity studies have demonstrated alterations in clinical populations. For instance, lower functional connectivity was shown in the left hemisphere language network during irony processing for autism patients [25], greater functional connectivity was shown from the putamen to other front-striatal regions for Obsessive-Compulsive Disorder (OCD) participants [26], reduced frontal connectivity was shown in individuals with Major Depressive Disorder (MDD) [27], clear increase of connectivity was demonstrated between work memory regions and language regions as the processing load increases for syntactically complex sentences [28], etc. However, given the fact that clinical populations have difficulty performing tasks inside the MRI scanner, distributed connectivity in resting state brain networks, as opposed to various task states, have previously been shown to be very sensitive to baseline alterations in various disorders such as cocaine abuse [29], multiple sclerosis [30] and Alzheimer’s disease [31], depression [32], autism [33] and Parkinson’s disease [34]. Specifically with respect to ADHD, it has also been confirmed that (1) resting state brain

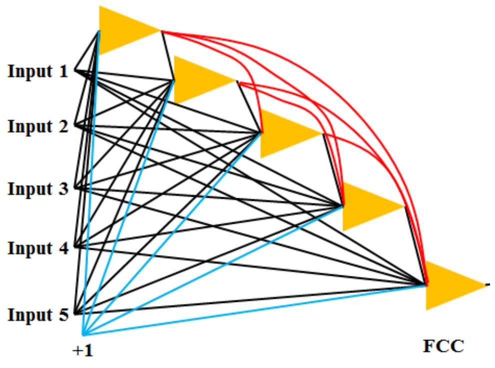

Fig.1.2 Fully Connected Cascade architecture (Denoted as: 5=1=1=1=1=1)

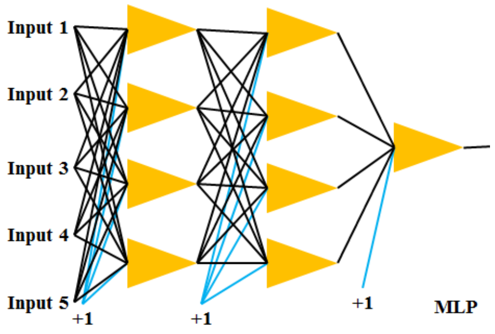

Fig.1.3 Multi-Layer Perceptron architecture (Denoted as: 5-4-4-1)

connectivity patterns in individuals are capable of differentiating the two most prominent ADHD subtypes using SVM-based multivariate pattern analysis (MVPA) [10], (2) compared to healthy subjects, patients with ADHD showed significantly reduced connectivity between bilateral pulvinar and right prefrontal regions, and significantly increased connectivity between the right pulvinar and bilateral occipital regions [35]. All of the studies cited above suffer from two major drawbacks. First, the connectivity metrics do not model directional interactions between regions. Rather, they calculated undirected association between signals from different brain regions, i.e. zero-lag synchronization, using statistical metrics such as Pearson's correlation coefficient. On the other hand, directional connectivity metrics such as Granger causality [36, 37, 38, 39, 40, 41, 42, 43, 44] model the causal interactions between brain regions and have been shown to be superior to correlation-based metrics for disease state classification [7]. Therefore, in this study, we derived directional connectivity metrics between brain regions from individual subjects and used them as features in our classification. Second, the connectivity metrics are based on linear models whereas biological processes are known to be non-linear. Therefore, in this study we estimate nonlinear directional connectivity using Kernel Granger causality [45, 46] and nonlinear undirected synchronous connectivity using correlation between probabilities of recurrences (CPR) [47]. Finally, we compared the performance of various connectivity-based metrics with that obtained by using raw fMRI data as features.

Even though the choice of brain connectivity-based metrics as features for disease state classification is biologically inspired, noise in the data (i.e. not all connectivities may be relevant for distinguishing the classes and thermal/physiological noise may impact those which do have the discriminatory power) and sample size may impact classifier performance. Therefore, feature selection plays a prominent role in determining the performance of a classifier. Therefore, we performed statistical tests on our features to pick the ones that are at least statistically different between the classes and adopted Principal Component Analysis (PCA) to reduce thermal/physiological noise and reduce the dimensionality of the features. PCA was invented in 1901 by Karl Pearson, is an eigenvector-based multivariate analyses that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. It reveals the internal structure of the data in a way that best explains the variance in the data. It shows the lower-dimension view of the high dimensional data with the most informative viewpoints. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to the preceding components. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute. Many previous studies have demonstrated the utility of PCA for feature selection and classification. For example, a classification accuracy of 90.6% was achieved using PCA of resting-state brain connectivity and SVM-based classification techniques for distinguishing healthy individuals from those with MDD [27]. PCA based metrics yielded a sensitivity of 82% and specificity of 86% for distinguishing between Multiple Sclerosis (MS) patients and healthy subjects [48]. In another study, multivariate pattern analysis using PCA was employed to classify depressed patients from healthy subjects with 100% specificity and 94.3% sensitivity [49]. PCA-based features also showed 80.4%, 77.6%, and 78.7% accuracies for classifying patients with schizophrenia & healthy controls, patients with schizophrenia & healthy siblings and healthy controls & healthy siblings [50], respectively. When multiple-kernel SVMs with PCA was applied, 96.3% classification accuracy was achieved for Mild Cognitive Impairment (MCI) diagnosis [51]. Therefore, we hypothesize that, when used in combination with a state-of-the-art deep neural network architecture and biological inspired features, the principal components and latent variables of those features are likely to boost the classification performance of even a difficult problem such as the one discriminating ADHD from healthy subjects. The second aspect of the effect of the sample size on classifier performance has been debated for a long time [52]. For example, the correspondence between the size of the sample and the number of support vectors in an SVM is a well-researched topic [53]. Some studies have shown that increasing training size will better performance [54]. Therefore in our study, we experimented with different sample sizes and tested whether classification results in our specific application was sensitive to it.

For my thesis, I propose a FCC deep NN architecture (Fig.1.2, compared with MLP in Fig.1.3) which overcomes MLP's limitations and possesses broad generalization ability. We demonstrate that the proposed NN architecture performs better than SVMs do, on classifying healthy subjects, i.e. TDC, from the ones with ADHD combined and ADHD inattentive for all combinations of metrics (i.e. raw data, linear/nonlinear directional/non-directional connectivity, principal components/latent variables of those features), and balanced/unbalanced samples of different sizes. We report that the overall performance of FCC deep NN architecture exceeds the best results obtained from the ADHD-200 Global Competition held in 2011 (best classification accuracy was 61%) [6], as well as results provided by Fair et al [10] in 2013 after their subsequent re-analysis of this dataset.

Chapter 2 Method

I. fMRI data

Pre-processed (head motion correction, spatial smoothing and normalization, frame-wise displacement adjustment, temporal band-pass filtering, etc.) fMRI time series data from 190 brain regions of 744 TDC subjects, 260 ADHD combined subjects, and 173 ADHD inattentive subjects, were obtained from the ADHD-200 Global Competition database [55]. The 190 brain regions were defined based on spectral clustering of resting state fMRI data. These regions contained voxels whose corresponding time series were most homogeneous [56]. All subjects were scanned on 3 Tesla scanners using standard resting T_2*-weighted echo-planar imaging, with sampling period (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle = 90 degree, and in-plane resolution = 64×64 mm^2 [10]. Table 2.1 shows the data acquisition sites which contributed to the composition of our sample.

	TDC	ADHD combined	ADHD inattentive
KENNDY KRIEGER INSTITUTE	62	17	6
NEW YORK UNIVERSITY CHILD STUDY CENTER	184	128	82
NEUROIMAGE SAMPLE	23	18	1
OREGON HEALTH & SCIENCE UNIVERISTY	124	69	38
PEKING UNIVERSITY_1	61	7	17
PEKING UNIVERSITY_2	32	15	20
PEKING UNIVERSITY_3	23	7	12
UNIVERSITY OF PITTSBURGH	89	0	0
WASHINGTON UNIVERSITY IN ST.LOUIS	151	0	0
Total	749	261	176

Table 2.1. Data sample composition

II. Feature Extraction and Selection

In addition to the raw image intensity values at each of the 190 brain regions, brain connectivity metrics between the 190 brain regions were extracted using three methods:

(1) Correlation between Probabilities of Recurrences (CPR): This is a non-parametric method for finding phase synchronization (PS) from two time series using temporal recurrence of patterns [47]. CPR is obtained from the phase space trajectory of the observed signal. CPR captures

higher order and potentially nonlinear synchronizations [47]. Complete PS occurs when the respective phases and frequencies of two signals are locked. CPR measures the degree of PS between two signals as ranging from 0 to 1, where 0 represents no PS and 1 represents complete PS.

Given a time series of length N as given below,

$$x_1, x_2, x_3, \cdots\cdots x_i, \cdots x_N \tag{1.1}$$

Vectors $\mathbf{y}_i$ of dimension D and lag (delay) d are defined as

$$\boldsymbol{y}_i = \begin{bmatrix} x_k \\ x_{k+d} \\ x_{k+2d} \\ \vdots \\ x_{k+(D-1)d} \end{bmatrix} \tag{1.2}$$

Where $D \geq 1$, $d \geq 1$, i=1 to N, k=i mod N-(D-1)d. The variable $\boldsymbol{y}$ gives the trajectory of x in phase space of dimension D and lag d. Based on this, we define the recurrence matrix as follows

$$R(i, j) = \Theta(r - \| \overrightarrow{y_i} - \overrightarrow{y_{i+\tau}} \|) \text{ ; i, j=1,2,3.......N} \tag{1.3}$$

Where N is the number of states considered, r is the threshold distance, Θ(.) is the Heaviside unit step function, and ||.|| is a norm. The trajectory returns to the neighborhood of i after a delay τ when j=i+τ, and $R(i,j)$=1. Considering the number of such recurrences for all *(i, i+τ),* relative to the total number N- τ, we get $P(\tau)$, which is an estimate of the probability that the system returns to a pre-defined state after a delay τ. The probability $P(\tau)$ that each of the samples of the trajectory returns to its own neighborhood after τ samples delay, is given by the equation

$$P(\tau) = \frac{1}{N_s - \tau} \sum_{i=1}^{N_s - \tau} R_{i,i+\tau} = \frac{1}{N_s - \tau} \sum_{i=1}^{N_s - \tau} \Theta(\in - \| \overrightarrow{y_i} - \overrightarrow{y_{i+\tau}} \|) \tag{1.4}$$

$P(\tau)$ can be viewed as the probability with which the trajectory has a period k. By using the probability of recurrence $P(\tau)$ of two signals, it is possible to detect PS between any two signals by a measure called Correlation between Probabilities of Recurrence (CPR) [47]. Evaluation of CPR consists of two steps:

- Compute probabilities of recurrence $P_1(\tau)$ and $P_2(\tau)$ for the signals 1 and 2 respectively
- Compute the correlation coefficient between probabilities of recurrence

$$CPR = \frac{\sum_{\tau=\tau_e}^{\tau_m - 1}\{P_1(\tau) - m_1\}\{P_2(\tau) - m_2\}}{\sigma_1\sigma_2} \tag{1.5}$$

Where m_1 and m_2 are the mean, and, σ_1 and σ_2 are the standard deviations of $P_1(\tau)$ and $P_2(\tau)$ respectively. τ ranges from τ_e to τ_m. Since $P(\tau)$ always has a value of 1 for τ =0, CPR is computed only over the segment starting when $P(\tau)$ falls below 1/e. τ_e is the value of τ for which $P(\tau)$=1/e.

(2) Correlation-purged Granger Causality (CPGC): The principle of Granger causality is that if the past and present state of time series A is able to predict the future state of time series B, then A is said to Granger-cause B. CPGC extends this concept by separately modeling out the effect of instantaneous correlation and hence gives linear causal influence between different brain regions [39]. For k fMRI time series $Y(t) = [y_1(t)\ y_2(t) y_3(t) \dots\ y_k(t)]$, the traditional multivariate Vector Autoregressive (VAR) model of order p is defined as:

$$Y(t) = \sum_{i=1}^{p} C(n)Y(t-n) + \Delta(t) \tag{2.1}$$

Where $\Delta(t)$ is the model error, $C(n)$ are the model coefficients and is defined as:

$$C(n) = \begin{bmatrix} c_{11}{}^{(n)} & \cdots & c_{1k}{}^{(n)} \\ \vdots & \ddots & \vdots \\ c_{k1}{}^{(n)} & \cdots & c_{kk}{}^{(n)} \end{bmatrix} \tag{2.2}$$

Eq.(2.2) can be rewritten as:

$$\Delta(t) = Y(t) - \sum_{i=1}^{p} C(n)Y(t-n) \tag{2.3}$$

Applying frequency transformation on Eq.(2.2), we would have

$$\Delta(f) = Y(f)\left[\beta_{ij} - \textstyle\sum_{i=1}^{p} c_{ij}(n)e^{-i2\pi fn}\right] \tag{2.4}$$

Where $c(f)$ can be defined as:

$$c_{ij}(f) = \beta_{ij} - \sum_{i=1}^{p} c_{ij}(n)e^{-i2\pi fn} \tag{2.5}$$

Where c_{ij} is an element of $C(f)$ and β_{ij} is the Kroenecker-delta function. Therefore, the transfer matrix of the VAR model can be defined as:

$$H(f) = C^{-1}(f) \tag{2.6}$$

Hence the correlation-purged Granger Causality (CPGC) can be calculated as:

$$CPGC_{ij} = \sum_{f} h_{ij}{}'(f) \tag{2.7}$$

(3) Kernel Granger Causality (KGC): This is a nonlinear extension of Granger causality, similar to the linear case described above. Consider both a univariate (Eq.3.1 and a bivariate (Eq.3.2) linear autoregressive models between two fMRI time series x and y. (Note that $x_n{}'$ is not the derivative of x_n; rather $x_n{}'$ represents a different estimation of x_n. Same notion will follow through the rest of this section)

$$x_n = \sum_{j=1}^{m} \alpha_j \, x_{n-j} + e_n \tag{3.1}$$

$$x_n{}' = \sum_{j=1}^{m} \alpha_j{}' \, x_{n-j} + \sum_{j=1}^{m} \beta_j{}' \, y_{n-j} + e_n{}' \tag{3.2}$$

Where $\alpha_j, \alpha_j{}'$, and $\beta_j{}'$ are coefficients for the autoregressive models, m is the model order, and $e_n, e_n{}'$ are residual errors. Granger causality index γ for y causing x: $y \rightarrow x$ can be then defined as:

$$\gamma_{y\rightarrow x} = \frac{\sum_n \|e_n\|^2 - \sum_n \|e_n{}'\|^2}{\sum_n \|e_n\|^2} \tag{3.3}$$

Let’s define a couple of new vectors:

$$A_i = (x_i, \ldots x_{i+m-1})^T \tag{3.4}$$

$$A_i{}' = (x_i{}', \ldots x_{i+m-1}{}')^T \tag{3.5}$$

$$B_i = (A_i{}^T, A_i{}'^T)^T \tag{3.6}$$

$$\varphi = (x_{1+m}, \ldots x_{N+m})^T \tag{3.7}$$

$$\tilde{x}_j = \textstyle\sum_{j=1}^{m} \alpha_j \, x_{n-j} \tag{3.8}$$

$$\widetilde{x'}_j = \textstyle\sum_{j=1}^{m} \alpha_j{}' \, x_{n-j} + \sum_{j=1}^{m} \beta_j{}' \, y_{n-j} \tag{3.9}$$

We then can construct four matrices:

$$\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N) \tag{3.10}$$

$$\widetilde{X'} = \left(\widetilde{x'}_1, \widetilde{x'}_2, \dots, \widetilde{x'}_N\right) \tag{3.11}$$

$$C = A^T A \tag{3.12}$$

$$D = B^T B \tag{3.13}$$

Let H be the range of matrix C, H' be the range of matrix D. Then $\tilde{X}$ can be considered as the projection of φ on H, $\widetilde{X'}$ can be considered the projection of φ on H'. If we define P as the projector on the space H and P' as the projector on the space H', considering matrix multiplication as the process of calculating projection of one matrix on the other, we will have

$$\tilde{X} = P\varphi \tag{3.14}$$

$$\widetilde{X'} = P'\varphi \tag{3.15}$$

We can further define

$$\mu = \varphi - P\varphi \tag{3.16}$$

$$\mu' = \varphi - P'\varphi \tag{3.17}$$

Where μ is orthogonal to $\tilde{X}$ and represents the error e_n space, which should be orthogonal to H. μ' is orthogonal to $\widetilde{X'}$ and represents the error $e_n{}'$ space which, should be orthogonal to H'. Let's decompose H' into two orthogonal parts: $H' = H \oplus H^{\perp}$, where $H^{\perp}$ is orthogonal to H and corresponds to the additional features due to the inclusion of y in Eq.(3.2). Also let's define $P^{\perp}$ as the projector on $H^{\perp}$ space, then $P^{\perp}\mu$ should represent the projection of μ on $H^{\perp}$. Further $P^{\perp}\mu$

represents the difference vector between e_n and $e_n{}'$. Therefore we can calculate the numerator $\sum_n \|e_n\|^2 - \sum_n \|e_n{}'\|^2$ in Eq.(3.3) by calculating $\|P^{\perp}\mu\|^2$; If we normalize and zero-mean φ without losing generality, the denominator $\sum_n \|e_n\|^2$ in Eq.(3.3) can be calculated by $1 - \tilde{X}^T\tilde{X}$. Therefore Eq.(3.3) can be re-written as:

$$\gamma_{y\rightarrow x} = \frac{\|P^{\perp}\mu\|^2}{1 - \tilde{X}^T\tilde{X}} \tag{3.18}$$

$H^{\perp}$ can be spanned by chosen bases such as eigenvectors. Let's call the eigenvector in $H^{\perp}$ t_i. The projection $P^{\perp}\mu$ can be calculated using the Pearson correlation coefficients for μ and t_i by

$$r_i = \frac{(n\mu t_i - \sum \mu \sum t_i)}{\sqrt{[n\mu\mu^T - (\sum \mu)^2][n t_i{}^T t_i - (\sum t_i)^2]}} \tag{3.19}$$

Where n is the dimension for either in μ or t_i. We assume μ is a row vector and t_i is a column vector. Replace $P^{\perp}\mu$ with r_i, Eq.(3.18) can be re-written as Eq.(3.20)

$$\gamma_{y\rightarrow x} = \frac{\sum_i r_i{}^2}{1 - \tilde{X}^T\tilde{X}} \tag{3.20}$$

Applying Bonferroni correction to select significant $t_{i'}$ with false positive threshold to be 0.05, we can re-calculate Granger causality index by:

$$\gamma_{y\rightarrow x} = \frac{\sum_{i'} r_{i'}{}^2}{1 - \tilde{X}^T\tilde{X}} \tag{3.21}$$

For kernel Granger causality, we replace the linear autoregressive models in Eq.(3.1) and (3.2) with $k(x, x')$. Each step for deriving linear Granger causality should be correspondingly followed in the derivation of kernel Granger causality. Two commonly used kernels are worth mentioning:

(1) Inhomogeneous polynomial kernel:

$$k_p(x, x') = (1 + x^T x')^p \tag{3.22}$$

(2) Gaussian kernel:

$$k_\sigma(x, x') = \exp(-\frac{(x - x')^T (x - x')}{2\sigma^2}) \tag{3.23}$$

Further details can be obtained from Liao et al [57]. Specifically, we used a linear Kernel as well as an inhomogeneous polynomial nonlinear Kernel to obtain two different values for KGC.

CPR represents non-directional connectivity while CPGC and KGC represent directional connectivity. We obtained two measures of KGC using polynomial orders 1 (called KGC_par1, corresponding to linear Granger causality) and 2 (called KGC_par2, corresponding to linear Granger causality). The model order was chosen to be 5 for both CPGC and KGC using the Bayesian information criterion [58]. Latent variables (Eigen values) and principal components were extracted from the raw fMRI time series derived from 190 brain regions, as well as the 4 sets of connectivity metrics obtained from each subject using MATLAB. Consequently, two different feature sets were derived for each of the five metrics. First, we chose the top 20 latent variables, which explained most of the variance in the data, as feature inputs to the classifiers. Second, we performed a t-test to find principal components which were significantly different among the groups and chose the 200 most significant ones as input features.

III. Classification

(1) Neural Networks

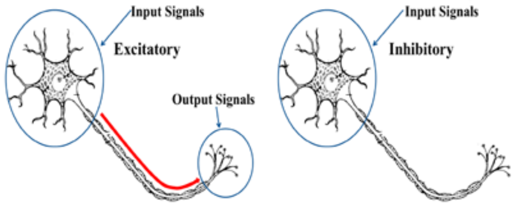

Fig.2.1 Biological neurons

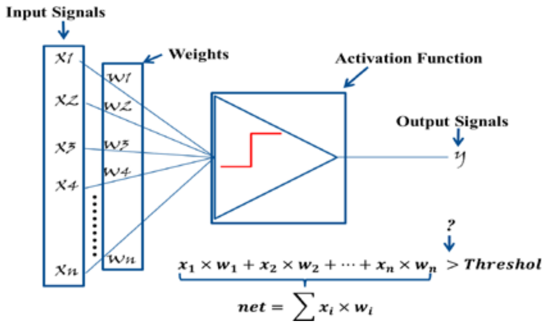
Input Signals
x1
x2
x3
x4
xn
w1
w2
w3
w4
wn
Weights
Activation Function
Output Signals
y
?
$x_1 \times w_1 + x_2 \times w_2 + \cdots + x_n \times w_n$ > Threshol
$net = \sum x_i \times w_i$

Fig.2.2 Artificial neurons

Artificial Neural Network, commonly known as Neural Network, consists of interconnecting artificial neurons that mimic the functionality of biological neurons such as firing upon receiving excitatory signals and not firing upon receiving inhibitory signals (illustrated in Fig.2.1). Therefore it possesses human-like decision making ability. A bipolar artificial neuron (or a bipolar perceptron) was originally designed as a classifier with linear activation functions, generating a positive output if the net value exceeded the threshold and a negative output otherwise (illustrated in Fig.2.2). The backbones of NN are training algorithms and architectures. Various types of NN training algorithms and architectures have been developed along the course of implementing NN on artificial computation. Training algorithms include Error Back Propagation (EBP), Levenberg-Marquardt (LM), Neuron by Neuron (NBN), etc. Architectures include Multi-Layer Perceptron (MLP) shown in Fig.1.3, Fully Connected Cascade (FCC) shown in Fig.1.2, Cascade Correlation (CC), etc. Multi-Layer Perceptron (MLP) Neural Network architecture was designed to handle nonlinear mapping between inputs and outputs, and is the most popular neural network architecture. It utilizes the Error Back Propagation (EBP) algorithm for training. For a single output case, EBP can be performed using

$$\Delta w_p = \alpha \sum_{p=1}^{P} \left[(d_p - o_p) f'(net_p) x_p \right] \tag{4.1}$$

where w is the weight vector, α is the learning constant, P is the number of input patterns, d is the desired output, o is the actual output, $f'(net_p)$ is the first derivative of the activation function *f* (normally hyperbolic), net is the output of *f*, and *x* is the input. This method becomes slow and has convergence issues when dealing with large data [23]. A few remedies were later developed including:

(i) EBP with variable learning rate [59]: If error increased by 5%, the updated values are ignored and learning constant is reduced; if error decreased by more than 5%, the learning constant is increased

(ii) Steepest descent EBP [60], defined as:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha g \tag{4.2}$$

where $\boldsymbol{w}_k$ is the weight vector, α is the learning constant, and g is the gradient vector given by

$$g = (\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots \frac{\partial E}{\partial w_n})^T \tag{4.3}$$

(iii) Newton method [61], defined as:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - A_k^{-1} g \tag{4.4}$$

where A_k is the Hessian matrix given by

$$A = \begin{pmatrix} \frac{\partial^2 E}{\partial {w_1}^2} & \frac{\partial^2 E}{\partial w_2 \partial w_1} & \cdots & \frac{\partial^2 E}{\partial w_n \partial w_1} \\ \frac{\partial^2 E}{\partial w_1 \partial w_2} & \frac{\partial^2 E}{\partial {w_2}^2} & \cdots & \frac{\partial^2 E}{\partial w_n \partial w_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_1 \partial w_n} & \frac{\partial^2 E}{\partial w_2 \partial w_n} & \cdots & \frac{\partial^2 E}{\partial {w_n}^2} \end{pmatrix} \tag{4.5}$$

where *E* is the output error, *n* is the number of weights.

(iv) Gauss-Newton algorithm [62], defined as:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - ({J_k}^T J_k)^{-1} {J_k}^T e \tag{4.6}$$

where $\boldsymbol{J}$ is the Jacobian matrix given by

$$\boldsymbol{J} = \begin{pmatrix} \frac{\partial e_{11}}{\partial w_1} & \frac{\partial e_{11}}{\partial w_2} & \cdots & \frac{\partial e_{11}}{\partial w_n} \\ \frac{\partial e_{21}}{\partial w_1} & \frac{\partial e_{21}}{\partial w_2} & \cdots & \frac{\partial e_{21}}{\partial w_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial e_{M1}}{\partial w_1} & \frac{\partial e_{M1}}{\partial w_2} & \cdots & \frac{\partial e_{M1}}{\partial w_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial e_{1p}}{\partial w_1} & \frac{\partial e_{1p}}{\partial w_2} & \cdots & \frac{\partial e_{1p}}{\partial w_n} \\ \frac{\partial e_{2p}}{\partial w_1} & \frac{\partial e_{2p}}{\partial w_2} & \cdots & \frac{\partial e_{2p}}{\partial w_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial e_{Mp}}{\partial w_1} & \frac{\partial e_{Mp}}{\partial w_2} & \cdots & \frac{\partial e_{Mp}}{\partial w_n} \end{pmatrix} \tag{4.7}$$

Where n is the number of weights, M is the number of outputs, and $\boldsymbol{e}$ is the vector for output error

$$\boldsymbol{e} = \begin{pmatrix} e_{\mathbf{11}} \\ e_{\mathbf{21}} \\ \vdots \\ e_{\boldsymbol{M1}} \\ \vdots \\ e_{\boldsymbol{1P}} \\ e_{\boldsymbol{2P}} \\ \vdots \\ e_{\boldsymbol{MP}} \end{pmatrix} \tag{4.8}$$

$\boldsymbol{e}$ is defined as

$$\boldsymbol{e} = \sum_{p=1}^{P} \sum_{m=1}^{M} (d_{pm} - o_{pm})^2 \tag{4.9}$$

Where d is the desired output, o is the actual output, M is the number of outputs and P is the number of inputs.

(v) LM training algorithm [63, 64], defined as:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - ({J_k}^T J_k \ + \mu I\ )^{-1} {J_k}^T e \tag{4.10}$$

Where μ is a learning parameter, and I is the identity matrix. Between Newton algorithm Eq.(4.4) and Gauss-Newton algorithm Eq.(4.6), the improvement is made on the Hessian matrix and the gradient matrix computation. Gauss-Newton algorithm calculates the quasi-Hessian matrix using the Jacobian matrix, which makes the computation faster since it is a first order derivative. For gradient matrix, derivative is not needed anymore, which saves more computation time. A_k^{-1} in Eq.(4.4) is replaced by $({J_k}^T J_k)^{-1}$ in Eq.(4.6). ${J_k}^T J_k$ is considered as the quasi-Hessian; g in Eq.(4.4) is replaced by ${J_k}^T e$ in Eq.(4.6). Notice that a learning parameter is added to the LM algorithm so that when learning constant is small, μI can be ignored comparing to ${J_k}^T J_k$; when μI is not small, it stays. This choice makes LM faster than the Gauss-Newton algorithm.

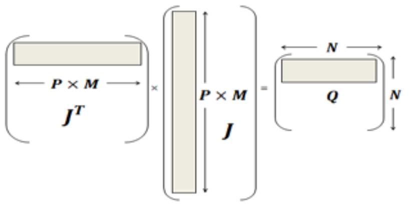

Fig.2.3 matrix multiplication routine 1

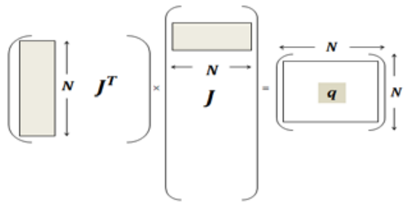

Fig.2.4 matrix multiplication routine 2

Traditional matrix multiplication $A \times B = C$ is done by multiplying one row from *A* with one column from *B* and producing a scalar for *C* (shown in Fig.2.3). However we can also perform multiplication for one column from *A* and one column from *B* (shown in Fig.2.4). If we examine the two routines carefully, we will find out that both of them require exactly the same numbers of operations (shown in Table 2.2 and 2.3). However for NN training only one row of Jacobian matrix is calculated when each input subject is applied. Therefore, the calculation of Hessian matrix can start right after the calculation of the first row and then finish after the calculation of the last row. This way, not only training time is shortened, but also imposes less storage requirement on a computer demonstrated in Table 2.4.

Number of Multiplication	**Number of Addition**
$(P \times M) \times N \times N$	$(P \times M - 1) \times N \times N$

Table 2.2 Number of operations needed for routine 1

Number of Multiplication	**Number of Addition**
$(N \times N) \times P \times M$	$(P \times M - 1) \times N \times N$

Table 2.3 Number of operations needed for routine 2

Multiplication routine	Number of elements for storage
Row-column (routine 1)	$(P \times M) \times N + N \times N + N$
Column-row (routine 2)	$N \times N + N$
Difference	$(P \times M) \times N$

Table 2.4 Storage for matrix multiplication routine 1 and 2

MLP uses LM training algorithm to reduce the computation cost by calculating Jacobian matrix instead of second derivative Hessian matrix. However there are two factors which can significantly increase MLP's computational complexity: (1) the interconnection among the neurons, (2) size of the input sample. Both these factors are predominant in fMRI based classification which may explain why SVMs are clearly preferred over MLP NNs in this field. A Neuron-By-Neuron method was developed by Wilamowski to address these issues. This method fully implemented the idea of interconnection by connecting each neuron to the output neuron while retaining the connections between contiguous neurons and shortened training time by incorporating a Neuron-by-Neuron architecture with improved LM training algorithm that implements matrix multiplication routine 2 (shown in Fig.4, compared with normal routine shown in Fig.1.3) to gain computation speed and storage advantages [23, 65]. Therefore, we have adopted this method in our study.

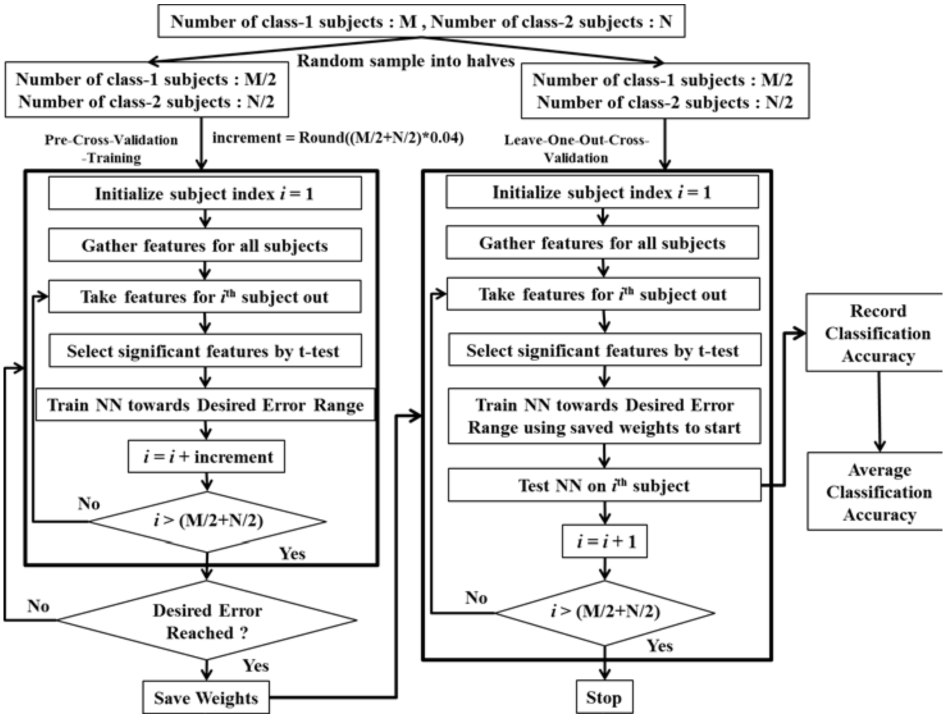
Number of class-1 subjects : M , Number of class-2 subjects : N
Random sample into halves
Number of class-1 subjects : M/2
Number of class-2 subjects : N/2
Number of class-1 subjects : M/2
Number of class-2 subjects : N/2
Pre-Cross-Validation -Training
increment = Round((M/2+N/2)*0.04)
Leave-One-Out-Cross-Validation
Initialize subject index i = 1
Gather features for all subjects
Take features for ith subject out
Select significant features by t-test
Train NN towards Desired Error Range
i = i + increment
No
i > (M/2+N/2)
Yes
No
Desired Error Reached ?
Yes
Save Weights
Initialize subject index i = 1
Gather features for all subjects
Take features for ith subject out
Select significant features by t-test
Train NN towards Desired Error Range using saved weights to start
Test NN on ith subject
i = i + 1
No
i > (M/2+N/2)
Yes
Stop
Record Classification Accuracy
Average Classification Accuracy

Fig.2.5 Proposed two-stage cross validation procedure. Left: Pre-Cross-Validation- Training; right: Leave-One-Out-Cross-Validation

Fig.2.5 illustrates a schematic of the proposed two-stage classification procedure consisting of Pre-Cross-Validation-Training and leave-one- out cross validation (LOOCV). We first randomly sampled the data into two equal halves. One half of the data was for Pre-Cross-Validation-Training, the other for LOOCV. In Pre-Cross-Validation-Training stage, we gathered features for all subjects and took features for the i^{th} subject out. Therefore, the training data consisted of features from all subjects except subject *i*. We selected significant features by t-test for principal components and connectivity path weights. For latent variables, t-test was not applied and the top 20 latent variables were chosen since they contained most energy. We trained the NNs towards the desired error rage using the training set. The subject index *i* was then incremented by 4% of the size of the training data. For example, if *i*=1 in the first iteration and the size of the training data is 500, then *i*=21 in the next iteration. We experimented with various values for the increment and found that an increment of around 4-5 % lead to more generalizability. After the subject index *i* iterated through the entire training data, we checked whether the desired error range (for example, ≤ 0.75) was reached. If yes, the corresponding weights were saved and passed onto the LOOCV loop. In LOOCV stage, the training data comprised of all but one subject in every loop. We repeated the same steps as we did in Pre-Cross-Validation-Training stage but using the saved weights to train the NN. Also, we tested the trained NN using the i^{th} subject, which was left out of training in each loop, as the testing set. Classification accuracy for each LOOCV iteration was recorded and averaged after LOOCV was finished.

In the Pre-Cross-Validation-Training loop, the root mean square training error was restricted to be ≤ 0.75 (+1 and -1 being the two labels representing the two groups of subjects) when both principal components and connectivity path weights ranked by the t-test were used as the feature inputs and ≤ 0.70 when latent variables were the feature inputs. In the training stage of the LOOCV loop, the root mean square error was also restricted to be ≤ 0.75 for principal components and connectivity path weights and ≤ 0.70 for latent variables on the training data while the error was restricted to be ≤ 90% of the root mean square training error on testing data. The smaller the range of training error, higher the classification accuracy for the given sample, but the NN may lose broader generalizability. We tried the error range from 0.25 to 0.90 by 0.05 increments. When the training error was 0.25 and 0.90, the trained NN produced ≤ 25 % classification accuracy. When the training error was in the range of [0.7,0.75], the classification accuracy was boosted over chance. We chose 0.75 for principal components and connectivity path weights, and 0.70 for latent variables because it showed least standard deviation and >80% accuracy, indicating generalizability. It also was reasonable to confine the testing error a little more than the training error in order to be conservative. For the error on testing data, we tried from 95% to 80% of the training error threshold by 5% decrements. 90% generally produced the best results with least standard deviation. Instead of starting with random weights for each LOOCV iteration, this two-stage classification process demanded that the NN have to be trained previous to cross validation to get a set of weights, which generally met the required training error range. Therefore, each LOOCV loop started with the saved weights from the Pre-Cross-Validation-Training stage, not the weights from the previous LOOCV iteration. However, note that the desired training error range was not always met. In such a situation, the trained NN weights were still accepted because in the LOOCV stage, we restricted the testing error to be within 90% of the training error, which guaranteed the result to be conservative. These steps ensured complete separation of training and testing data. Therefore the results obtained were conservative.

The NBN NN architecture we used consisted of only bipolar neurons (with +1 and -1 being the two labels representing the two groups of subjects being compared at once, bipolar neuron was the best choice for outputs) with hyperbolic activation function (hyperbolic activation function was preferred over linear activation function because the output necessarily did not have a linear range), and formed a fully connected cascade (FCC) architecture (example of FFC architecture shown in Fig.1). This NN architecture was trained using the modified NBN software [66]. We used a fully connected cascade deep NN architecture which had 200=1=1=1=1=1=1=1=1 configuration (please refer to Fig.1 for an illustration) for both the stages of classification when the top ranked 200 principal components were used, and 20=1=1=1=1=1 configuration when the largest 20 latent variables were used. For comparison, we also performed classification using a multi-layer perceptron (MLP) NN which is traditionally used in many applications. We used 20-10-10-10-1 configuration (refer to Fig.1.3) MLP architecture for the top 20 latent variables, 200-10-10-10-10-1 configuration (refer to Fig.1.3) for the top ranked 200 principal components. However the MLP architecture did not converge.

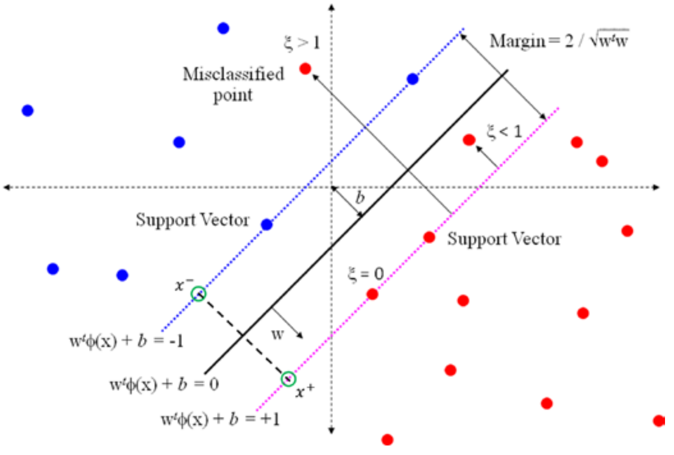

Fig.2.6 2D SVMs

(2) Support Vector Machines (SVMs)

SVM classifier was introduced in 1998. It is supervised learning model with associated learning algorithms that analyze data and recognize patterns. To classify data points into linear separable data sets, SVMs find out the optimal line for 2D space, plane for 3D space, and hyperplane for any higher dimension space by minimizing an upper bound of the training errors and maximizing the margin between the separating lines, planes, or hyperplanes. SVM is non-probabilistic classifier and well suited for both classification and regression analysis. For instance, given a set of training examples, each marked as belonging to one of two categories, a two-class SVM classifier is trained to build a model that assigns new examples into one category or the other. SVMs model represents the examples as mapped points in space so that the examples of the separate categories are divided by a clear gap which is as wide as possible (shown in Fig.2.6). New examples are mapped into the same space and predicted to belong to a category based on which side of the gap they fall on. SVMs essentially solve optimization problems.

Given a training set of the form (x_i, y_i) with x_i being the data points and y_i being the class labels, the SVMs solve the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \tag{5.1}$$

subject to $y_i(w \times x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where ξ_i is the slack variable, measuring the degree of a data point's misclassification, w is the weight defining the hyperplane and $C > 0$ is the penalty parameter of the error term. The first term in Eq.(5.1) is the reciprocal of the margin, and can be derived as follows: select a data point x^-, indicated by a green circle, on the negative separating line in Fig.16. Then select the corresponding data point x^+, indicated by a green circle, on the positive separating line. The connection of the two data points should be perpendicular to the optimal line in the middle. From Fig.7, we know that

$$\boldsymbol{w}\boldsymbol{x}^- + b = -1 \tag{5.2}$$

$$\boldsymbol{w}\boldsymbol{x}^+ + b = +1 \tag{5.3}$$

Let's further define

$$\boldsymbol{x}^+ - \boldsymbol{x}^- = \lambda\boldsymbol{w} \tag{5.4}$$

Substitute Eq.(5.4) back into Eq.(5.2), we get

$$\boldsymbol{w}(\boldsymbol{x}^+ - \lambda\boldsymbol{w}) + b = -1 \tag{5.5}$$

From Eq.(5.5), we can solve for λ:

$$\lambda = \frac{2}{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}} \tag{5.6}$$

The margin M can be calculated as:

$$M = \|\boldsymbol{x}^+ - \boldsymbol{x}^-\| = \|\lambda\boldsymbol{w}\| = \lambda\|\boldsymbol{w}\| = \lambda\sqrt{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}} = \frac{2}{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}}\sqrt{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}} = \frac{2}{\sqrt{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}}} \tag{5.7}$$

Applying *Lagrangian Multiplier*, Eq.(5.1) can be transformed into Eq.(5.8) defined below:

$$L_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \, y_i(x_i w + b) + \sum_{i=1}^{l} \alpha_i \tag{5.8}$$

We need to minimize L_P with respect to w and b, and simultaneously require that the derivative of L_P with respect to all the α_i vanish. This means we can equivalently solve the following "dual" problem: maximize L_P, subject to the constraints that the gradient of L_P with respect to $w \; and \; b$ vanish.

Take the partial derivative of L_P with respect to w, we get Eq.(5.9) as below:

$$\frac{\partial L_P}{\partial w} = \frac{1}{2} 2\|w\| - \sum_{i=1}^{l} \alpha_i \, y_i(x_i + 0) + 0 \tag{5.9}$$

Take the partial derivative of L_P with respect to b, we get Eq.(5.10) as below:

$$\frac{\partial L_P}{\partial b} = 0 - \sum_{i=1}^{l} \alpha_i \, y_i(0 + 1) + 0 \tag{5.10}$$

For w to vanish in Eq.(5.9), Eq.(5.11) has to be true:

$$w = \sum_{i=1}^{l} \alpha_i \, y_i x_i \tag{5.11}$$

For b to vanish in Eq.(5.10), Eq.(5.12) has to be true:

$$\sum_{i=1}^{l} \alpha_i \, y_i = 0 \tag{5.12}$$

The decision function implemented by SVMs is originally written as:

$$f(x) = sign(wx \; + b) \tag{5.13}$$

Substitute Eq.(5.11) and (5.12) into Eq.(5.13), we get Eq.(5.14) as below:

$$f(x) = sign\left(\sum_{i}^{n} y_i \, \alpha_i x x_i \; + b\right) \tag{5.14}$$

The decision function implemented by SVM can be written as:

$$f(x) = sign(\textstyle\sum_{i}^{n} y_i \, \alpha_i K(x, x_i) \; + b) \tag{5.15}$$

Where $K(x_i, x_j)$ is the kernel function. We used the radial basis kernel (RBF) of the form

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left(x_i - x_j\right)^2}{2\sigma^2}\right) \tag{5.16}$$

In the study, I used *Spider* [67], a MATLAB toolbox for implementing the RBF-based SVM explained above using the following parameters: soft margin C = Infinity and RBF sigma = 11. Fig.2.7 shows a schematic for this procedure wherein we first sampled half of class 1 and class 2 subjects (for example, 372 from 744 TDC subjects and 130 from 260 ADHD combined subjects). Though this is strictly not required for LOOCV of SVM, we did it in order to keep the data sizes consistent with our NN procedure and so that the results across both methods are comparable. Second, we conducted LOOCV. Inside the LOOCV loop, we gathered features for

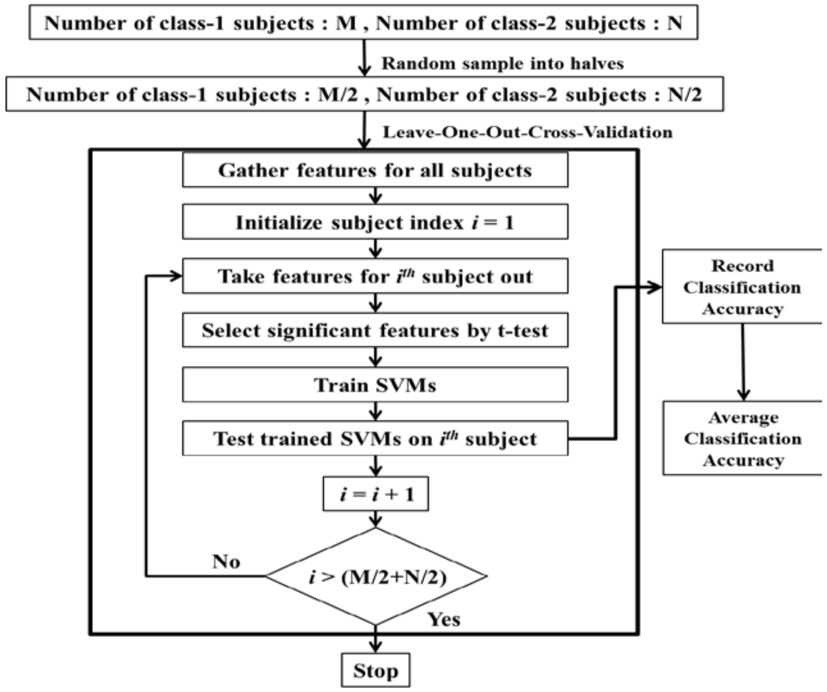
Number of class-1 subjects : M , Number of class-2 subjects : N
Random sample into halves
Number of class-1 subjects : M/2 , Number of class-2 subjects : N/2
Leave-One-Out-Cross-Validation
Gather features for all subjects
Initialize subject index $i = 1$
Take features for i^{th} subject out
Select significant features by t-test
Train SVMs
Test trained SVMs on i^{th} subject
$i = i + 1$
No
$i > (M/2+N/2)$
Yes
Stop
Record Classification Accuracy
Average Classification Accuracy

Fig.2.7 Proposed SVMs classification procedure

all subjects and took features for the i^{th} subject out to get the training set and testing set. We selected significant features by t-test for principal components and connectivity path weights. For latent variables, t-test was not applied and the top 20 latent variables were chosen because they contain the most energy. We trained the SVMs using the training set and tested the trained SVMs using the i^{th} subject. Classification accuracy for each LOOCV iteration was recorded and averaged after LOOCV was finished.

IV. Features Important for Classification

While it is possible to find the features which are most discriminative, and hence, important for classification, those features do not always carry meaningful interpretation. For example, the latent variables and principal component features cannot be interpreted because they are drawn from the entire data set. However, raw features, i.e. connectivity path weights, represent specific interactions between brain regions can be informative for inferring the underlying neuronal alterations in ADHD. Therefore, we ranked the directional connectivity features which gave maximum accuracy using the following procedure. First, we calculated the number of times each feature was picked across all iterations and divided that number by the maximum number of LOOCV iterations. This gave the frequency of occurrence of features across all iterations with a range [0,1]. We picked the top ranked paths for each comparison. They were 1, 0.9989 and 0.9931 for comparisons of TDC & ADHD combined, TDC & ADHD inattentive and ADHD combined & ADHD inattentive, respectively. This guaranteed that the picked features occurred in almost all iterations.

The most discriminative features represent a metric of generalizability, but they do not necessarily indicate statistical separation between the classes. Therefore, we also performed a t-test using the entire subject sample and created a mask of features which were most significantly different between the groups ($p<0.0001$). This mask was applied to the ranked features and the surviving features were used to infer differences between the groups.

Chapter 3 Results & Discussion

I. Results

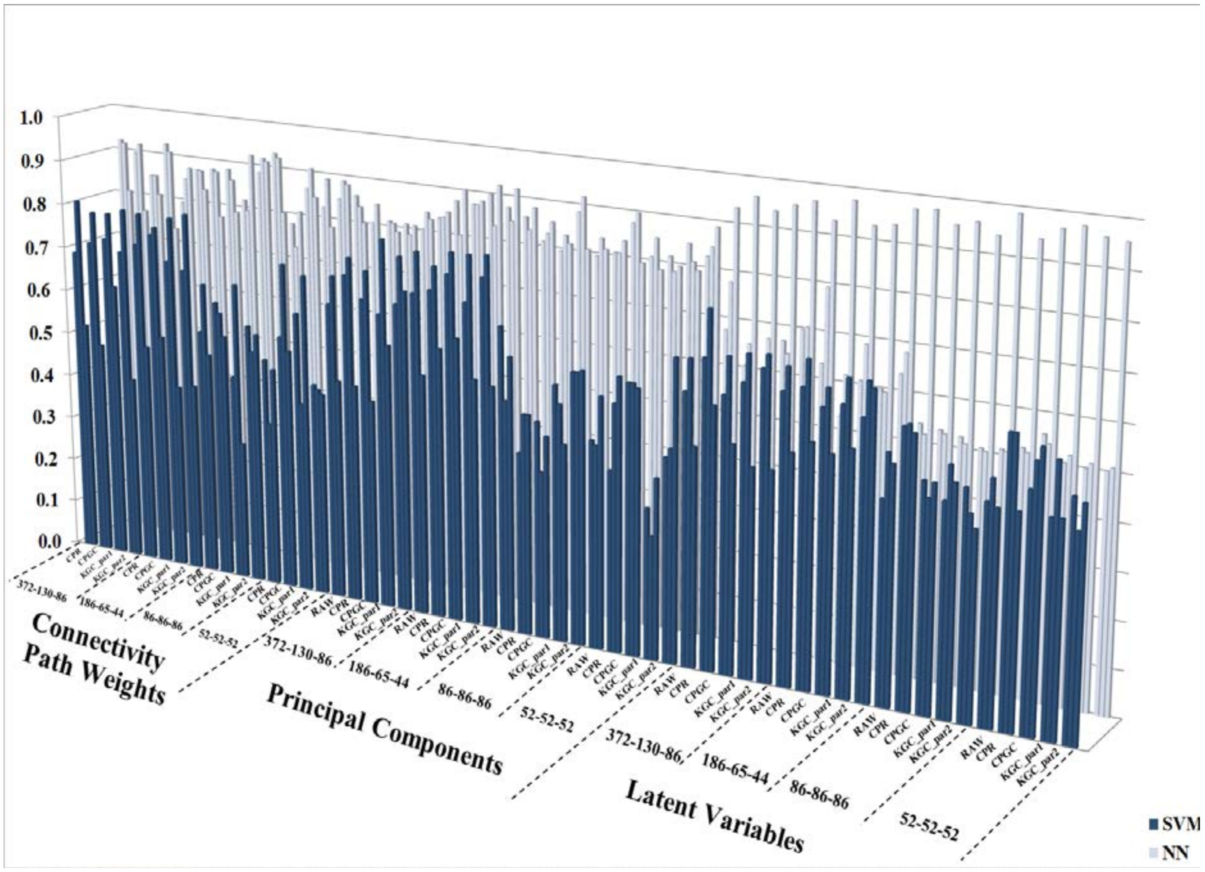

Fig.3.1 Comparison of classification accuracies between SVM and NN. There are 3 feature categories: (1) Raw Connectivity Path Weights, (2) Principal Components, and (3) Latent Variables. In each feature category, there are 4 size categories representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 372-130-86, (2) 186-65-44, (3) 86-86-86, (4) 52-52-52. In each size category, there are features obtained from raw data, CPR, CPGC, KGC_par1 and KGC_par2. For each feature, the average accuracies from left to right represent the following three pairwise group comparisons: TDC & ADHD combined, TDC & ADHD inattentive, ADHD combined & ADHD inattentive

SVM versus NN: As shown in Fig.3.1 that FCC deep architecture NN classifier consistently gives higher classification accuracies than SVM classifiers do across all combinations. For Connectivity Path Weights and Principal Components, SVMs classifier's performance is not consistently high. In contrast, NN classifier's performance is consistently high and stable. For Latent Variable features obtained from PCA, NN classifier's accuracies between ADHD combined & ADHD inattentive are significantly better than that of SVM.

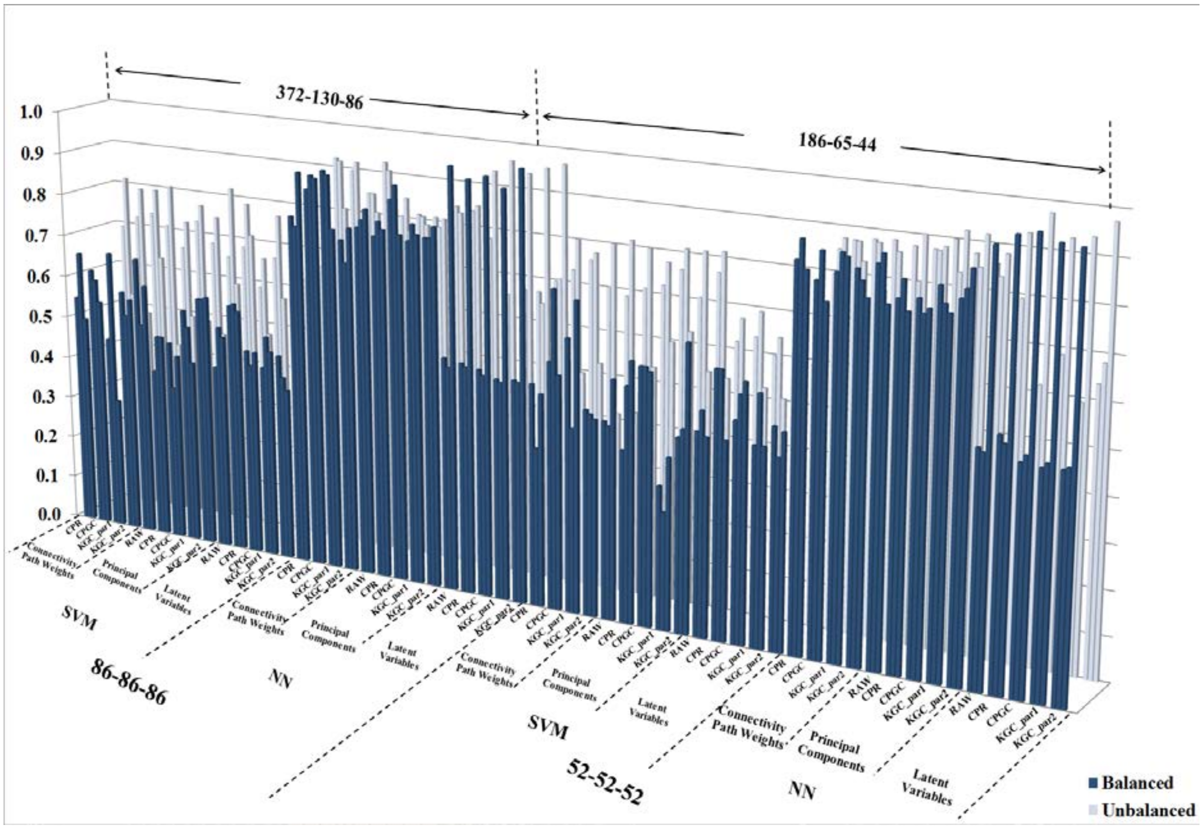

Fig.3.2 Comparison of classification accuracies between balanced and unbalanced data. There are 4 sample size categories representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 86-86-86 (front left), (2) 52-52-52 (front right), (3) 372-130-86 (back left), (4) 186-65-44 (back right); two classifier categories: (1) SVM, (2) NN; 3 feature categories: (1) Connectivity Path Weights, (2) Principal Components, (3) Latent Variables. In each feature category, there are features obtained from raw data, CPR, CPGC, KGC_par1 and KGC_par2. For each feature, the average accuracies from left to right represent the following three pairwise group comparisons: TDC & ADHD combined, TDC & ADHD inattentive, ADHD combined & ADHD inattentive

Balanced versus Unbalanced Sample Size: For SVM, unbalanced data (back row), i.e. unequal number of subjects in each class, generate better classification accuracies than balanced data (front row) as demonstrated in Fig.3.2. For NN classifier, balanced or unbalanced sample sizes generally do not make much difference except when latent variables are used as features.

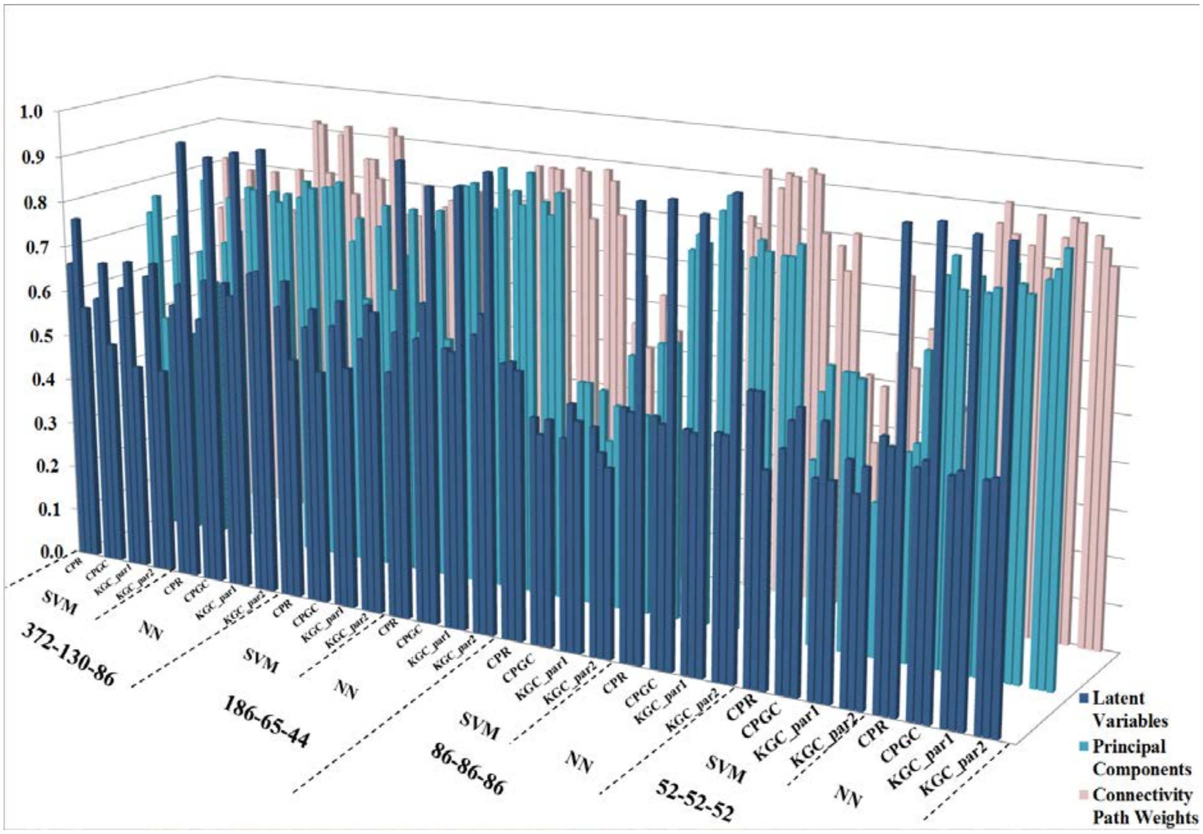

Fig.3.3 Comparison of accuracies between feature categories of latent variables (front row), principal components (middle row), and connectivity path weights (back row). From left to right, there are 4 sample size categories representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 372-130-86, (2) 186-65-44, (3) 86-86-86 , (4) 52-52-52 ; two classifier categories: (1) SVM, (2) NN; four feature categories: (1) CPR, (2) CPGC, (3) KGC_par1, (4) KGC_par2. For each feature, the average accuracies from left to right represent the following three pairwise group comparisons: TDC & ADHD combined, TDC & ADHD inattentive, ADHD combined & ADHD inattentive

Feature Category (latent variables vs principal components vs raw connectivity): It can be seen in Fig.3.3 that Principal Components (middle row) and Connectivity Path Weights (back row) generally provide better classification accuracies than Latent Variables (front row) do. For ADHD combined & ADHD inattentive comparison, all 3 feature categories provide very high accuracy.

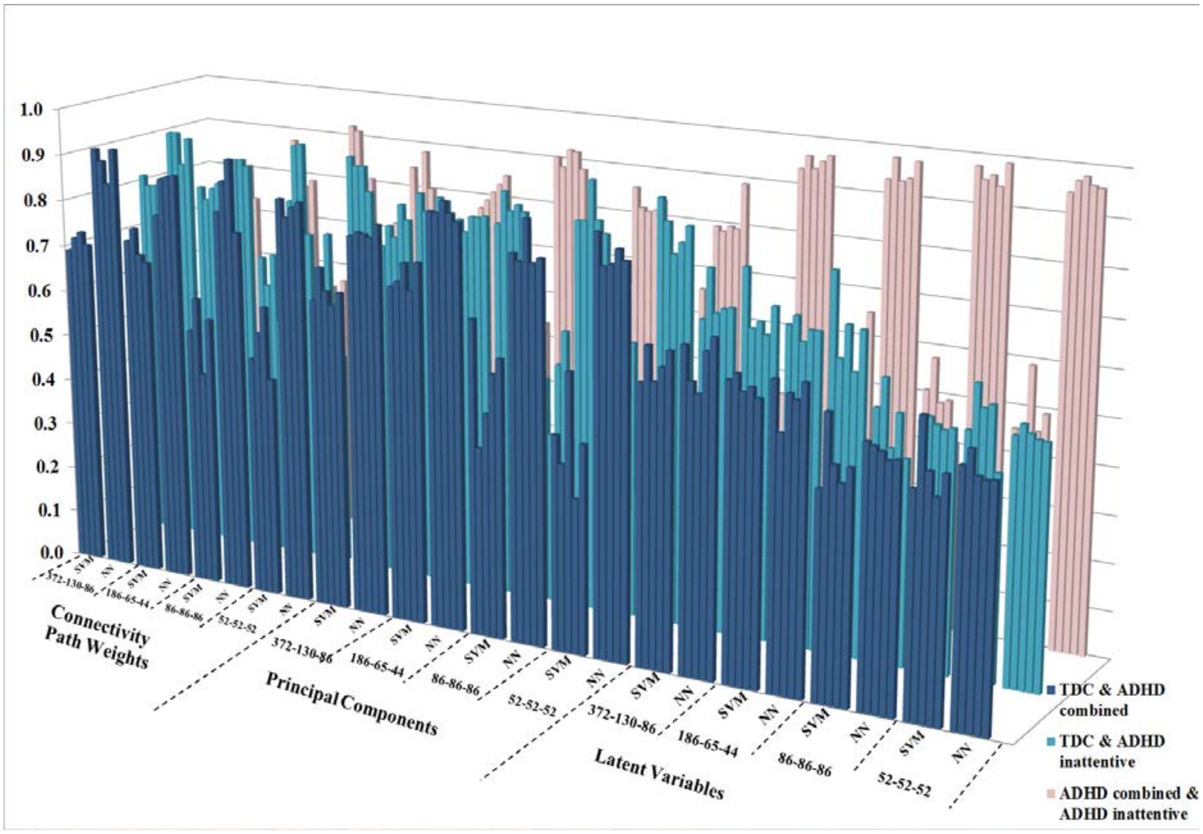

Fig.3.4 Comparison of accuracies obtained from TDC & ADHD combined (front row), TDC & ADHD inattentive (middle row), and ADHD combined & ADHD inattentive (back row). There are 3 feature categories: (1) Connectivity Path Weights (front row left), (2) Principal Components (front row middle), (3) Latent Variables (front row right); 4 sample size categories from left to right representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 372-130-86, (2) 186-65-44, (3) 86-86-86 , (4) 52-52-52 ; two classifier categories: (1) SVM, (2) NN. Note that for Connectivity Path Weights, the accuracy bars represent 4 features: (1) CPR, (2) CPGC, (3) KGC_par1, (4) KGC_par2; for Principal Components and Latent Variables, the accuracy bars include performance using raw intensities of fMRI images in addition to the 4 connectivity-based features.

TDC and ADHD group comparisons: For both SVMs and NN classifiers, the accuracies for classifying TDC group from ADHD groups were generally better than the accuracy for classifying between ADHD sub-groups in Fig.3.4. In fact, connectivity path weights with NN gave over 90% accuracy in classifying between TDC & ADHD inattentive, and between TDC & ADHD combined. However, NN classifier with latent variable features performed extremely well (close to 95%) in classifying between the ADHD subgroups.

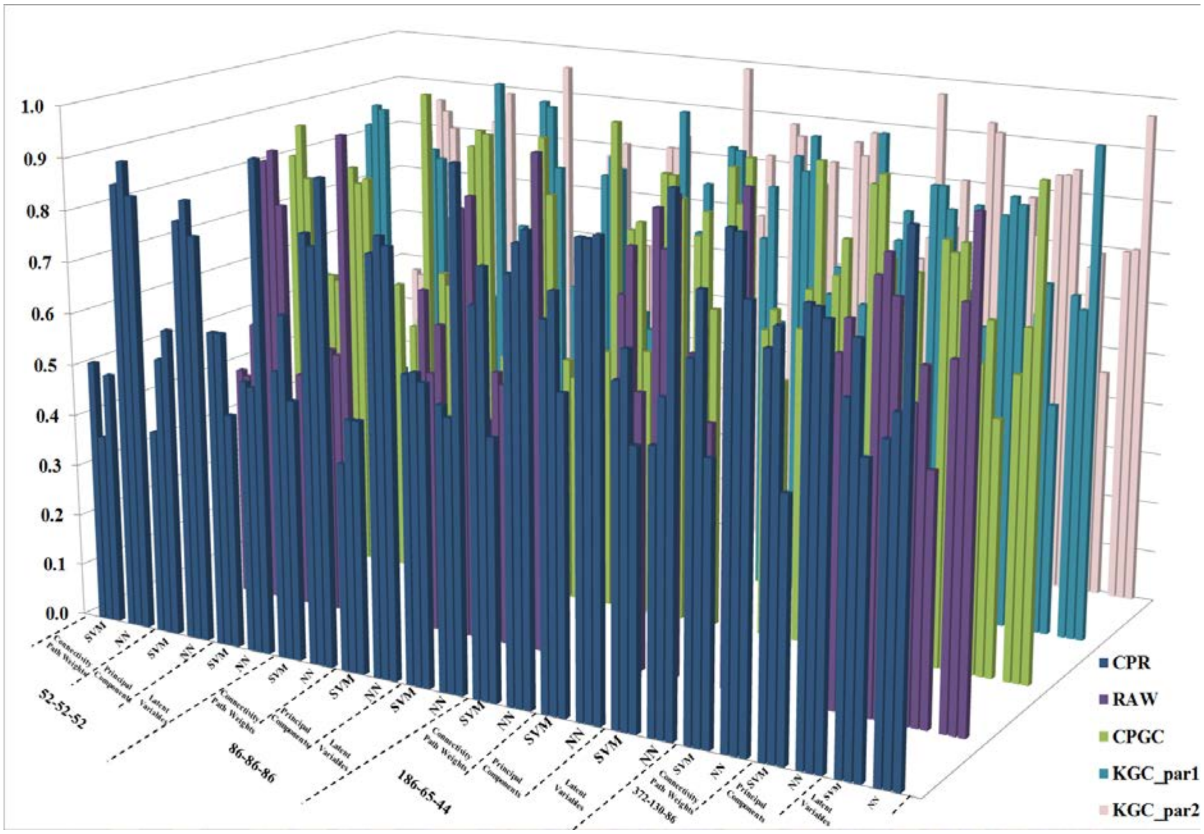

Fig.3.5 Comparison of accuracies obtained from the 5 types of features (1) RAW data, (2) CPR, (3) CPGC, (4) KGC_par1, (5) KGC_par2. There are 4 sample sizes from left to right representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 52-52-52, (2) 86-86-86, (3) 186-65-44, (4) 372-130-86; 3 feature categories from left to right: (1) Connectivity Path Weights, (2) Principal Components, (3) Latent Variables; two classifier categories: (1) SVM, (2) NN.

Comparisons between performances across different features: From Fig.3.5, there is no significant trend showing one connectivity-based feature outperforming others. In order to illustrate this, we sorted the accuracies for each feature and plotted them in Fig.3.6. It can be seen that all connectivity path weights gave similar classification accuracies. However raw image intensities generally gave lower classification accuracies than the connectivity path weights.

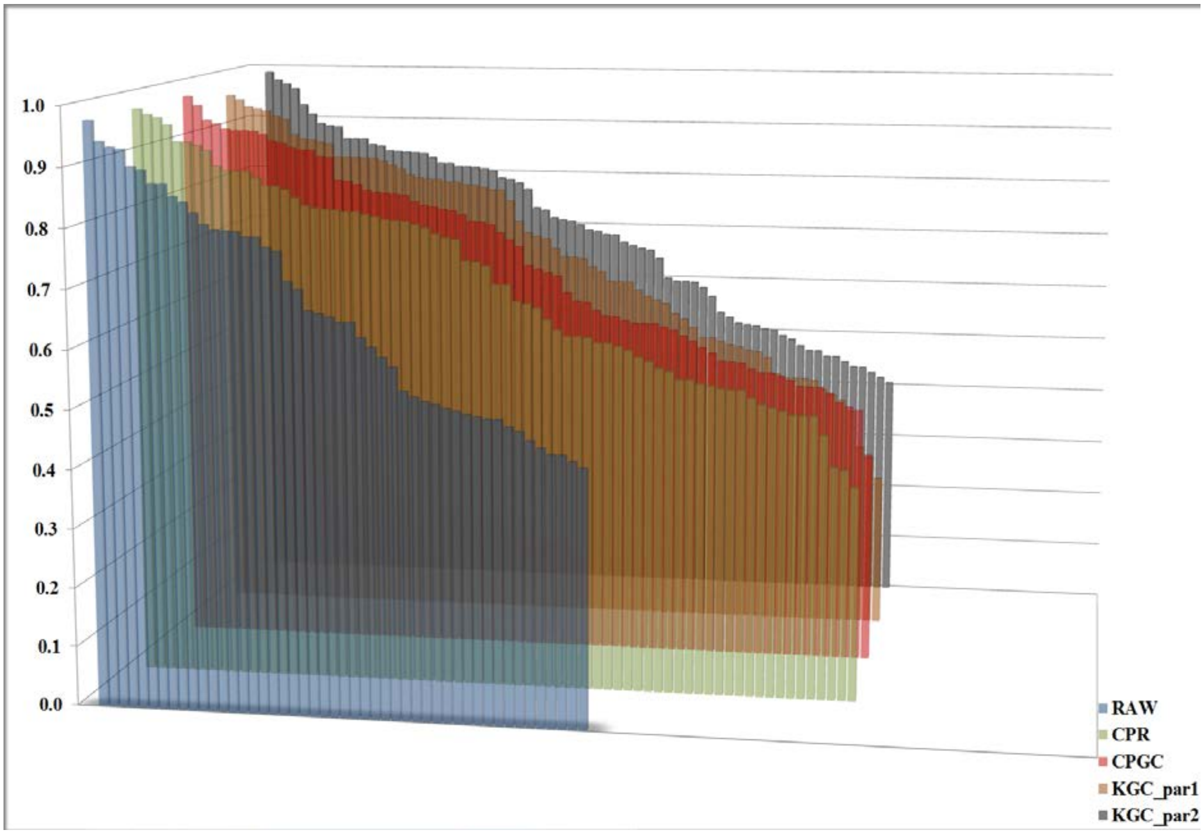

Fig.3.6 Comparison of accuracies obtained from the following features: (1) RAW, (2) CPR, (3) CPGC, (4) KGC_par1, (5) KGC_par2. All accuracies have been sorted in descending order. Note that for raw data, we only had features based on principal components and latent variables as against connectivity-based features wherein even the path weights were used as features. Therefore the total number of features for the former is lesser than the latter.

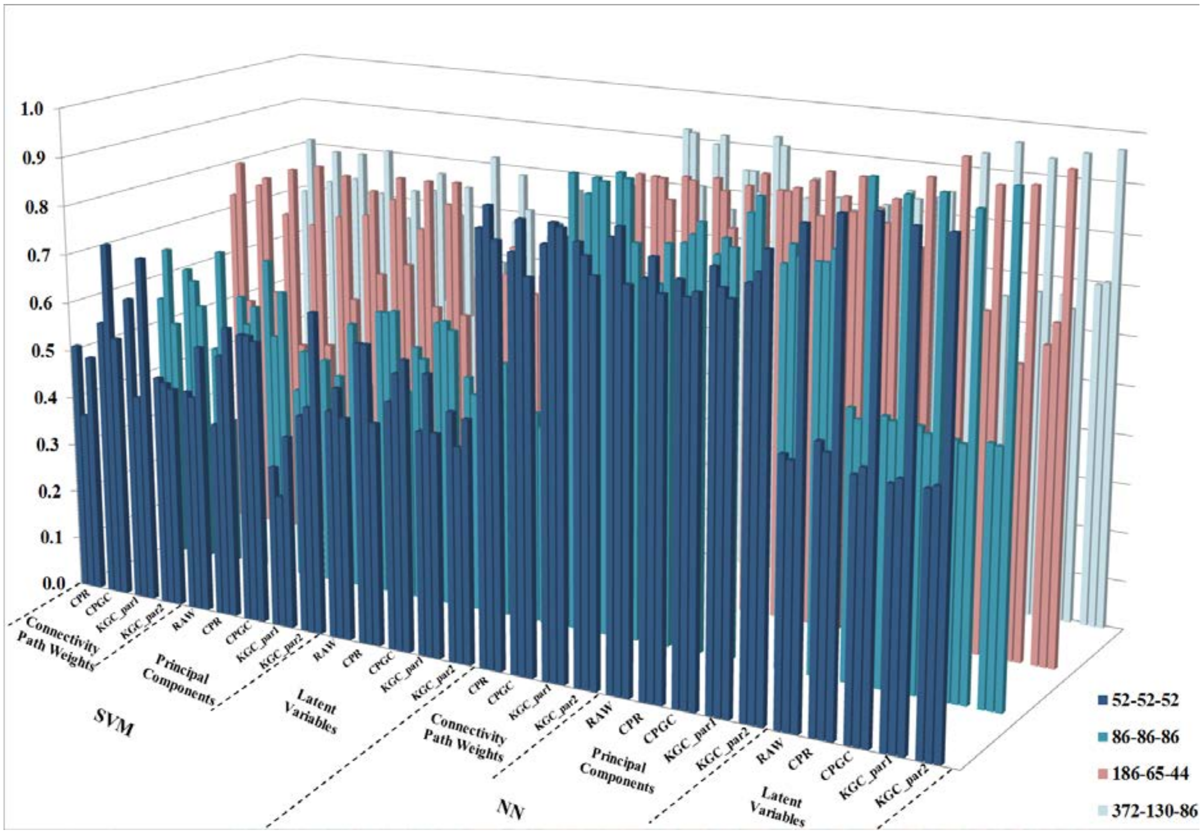

Fig.3.7 Comparison of accuracies for different sample sizes representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 52-52-52, (2) 86-86-86, (3) 186-65-44, (4) 372-130-86. There are two classifier categories: (1) SVM, (2) NN. For each of them, there are 3 feature categories from left to right: (1) Connectivity Path Weights, (2) Principal Components, (3) Latent Variables. Note that for Connectivity Path Weights, the accuracy bars correspond to the following features: (1) CPR, (2) CPGC, (3) KGC_par1, (4) KGC_par2 whereas for Principal Components and Latent Variables, there is an additional accuracy bar for raw data.

Effect of Sample Size on Performance: As seen from Figs.3.7 and 3.8. Larger sample sizes generally gave higher classification accuracy. The exception to this was the NN classification accuracies for ADHD combined & ADHD inattentive, especially with Latent Variables, wherein the accuracy obtained by lower sample sizes was comparable to those obtained from larger sample sizes. However, it is noteworthy that the peak accuracies obtained from different sample sizes are not significantly different from each other.

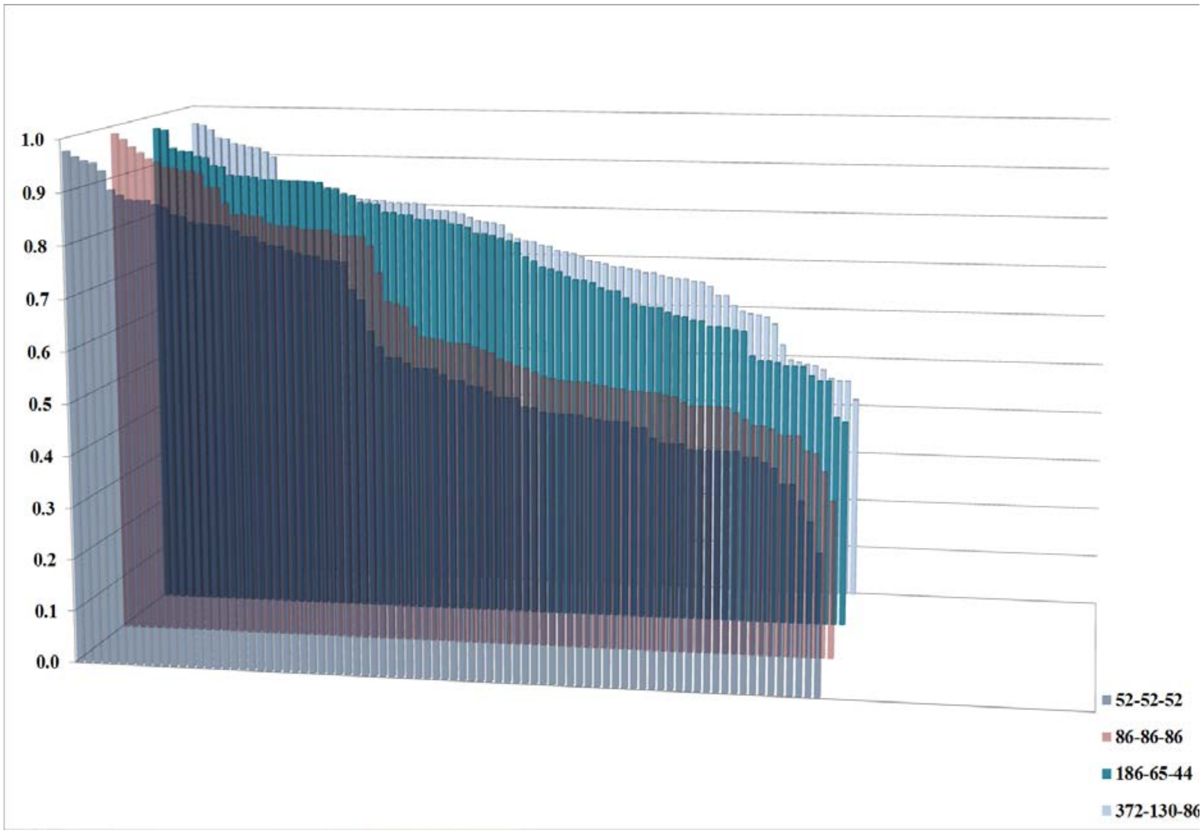

Fig.3.8 Comparison of accuracies for different sample sizes representing the number of subjects taken from TDC, ADHD combined and ADHD inattentive groups: (1) 52-52-52, (2) 86-86-86, (3) 186-65-44, (4) 372-130-86, with the accuracies having been sorted in descending order.

Features Important for Classification: Figures 3.9, 3.10 and 3.11 show the directional connectivity features with the highest discriminative power obtained by the ranking procedure described in the methods section for the following comparisons between the groups, respectively: TDC & ADHD combined, TDC & ADHD inattentive, ADHD combined & ADHD inattentive. Fig.3.12 shows the regions which drove L mid orbitofrontal cortex and which were greater in TDC as compared to ADHD combined as well as those which drove the same region, but were greater in ADHD combined as compared to TDC. It can be seen that differential inputs to the L mid orbitofrontal cortex and higher drive to the Vermis region of the cerebellum mainly distinguishes these two groups. On the other hand, large numbers of inputs to many cerebellar regions were higher in TDC as compared to ADHD inattentive and this was discriminative. Finally, higher drive of L Mid orbitofrontal cortex in ADHD combined distinguished it from the ADHD inattentive group.

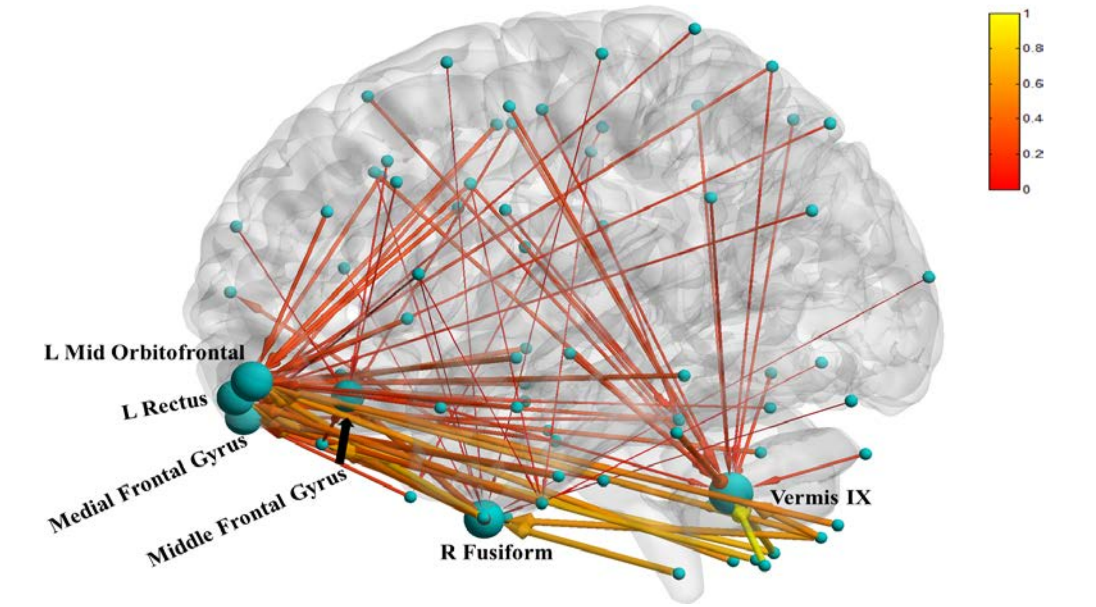

Fig.3.9.a

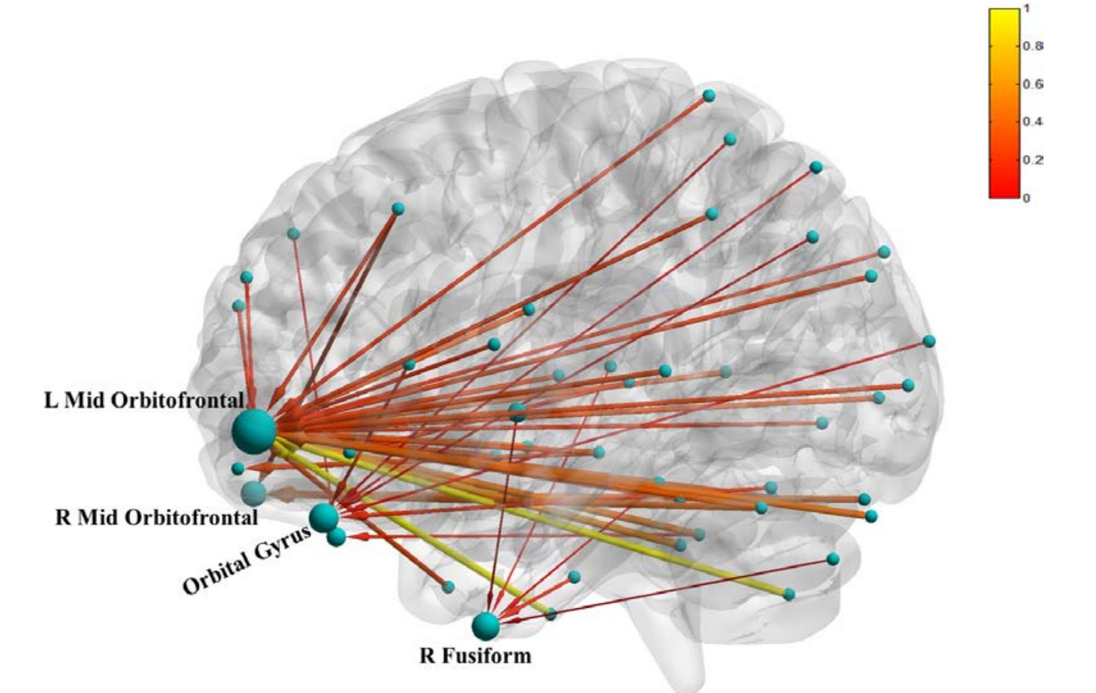

Fig.3.9.b

Top ranked connectivity paths which were greater in TDC as compared to ADHD combined group (Fig.15a) and vice versa (Fig.15b). The width and the color of each path represent the rank of the path, with thinner and darker color path representing lower rank, bigger and lighter color path representing higher rank. The radius of the nodes is proportional to the number of inputs/outputs to/from them.

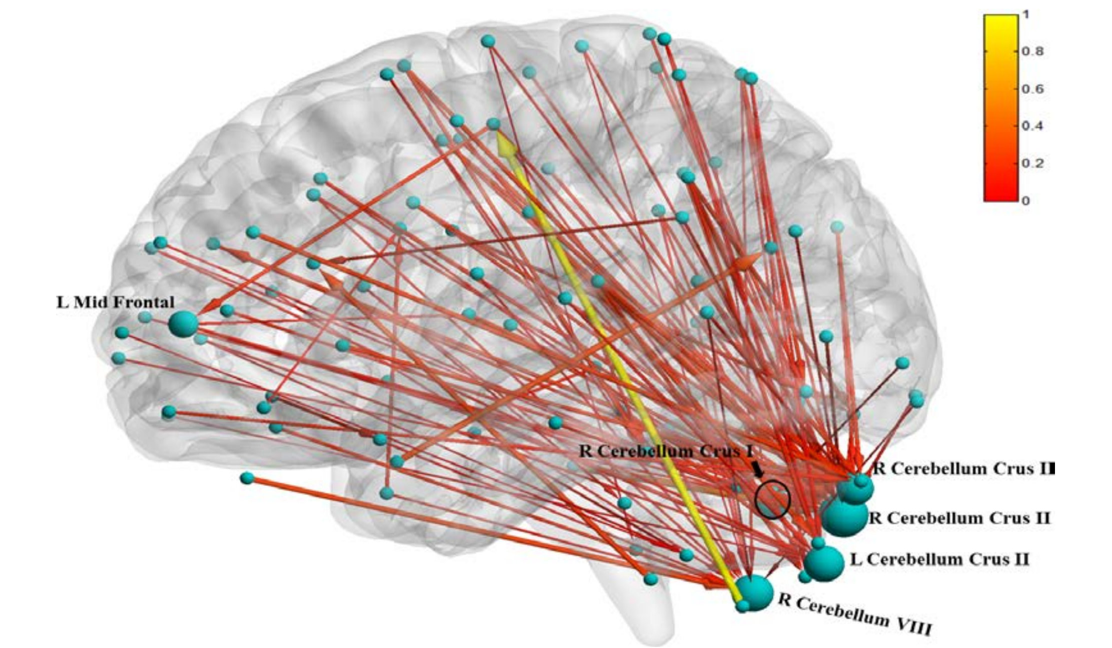

Fig.3.10.a

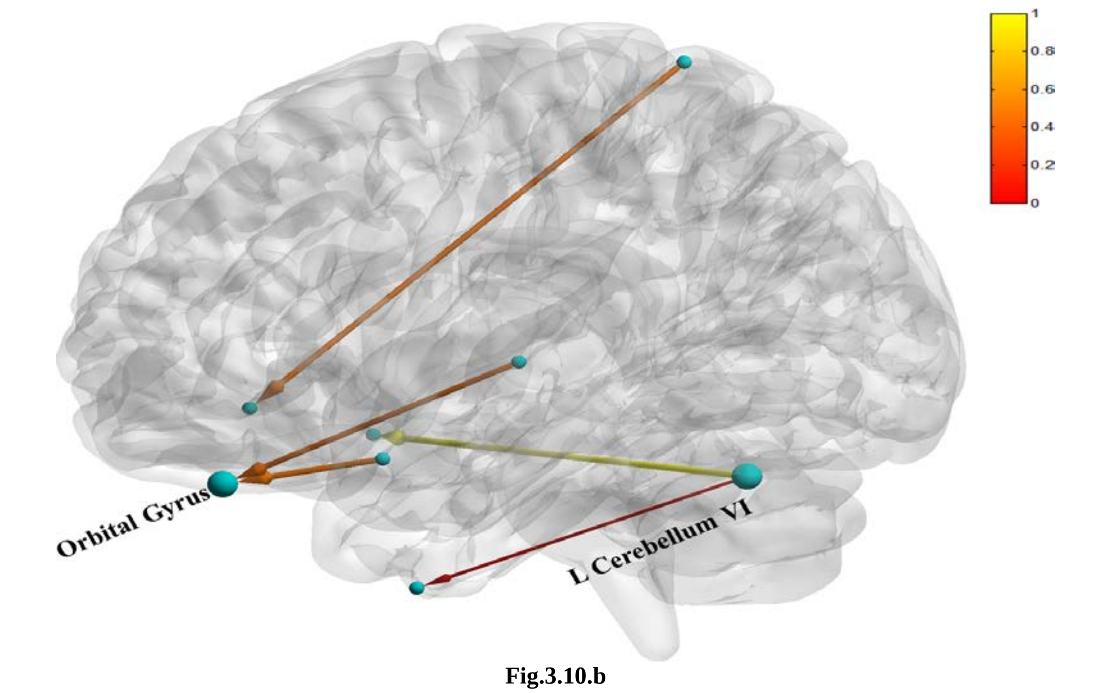

Fig.3.10.b

Top ranked connectivity paths which were greater in TDC as compared to ADHD inattentive group (Fig.16a) and vice versa (Fig.16b). The width and the color of each path represent the rank of the path, with thinner and darker color path representing lower rank, bigger and lighter color path representing higher rank. The radius of the nodes is proportional to the number of inputs/outputs to/from them.

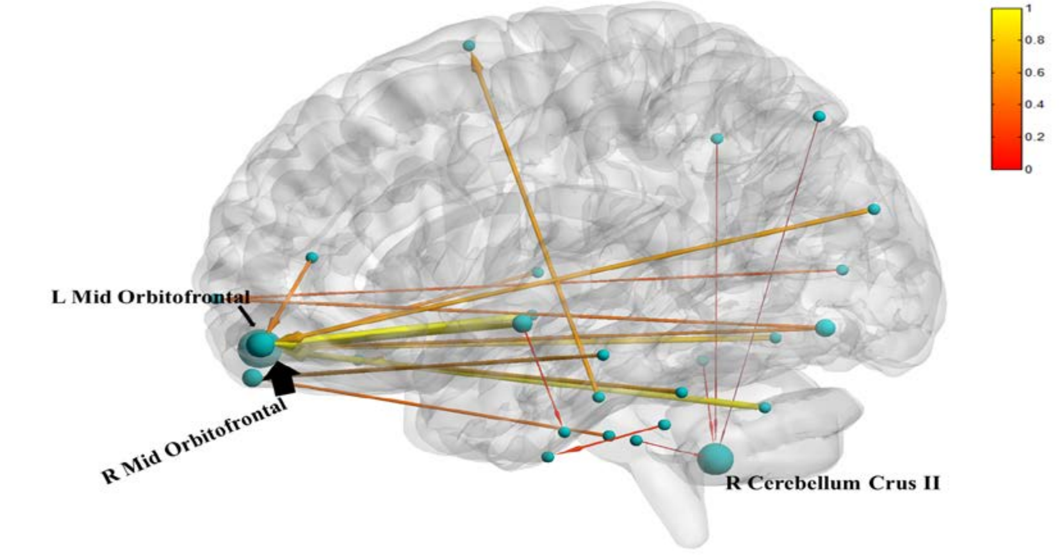

Fig.3.11.a

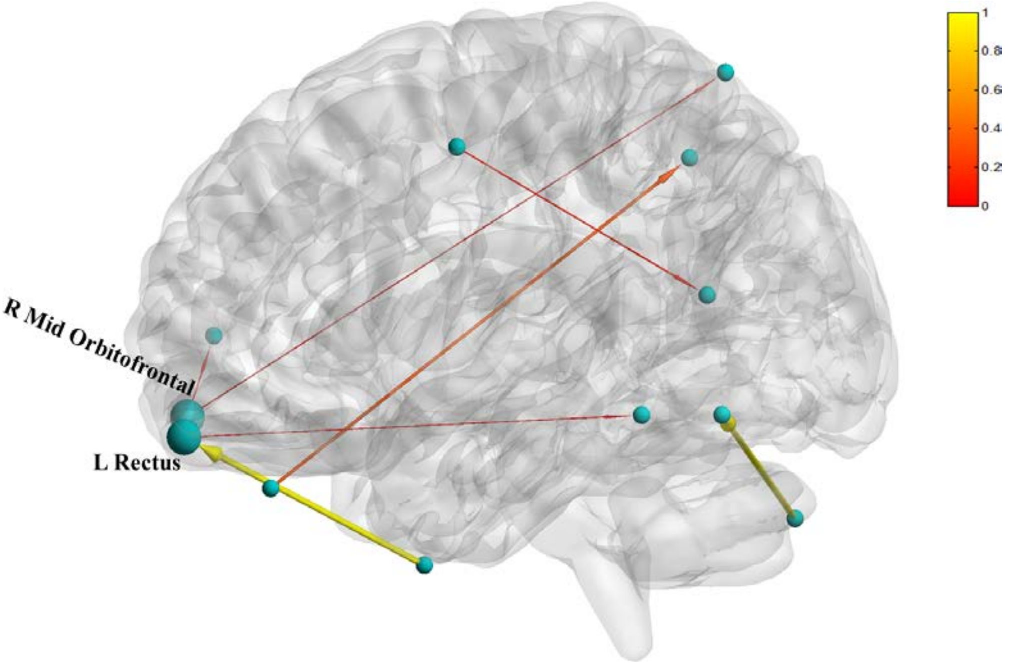

Fig.3.11.b

Top ranked connectivity paths which were greater in ADHD combined as compared to ADHD inattentive group (Fig.17a) and vice versa (Fig.17b). The width and the color of each path represent the rank of the path, with thinner and darker color path representing lower rank, bigger and lighter color path representing higher rank. The radius of the nodes is proportional to the number of inputs/outputs to/from them.

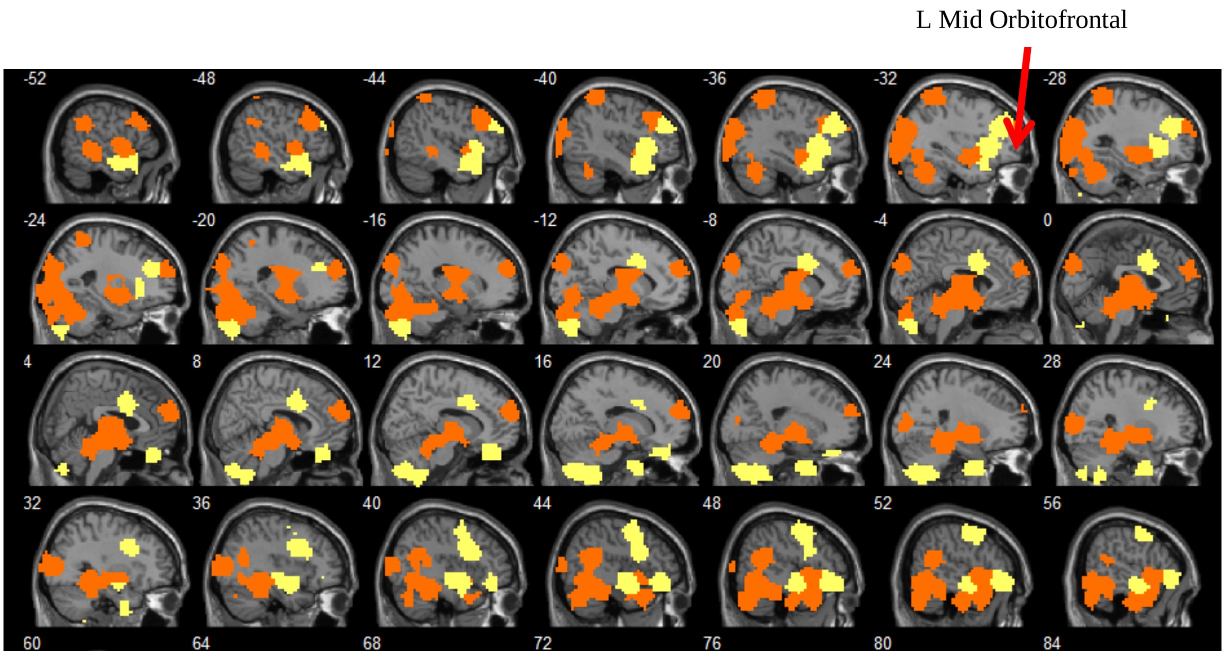

Fig.3.12 Regions which drove L mid orbitofrontal cortex (shown by red arrow) with greater influence in TDC as compared to ADHD combined (yellow) and with greater influence in ADHD combined as compared to TDC (orange).

II. Discussion

The main contributions of this paper are multifold. First, we have shown that the fully connected cascade deep neural network architecture outperforms support vector machines, which are popular in neuroimaging, for disease state classification using fMRI. In doing so, we have bested the top accuracy previously reported for the ADHD-200 global competition dataset [6]. Second, we have shown that connectivity-based features have higher discriminatory power as compared to raw data. Third, we have investigated the effects of principal component analysis of the features, as well as sample size and balance, on the performance of classifiers. Finally, the top ranked discriminative features inform us about the neural underpinnings, specifically altered brain connectivity, of ADHD. Below, we discuss each of these themes.

Introduced in 1998, SVMs are capable of handling data of high dimensionality and generally offer very good classification accuracy. Techniques of feature extraction and selection have been developed accordingly to optimize SVMs’ performance depending on the informative feedback on the significance of each feature to classification [68, 69]. Consequently, SVMs have been ubiquitously used in brain state and disease state classification using neuroimaging data [10, 70, 71]. On the other hand, NNs have been handicapped by issues relating to training, computational complexity and convergence of the traditional MLP architecture [21, 22, 23] . In this study, trained a fully connected cascade deep neural network architecture trained by the NBN algorithm and have proposed a two step-procedure for cross-validation maintaining complete separation of training and testing data. Further, we have demonstrated that the FCC NNs outperform SVMs irrespective of sample size, the type of feature or the classes being compared. This is a remarkable result which should pave the way for increased future usage of FCC NNs for brain state and disease state classification problems in neuroimaging.

Since ADHD is a spectrum disorder with heterogeneity inherent in its clinical definition, previous efforts to classify TDC from ADHD, as well as between ADHD sub-types, have not met with a lot of success. In response to this challenge, a global competition for classification of ADHD based on resting state fMRI data and phenotypic information (including age, gender, handedness, verbal and performance IQ) was announced [5]. Most researchers in the field were surprised by the poor overall accuracy (61%) of the winning fMRI-based classification approach. This was aggravated by the fact that phenotypic features performed better than fMRI (64%) [6, 72], undercutting the argument that directly measuring brain activity should give rise to more discriminative features. Subsequent careful re-analysis of the data by Fair *et al* [10] using a smaller sample size of 52 subjects in each class, functional connectivity patterns as features and a

SVM classifier, increased the accuracy to around 80%. We report around 90% accuracy in classifying between TDC & ADHD inattentive, and TDC & ADHD combined, and close to 95% accuracy in classifying the ADHD sub-types. This clearly demonstrates the utility of FCC deep architecture NNs.

Comparing the classification accuracy between different features, we found that connectivity-based features outperform raw image data. This is unsurprising given the fact that this has been demonstrated previously in other disorders [7]. It also makes sense specifically with respect to ADHD since it is a diffuse disorder with no focal point in the brain being solely involved in its pathology. Rather, different brain systems covering large parts of the brain have been implicated in ADHD [73, 74, 75, 76, 77]. However, we did hypothesize that directional connectivity features will perform better than non-directional ones and nonlinear connectivity features may show better accuracy as compared to linear ones. We did not find a strong evidence for this with respect to this data set. It is not possible to generalize this finding since other studies have shown that directional connectivity features perform better than non-directional connectivities [7].

Noise in the data can adversely affect the discriminatory power of features. Therefore, previous studies have used PCA as a means to separate the noise from the signal of interest and extract features such as principal components and latent variables which are likely to have more discriminatory power [78]. We did not find strong support for this notion from our data set. Though principal components gave much better accuracy than latent variables, and were comparable to raw connectivity path weights, they lack neuroscientific interpretation and hence are less preferable to connectivity path weights with comparable accuracy.

The relative sample sizes of classes can have a large bearing on the performance of a classifier. For example, suppose we have a three-class problem with 40%, 45% and 5% of the sample drawn from classes 1, 2 and 3, respectively. A classifier misclassifying all samples in class 3 may still give 90% overall classification accuracy. However this classifier is not desirable because it is unable to discriminate class 3. This phenomenon becomes problematic if class 3 is of primary interest. Such a scenario is not uncommon in disease state classification because more often than not, including the ADHD-200 data set, the size of the disease class is smaller than the healthy class because it is difficult to recruit and acquire data from a patient population. Our investigation of the performance of SVM and FCC NN for balanced and unbalanced data sets showed that the former was sensitive to it with higher accuracy for unbalanced data while the latter was relatively insensitive with high accuracy for both balanced and unbalanced data sets. However, it is noteworthy that weighted-SVMs can be employed to overcome SVM's sensitivity to unbalanced data [79].

The top ranked features obtained from connectivity path weights highlight dysfunction of causal pathways associated with frontal and cerebellar regions. Specifically, there appeared to be a large reduction of the causal influence of many cerebellar regions from other cortical areas in ADHD inattentive as compared to TDC. There were more limited reductions of the input to the vermis region of the cerebellum in ADHD combined as compared to TDC. These reductions in the causal input to the cerebellum are consistent with previously observed structural deficits in cerebellar white matter pathways [80, 81] as well as focal and distributed functional abnormalities involving the cerebellum [80, 82, 83, 84]. There were specific increases in the causal input to left mid orbitofrontal cortex in TDC as compared to ADHD combined while certain other regions had a higher causal influence on the same region in ADHD combined as compared to TDC. This shows that dysregulation of frontal circuitry in ADHD is not unidimensional. Rather it is more complex, with re-organization of regions driving this region in ADHD as compared to controls. These findings are consistent with previously reported structural [80, 81, 85] and functional [80, 82, 83, 86] alterations of frontal circuitry in ADHD. It is particularly noteworthy that previous studies found more alterations with left, rather than right, frontal cortex [86]. This is corroborated by our results. Comparison of the ADHD subtypes also show specific differences in frontal and cerebellar connectivity. This shows that there may be a neurological basis for the sub-types based on which regions drive frontal and cerebellar regions and by how much.

Chapter 4 Conclusion

In this study, I have demonstrated the utility of fully connected cascade deep architecture neural network for classifying subjects with attention deficit hyperactivity disorder from typically developing subjects. Given that support vector machines are predominantly used as classifiers in neuroimaging, my major contribution is to introduce FCC NN to the neuroimaging community. My second major contribution is to dispel the pessimism about neuroimaging based ADHD classification borne out of the final results of the ADHD-200 competition. I have shown that with improved classifier design and discriminative connectivity-based features, the classification accuracy can be greatly improved.

BIBLIOGRAPHY

[1] S. Ogawa, D. Tank, R. Menon, J. Ellermann, S. Kim, H. Merkle and K. Ugurbil, "Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging," *Proceedings of the National Academy of Sciences USA,* vol. 89, no. 13, pp. 5951-5955, 1992.

[2] S. Ogawa, T. Lee, A. Kay and D. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences USA,* vol. 87, no. 24, pp. 9868-9872, 1990.

[3] S. Kim and K. Ugurbil, "Functional magnetic resonance imaging of the human brain," *Journal of Neuroscience Methods,* vol. 74, no. 2, pp. 229-243, 1997.

[4] P. Fox and M. Raichle, "Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects," *Proceedings of the National Academy of Sciences USA,* vol. 83, no. 4, pp. 1140-1144, 1986.

[5] "The ADHD-200 Global Competition SAMPLE Home," [Online]. Available: http://fcon_1000.projects.nitrc.org/indi/adhd200.

[6] "The ADHD-200 Global Competition Result," [Online]. Available: http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html.

[7] G. Deshpande, Z. Li, P. Santhanam, C. Coles, M. Lynch, S. Hamann and H. X., "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.

[8] J. Zhang, W. Cheng, G. Wang, Z. Zhang, W. Lu, G. Lu and J. Feng, "Pattern Classification of Large-Scale Functional Brain Networks: Identification of Informative Neuroimaging Markers for Epilepsy," *PLoS ONE,* vol. 7, no. 5, p. e36733, 2012.

[9] J. Bohland, S. Saperstein, F. Pereira, J. Rapin and L. Grady, "Network, anatomical, and non-imaging measures for the prediction of ADHD diagnosis in individual subjects," *Frontiers in Systems Neuroscience,* vol. 6, no. 78, pp. 1-28, 2012.

[10] D. Fair, J. Nigg, S. Iyer, D. Bathula, K. Mills, N. Dosenbach, B. Schlaggard, M. Mennes, D. Gutman, S. Bangaru, J. Buitelaar, D. Dickstein, A. Martino, D. Kennedy, C. Kelly, B. Luna, J. Schweitzer, K. Velanova, Y. Wang, S. Mostofsky, F. Castellanos and M. 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, no. 80, pp. 1-31, 2013.

[11] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys,* vol. 34, no. 1, pp. 1-47, 2002.

[12] J. Dong, A. Krzyzak and S. CY, "An improved handwritten Chinese character recognition system using support vector machine," *Pattern Recognition Letters,* vol. 26, no. 12, pp. 1849-1856, 2005.

[13] S. Baluja and H. Rowley, "Boosting Sex Identification Performance," *International Journal of Computer Vision,* vol. 71, no. 1, pp. 111-119, 2007.

[14] Z. Lei and Y. Dai, "An SVM-based system for predicting protein subnuclear localizations," *BMC Bioinformatics,* vol. 6, no. 291, p. 291, 2005.

[15] S. Osowski, K. Siwek and T. Markiewicz, "MLP and SVM Networks – a Comparative Study," in *6th Nordic Signal Processing Symposium*, Espoo, Finland, 2004.

[16] M. Lee and C. To, "Comparison of Support Vectore Machine and Back Propagation Neural network in Evaluating the Enterprise Financial Distress," *International Journal of Artificial Intelligence & Applications,* vol. 1, no. 3, pp. 31-43, 2010.

[17] E. Monte-Moreno, "Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine," *Artificial Intelligence in Medicine,* vol. 53, no. 2, pp. 127-138, 2011.

[18] O. Sadik, W. Land, A. Wanekaya, M. Uematsu, M. Embrechts, L. Wong, D. Leibensperger and A. Volykin, "Detection and Classification of Organophosphate Nerve Agent Simulants Using Support Vectore machines with," *Journal of Chemical Information and Modeling,* vol. 44, no. 2, pp. 499-507, 2004.

[19] Chang,RF; Wu, WJ; Moon, WK; Chou,YH; Chen,DR;, "Support vector machines for diagnosis of breast tumors on US images," *Academic Radiology,* vol. 10, no. 2, pp. 189-197, 2003.

[20] A. J. a. N. Bogunovic, "Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features," *Artificial Intelligence in Medicine,* vol. 51, no. 3, pp. 175-186, 2011.

[21] B. Wilamowski, "Efficient neural network architectures and advanced training algorithms," *Gdańsk University of Technology Faculty of ETI Annals,* vol. 18, pp. 345-352, 2010.

[22] B. Wilamowski, "Can computers be more intelligent than humans?," in *4th International Conference on IEEE,* Yokohama, Japan, 2011.

[23] B. Wilamowski, "Neural Network Learning without Error Back Propagation," *IEEE Transactions on Neural Networks,* vol. 21, no. 11, pp. 1793-1803, 2010.

[24] B. Biswal, M. Mennes, X. Zuo, S. Gohel, C. Kelly, S. Smith, C. Beckmann, J. Adelstein, R. Buckner, S. Colcombe, A. Dogonowski, M. Ernst, D. Fair, M. Hampson, M. Hoptman, J. Hyde, V. Kiviniemi, R. Kotter, S. Li, C. Lin, M. Lowe, C. Mackay, D. Madden, K. Madsen, D. Margulies, H. Mayberg, K. McMahon, C. Monk, S. Mostofsky, B. Nagel, J. Pekar, S. Peltier, S. Petersen, V. Riedl, S. Rombouts, B. Rypma, B. Schlaggar, S. Schmidt, R. Seidler, G. Siegle, C. Sorg, G. Teng, J. Veijola, A. Villringer, M. Walter, L. Wang, X. Weng, S. Whitfield-Gabrieli, P. Williomson, C. Windischberger, Y. Zang, H. Zhang, F. Castellanos and M. Milham, "Toward Discovery Science of Human Brain Function," *Proceedings of the National Academy of Sciences USA,* vol. 107, no. 10, pp. 4734-4739, 2010.

[25] D. Williams, V. Cherkassky, R. Mason, A. Keller, N. Minshew and J. MA, "Brain Function Differences in Language Processing in Children and Adults with Autism," *Autism Research,* in press, 2013.

[26] R. Marsh, G. Horga, N. Parashar, Z. Wang, B. Peterson and H. Simpson, "Altered Activation in Fronto-Striatal Circuits During Sequential Processing of Conflict in Unmedicated Adults with Obsessive-Compulsive Disorder," *Biological Psychiatry,* in press, 2013.

[27] I. Strigo, S. Matthews and A. Simmons, "Decreased frontal regulation during pain anticipation in unmedicated subjects with major depressive disorder," *Translational Psychiatry,* vol. 3, no. 3, p. e239, 2013.

[28] M. Makuuch and A. Friederici, "Hierarchical functional connectivity between the core language system and the working memory system," *Cortex,* in press, 2013.

[29] S. Li, B. Biswal, Z. Li, R. Risinger, C. Rainey, J. Cho, B. Salmeron and E. Stein, "Cocaine administration decreases functional connectivity in human primary visual and motor cortex as detected by functional MRI," *Magnetic Resonance in Medicine,* vol. 43, no. 1, pp. 45-51, 2000.

[30] M. Lowe, M. Phillips, J. Lurito, D. Mattson, M. Dzemidzic and V. Mathews, "Multiple sclerosis: Low-frequency temporal blood oxygen level-dependent fluctuations indicate reduced functional connectivity - Initial results," *Radiology,* vol. 224, no. 1, pp. 184-192, 2002.

[31] S. Li, Z. Li, M. Zhang, M. Franczak and P. Antuono, "Alzheimer disease: Evaluation of a Functional MR Imaging Index as a Marker," *Radiology,* vol. 225, no. 1, pp. 253-259, 2002.

[32] Q. Wu, D. Li, W. Kuang, T. Zhang, S. Lui, X. Huang, R. Chan, G. Kemp and Q. Gong, "Abnormal regional spontaneous neural activity in treatment-refractory depression revealed by resting-state fMRI," *Human Brain Mapping,* vol. 32, no. 8, pp. 1290-1299, 2011.

[33] D. Shukla, B. Keehn and R. Muller, "Regional homogeneity of fMRI time series in autism spectrum disorders," *Neuroscience Letters,* vol. 476, no. 1, pp. 46-51, 2010.

[34] T. Wu, X. Long, Y. Zang, L. Wang, M. Hallett, K. Li and P. Chan, "Regional homogeneity changes in patients with Parkinson's disease," *Human Brain Mapping,* vol. 30, no. 5, pp. 1502-1510, 2009.

[35] X. Li, A. Sroubek, M. Kelly, I. Lesser, E. Sussman, Y. He, C. Branch and J. Foxe, "Atypical pulvinar-cortical pathways during sustained attention performance in children with attention-deficit/hyperactivity," *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 51, no. 11, pp. 1197-1207, 2012.

[36] G. Deshpande and X. Hu, "Investigating effective brain connectivity from FMRI data: past findings and current issues with reference to granger causality analysis," *Brain Connectivity,* vol. 2, no. 5, pp. 235-245, 2012.

[37] G. Deshpande, K. Sathian, X. Hu and K. Buckhalt, "A rigorous approach for testing the constructionist hypotheses of brain function," *Behavioral and Brain Sciences,* vol. 35, no. 3, pp. 148-149, 2012.

[38] G. Deshpande, P. Santhanam and X. Hu, "Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data," *NeuroImage,* vol. 54, no. 2, pp. 1043-1052, 2011.

[39] G. Deshpande, K. Sathian and X. Hu, "Assessing and Compensating for Zero-Lag Correlation Effects in Time-lagged Granger Causality Analysis of fMRI," *IEEE Transactions on Biomedical Engineering,* vol. 57, no. 6,

pp. 1446-1456, 2010.

[40] G. Deshpande, X. Hu, S. Lacey, R. Stilla and K. Sathian, "Object familiarity modulates effective connectivity during haptic shape perception," *NeuroImage,* vol. 49, no. 3, pp. 1991-2000, 2010.

[41] G. Deshpande, S. LaConte, G. James, S. Peltier and X. Hu, "Multivariate Granger causality analysis of fMRI data," *Human Brain Mapping,* vol. 30, no. 4, pp. 1361-1373, 2009.

[42] G. Deshpande, X. Hu, R. Stilla and K. Sathian, "Effective Connectivity during Haptic Perception: A study using Granger causality analysis of functional magnetic resonance imaging data," *NeuroImage,* vol. 40, no. 4, pp. 1807-1814, 2008.

[43] R. Goebel, A. Roebroeck, D. Kim and E. Formisano, "Invetigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping," *Magnetic Resonance Imaging,* vol. 21, no. 10, pp. 1251-1261, 2003.

[44] Q. Gao, H. Chen and Q. Gong, "Evaluation of the effective connectivity of the dominant primary motor cortex during bimanual movement using Granger causality," *Neuroscience Letters,* vol. 443, no. 1, pp. 1-6, 2008.

[45] D. Marinazzo, M. Pellicoro and S. Stramaglia, "Kernel-Granger causality and the analysis of dynamical networks," *Physical review E,* vol. 77, no. 5, p. 056215, 2008.

[46] D. Marinazzo, M. Pellicoro and S. Stramaqlia, "Kernel method for nonlinear granger causality," *Physical Review Letters,* vol. 100, no. 14, p. 144103, 2008.

[47] D. Rangaprakash, X. Hu and G. Deshpande, "Phase synchronization in brain networks derived from correlation between probabilities of recurrences in functional MRI data," *Internation Journal of Neural Systems,* in press, 2013.

[48] J. Richiardi, M. Gschwind, S. Simioni, J. Annoni, B. Greco, P. Hagmann, M. Schluep, P. Vuilleumier and D. Ville, "Classifying minimally disabled multiple sclerosis patients from resting state functional connectivity," *NeuroImage,* vol. 62, no. 3, pp. 2021-2033, 2012.

[49] L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li and D. Hu, "Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis," *Brain,* vol. 135, no. 5, pp. 1498-1507, 2012.

[50] M. Liu, L. Zeng, H. Shen, Z. Liu and D. Hu, "Potential risk for healthy siblings to develop schizophrenia: evidence from pattern classification with whole-brain," *NeuroReport,* vol. 23, no. 5, pp. 265-269, 2012.

[51] C. Wee, P. Yap, D. Zhang, K. Denny, J. Browndyke, G. Potter, K. 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.

[52] G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment,* vol. 37, no. 1, pp. 35-46, 1991.

[53] H. Kalayeh and D. Landgrebe, "Predicting the required number of training samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 5, no. 6, pp. 664-667, 1983.

[54] B. Brumen, M. Juric, T. Welzer, I. Rozman, H. Jaakkola and A. Papadopoulos, "Assessment of Classification Models with Small Amounts of Data," *Informatica,* vol. 18, no. 3, pp. 343-362, 2007.

[55] "ADHD-200 Global Competition database," [Online]. Available: http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline.

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

[57] Wei Liao, Daniele Marinazzo, Zhengyong Pan, Qiyong Gong, and Huafu Chen, "Kernel Granger Causality Mapping Effective Connectivity on fMRI Data," *IEEE TRANSACTIONS ON MEDICAL IMAGING,* vol. 28, no. 11, 2009.

[58] A. McQuarrie and C. Tsai, Regression and time series model selection, vol. 43, Singapor: World Scientific, 1998.

[59] Y. Yamamoto and P. Nikiforuk, "A New Supervised Learning Algorithm forMultilayered and Interconnected Neural Networks," *IEEE Transactions on Neural Networks,* vol. 11, no. 1, pp. 36-46, 2000.

[60] B. Wilamowski, S. Iplikci and M. Efe, "An algorithm for fast convergence in training neural networks," in *International Joint Conference on Neural Networks,* Boise, ID, USA, 2001.

[61] V. Singh, I. Gupta and H. Gupta, "ANN-based estimator for distillation using Levenberg–Marquardt approach," *Engineering Applications of Artificial Intelligence,* vol. 20, no. 2, pp. 249-259, 2007.

[62] P. Corral, O. Ludwig and A. Lima, "Time-varying channel neural equalisation using Gauss-Newton algorithm," *Electronics Letters,* vol. 46, no. 15, pp. 1055-1056, 2012.

[63] Ö. KIsI, "Multi-layer perceptrons with Levenberg-Marquardt training algorithm for suspended sediment concentration prediction and estimation," *Hydrological Sciences Journal,* vol. 49, no. 6, p. 1040, 2004.

[64] S. Basterrech, S. Mohammed, G. Rubino and M. Soliman, "Levenberg-Marquardt Training Algorithms for Random Neural networks," *The Computer Journal,* vol. 54, no. 1, pp. 125-135, 2011.

[65] B. Wilamowski, "Improved Computation for Levenberg-Marquardt Training," *IEEE Transactions on Neural Networks,* vol. 21, no. 6, pp. 930-937, 2010.

[66] "NBN software," [Online]. Available: http://www.eng.auburn.edu/~wilambm/nnt/index.htm.

[67] "MATLAB SVM toolbox-Spider," [Online]. Available: http://people.kyb.tuebingen.mpg.de/spider/main.html.

[68] T. Li, C. Zhang and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *International Society for Biocuration,* vol. 20, no. 15, pp. 2429-2437, 2004.

[69] X. Zhang, X. Lu, Q. Shi, X. Xu, H. Leung, L. Harris, J. Iglehart, A. Miron, J. Liu and W. Wong, "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC Bioinformatics,* vol. 7, no. 1, p. 197, 2006.

[70] A. Mueller, G. Candrian, V. Grane, J. Kropotov, V. Ponomarev and G. Baschera, "Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: a validation study," *Nonlinear Biomedical Physics,* vol. 5, no. 5, 2011.

[71] A. Mueller, G. Candrian, V. Grane, J. Kropotov, V. Ponomarev and G. Baschera, "Classification of ADHD patients on the basis of independent ERP components using a machine earning system," *Nonlinear Biomedical Physics,* vol. 4, no. 1, 2010.

[72] M. Brown, G. Sidhu, R. Greiner, N. Asgarian, M. Bastani, P. Silverstone, A. Greenshaw and S. Dursun, "ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements," *Frontiers in Systems Neuroscience,* vol. 6, no. 69, pp. 1-22, 2012.

[73] D. Fair, J. Posner, B. Naqel, D. Bathula, T. Dias, K. Mills, M. Blythe, A. Giwa, C. Schmitt and J. Nigg, "Atypical default network connectivity in youth with attention-deficit/hyperactivity disorder," *Biological Psychiatry,* vol. 68, no. 12, pp. 1084-1091, 2010.

[74] T. Costa Dias, V. Wilson, D. Bathula, S. Iyer, B. Thurlow, C. Stevens, E. Musser, S. Carpenter, D. Grayson, S. Mitchell, J. Nigg and D. Fair, "Reward circuit connectivity relates to delay discounting in children with attention-deficit/hyperactivity disorder," *European Neuropsychopharmacology,* vol. 23, no. 1, pp. 33-45, 2013.

[75] J. Epstein, M. Delbello, C. Adler, M. Altaye, N. Mills, S. Strakowski and S. Holland, "Differential patterns of brain activation over time in adolescents with and without attention deficit hyperactivity disorder (ADHD) during performance of a sustained attention task," *Neuropediatrics,* vol. 40, no. 1, pp. 1-5, 2009.

[76] F. Castellanos, D. Margulies, A. Kelly, L. Uddin, M. Ghaffari, A. Kirsch, D. Shaw, Z. Shehzad, A. Di Martino, B. Biswal, E. Sonuga-Barke, J. Rotrosen, L. Adler and M. Milham, "Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder," *Biological Psychiatry,* vol. 63, no. 3, pp. 332-337, 2008.

[77] K. Mills, D. Bathula, T. Dias, S. Iyer, M. Fenesy, E. Musser, C. Stevens, B. Thurlow, S. Carpenter, B. Naqel, J. Nigg and F. DA, "Altered cortico-striatal-thalamic connectivity in relation to spatial working memory capacity in children with ADHD," *Frontiers in Psychiatry,* vol. 3, no. 2, 2012.

[78] G. Sidhu, N. Asgarian, R. Greiner and M. Brown, "Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD," *Frontier Systems Neuroscience,* vol. 6, no. 74, 2012.

[79] A. Anand, G. Pugalenthi, G. Fogel, and P.N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino Acids,* vol. 39, pp. 1385-1391, 2010.

[80] P. Curatolo, E. D'Agati and R. Moavero, "The neurobiological basis of ADHD," *Italian Journal of Pediatrics,* vol. 36, no. 1, p. 79, 2010.

[81] H. Van Ewijk, D. Heslenfeld, M. Zwiers, J. Buitelaar and J. Oosterlaan, "Diffusion tensor imaging in attention

deficit/hyperactivity disorder: a systematic review and meta-analysis," *Neuroscience & Biobehavioral Reviews,* vol. 36, no. 4, pp. 1093-1106, 2012.

[82] S. Durston, J. Van Belle and P. De Zeeuw, "Differentiating frontostriatal and fronto-cerebellar circuits in attention-deficit/hyperactivity disorder," *Biological Psychiatry,* vol. 69, no. 12, pp. 1178-1184, 2011.

[83] A. Cubillo, R. Halari, A. Smith, E. Taylor and K. Rubia, "A review of fronto-striatal and fronto-cortical brain abnormalities in children and adults with Attention Deficit Hyperactivity Disorder (ADHD) and new evidence for dysfunction in adults with ADHD during motivation and attention," *Cortex,* vol. 48, no. 2, pp. 194-215, 2012.

[84] M. Cherkasova and L. Hechtman, "Neuroimaging in attention-deficit hyperactivity disorder: beyond the frontostriatal circuitry," *Canadian Journal of Psychiatry,* vol. 54, no. 10, pp. 651-664, 2009.

[85] A. Konrad, T. Dielentheis, D. El Masri, M. Bayerl, C. Fehr, T. Gesierich, G. Vucurevic, P. Stoeter and G. Winterer, "Disturbed structural connectivity is related to inattention and impulsivity in adult attention deficit hyperactivity disorder," *European Journal of Neuroscience* , vol. 31, no. 5, pp. 912-919, 2010.

[86] G. Bush, E. Valera and L. Seidman, "Functional neuroimaging of attention-deficit/hyperactivity disorder: a review and suggested future directions," *Biological Psychiatry,* vol. 57, no. 11, pp. 1273-1284, 2005.