Brain Informatics (2016) 3:145–155
DOI 10.1007/s40708-016-0047-1

An integrated feature ranking and selection framework for ADHD characterization

Cao Xiao · Jesse Bledsoe · Shouyi Wang · Wanpracha Art Chaovalitwongse · Sonya Mehta · Margaret Semrud-Clikeman · Thomas Grabowski

Received: 27 November 2015 / Accepted: 14 March 2016 / Published online: 2 April 2016

Abstract Today, diagnosis of attention deficit hyperactivity disorder (ADHD) still primarily relies on a series of subjective evaluations that highly rely on a doctor's experiences and intuitions from diagnostic interviews and observed behavior measures. An accurate and objective diagnosis of ADHD is still a challenge and leaves much to be desired. Many children and adults are inappropriately labeled with ADHD conditions, whereas many are left undiagnosed and untreated. Recent advances in neuroimaging studies have enabled us to search for both structural (e.g., cortical thickness, brain volume) and functional (functional connectivity) abnormalities that can potentially be used as new biomarkers of ADHD. However, structural and functional characteristics of neuroimaging data, especially magnetic resonance imaging (MRI), usually generate a large number of features. With a limited sample size, traditional machine learning techniques can be problematic to discover the true characteristic features of ADHD due to the significant issues of overfitting, computational burden, and interpretability of the model. There is an urgent need of efficient approaches to identify meaningful discriminative variables from a higher dimensional feature space when sample size is small compared with the number of features. To tackle this problem, this paper proposes a novel integrated feature ranking and selection framework that utilizes normalized brain cortical thickness features extracted from MRI data to discriminate ADHD subjects against healthy controls. The proposed framework combines information theoretic criteria and the least absolute shrinkage and selection operator (Lasso) method into a two-step feature selection process which is capable of selecting a sparse model while preserving the most informative features. The experimental results showed that the proposed framework generated the highest/comparable ADHD prediction accuracy compared with the state-of-the-art feature selection approaches with minimum number of features in the final model. The selected regions of interest in our model were consistent with recent brain–behavior studies of ADHD development, and thus confirmed the validity of the selected features by the proposed approach.

1 Introduction

Attention deficit hyperactivity disorder (ADHD) is among the most common child and adult neurodevelopmental disorder. ADHD symptoms include inattention, hyperactivity, and impulsivity. It affects approximately 5–10 % of all school-age children and nearly 5 % of adults on their motor, cognitive, and emotional development [1]. Diagnosis of ADHD still remains a challenge, requiring long-term and extended involvement from clinicians, parents, and teachers. Clinicians rely heavily on experiences and intuitions when conducting diagnostic interview and observational measures. A delay or incorrect diagnosis of ADHD could have a significant negative impact on a patient's social and emotional development, while an early and accurate detection of ADHD can strongly influence the

C. Xiao (✉) · J. Bledsoe · W. A. Chaovalitwongse ·
S. Mehta · T. Grabowski
University of Washington, Seattle, WA, USA
e-mail: danicaxiao@gmail.com

S. Wang
University of Texas, Arlington, Arlington, USA

M. Semrud-Clikeman
University of Minnesota, Minneapolis, USA

course of the condition development by delivery of appropriate treatments to the patient. In addition to the traditional clinical diagnosis, there is a pressing need to find a set of more discriminative and objective features to characterize ADHD that can be used to facilitate ADHD diagnosis.

Previous studies on the etiology of ADHD are mostly based on structural or functional neuroimaging research of group level (ADHD vs. control) differences. Some informative features extracted are blood oxygenation level-dependent (BOLD) signals from functional magnetic resonance imaging (fMRI) data [2], wavelet synchronization likelihoods extracted from electroencephalography (EEG) data [3], rolandic spikes from EEG data [4], brain volume measure extracted from magnetic resonance imaging (MRI) data [5]. The pursuit of neuroanatomical biomarkers has a great potential to facilitate new discriminative methods that are etiologically informed and validated by neuropsychological theories. However, due to high cost of neuroimaging data acquisition, most current ADHD studies are based on relatively small sample sizes, which reduce the statistical power needed to validate meaningful discriminative variable from a very large number of features extracted from structural MRI [6]. A limited sample size with equivalent number of features raises new challenges to traditional machine learning algorithms, such as logistic regression or support vector machines (SVM), as they tend to overfit and lack a generalization power when training on a dataset containing the number of features far larger than the sample size ($p \gg n$ problem). In previous work, some models either use hundreds of features as an input or exhaustively search on a preselected smaller subset of features. SVM is mostly favored [7] and some variant of feedforward neural networks [8] is also used. We believe that those methods are either susceptible to overfitting or too restrictive in the search space. The interpretation of the final models is very difficult to validate by existing neuropsychological theories.

In this study, we propose an integrated feature ranking and selection framework that uses brain cortical thickness, extracted from structural MRI data, as features and constructs a prediction model to identify ADHD subjects versus normal controls. The framework performs a two-step feature selection process based on both information theoretic criteria and regularization concept. To mitigate the inconsistent feature selection issue of regularization, especially the lasso method [9], the framework preanalyzes all features to rank informative features based on mutual information scores [10]. In feature selection, it extends the lasso method [11] to construct a prediction model by fixing those preselected highly informative features when performing regression. Tested on both simulated and real datasets, our framework is shown to effectively preserve highly informative features identified in the feature ranking step and improve the model accuracy while searching in a full-feature space and maintaining the sparsity in the feature selection step. With a prediction accuracy of 80.9 %, our framework selects two sparse models, each with only 4 or 5 cortical thickness features. Previous neurodevelopmental studies of ADHD also consistently suggest that the features selected in our models have a deeper connection to the neurodevelopmental basis of ADHD, and thus making the models highly interpretable to clinicians. The proposed feature selection and prediction framework is a necessary first step to help clinicians find more features of characterizing ADHD using an objective measure with high discriminative accuracy.

The rest of the paper is organized as follows. In Sect. 2, we introduce the background of ADHD, including the brain cortical thickness and its connection to ADHD. We also review the current feature ranking and selection algorithms. Section 3 presents the proposed two-step feature ranking and selection framework, including the model formulations and model validation using simulated datasets. Section 4 shows the experimental results of the proposed framework on ADHD characterization using a real MRI neuroimaging dataset. Finally, we conclude the study in Sect. 5.

2 Background

2.1 Feature extraction of ADHD

ADHD is considered a neurodevelopmental disorder given the age-related differences in cortical maturation that characterize ADHD. Researchers suggest that the origins of attention can be observed in infants as young as three months when the young infant is able to selectively attend (i.e., recognize and orient toward) to their caregiver's face [12]. According to these researchers, attention is composed of differential structures and circuits, called an organ system. Furthermore, as a child matures during preschool and early elementary school years, attention response grows into the ability to self-regulate (i.e., adjust one's emotional state/behavior depending on the demands of the environment) in a changing and dynamic environment. Those higher level attention abilities are often described with the term "executive functions." Such development not only relies on social demand, but also is due to the brain maturation of the prefrontal cortex. In Posner and Fan's (2008) model, self-regulation leads to the second stage in attention development, the executive network. During the ages of 5–9, children with deficits in self-regulation and attention are noticed by teachers and parents, as their behaviors deviate from what would be developmentally appropriate.

Choosing brain cortical thickness as the features in ADHD characterization is not only supported by theory, but also benefits from advances of neuroimaging techniques. Numerous theories have hypothesized the cause of ADHD [13–18]. Those hypothesis are further supported by neuroimaging research, which provides an accurate way to measure the relationship between behaviors or symptoms and underlying brain morphology and brain functioning. As structural and functional neuroimaging techniques have improved vastly over the last thirty years, MRI provides excellent spatial resolution, uses no ionizing radiation (unlike computed tomography, CT), and thus can be used in pediatric samples of clinical and non-clinical typically developing controls. Cortical and surfaced-based neuroimaging techniques improve on conventional volumetric analysis by allowing for a direct measure of cortical thickness in millimeters, thus may present a more sensitive tool for understanding and measuring brain abnormalities in children with ADHD. So far, a large number of neuroimaging studies have observed that ADHD manifests via a general deficit in the dopaminergic system of the brain including prefrontal cortex [13, 5] or abnormalities in brain structures rich in dopamine receptors in children and adults with ADHD [19–23, 5].

2.2 Feature selection

Although recent advances in neuroimaging studies have enabled us to search for structural brain abnormalities caused by the disease that can potentially be used as new biomarkers of ADHD, characterization using traditional machine learning techniques can be difficult because structural characteristics of neuroimaging data, especially MRI data, usually result in large number of features. Even grouping raw features into region of interests (ROI), finding discriminative features for ADHD is still not easy due to relative small sample size with a limited number of patients and healthy participants. Learning from limited sample size with equivalent feature size raises significant issues of overfitting and interpretability of the final model. This study is motivated by the challenge and is aimed to develop efficient feature selection approaches that can construct a sparse model with the most clinical meaningful features preserved. In particular, this paper proposes a novel integrated feature ranking and selection framework which combines information theoretic criteria and the least absolute shrinkage and selection operator (Lasso) method into a two-step feature selection process. The current information theory-based and the Lasso-based feature selection approaches will be discussed in the following.

2.2.1 Feature selection using mutual information

Mutual information [24, 10] is a measure of the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence. MI measures how much information a feature contains about the class without making any assumptions about the nature of their underlying relationships. It is formulated as

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right).$$

If the feature is a perfect indicator for the class membership, its MI reaches its maximum value. A basic intuition is that a stronger mutual information implies a greater predictive ability when using the feature. As an information theoretic criteria, MI have been applied in many feature selection problems [25]. To know whether a given candidate feature should be included, one must be able to evaluate the joint mutual information $I(X, Y)$. However, as feature matrix X is generally multi-dimensional with a continuous distribution, the joint mutual information $I(X, Y)$ is thus extremely difficult to reliably estimate. To solve the problem, one can assume each feature is independent of all other features, and rank the features in descending order according to their individual mutual information score $I(X_i, Y)$. The feature selection is simply picking the top K features, where K can be determined by a predefined certain number of features or some stopping criterion. The feature selection criterion based on mutual information score is commonly adopted in literature. It is often referred as Mutual Information Maximization (MIM) approaches [26]. However, the performance of this approach is known to be suboptimal if features are interdependent, which is a general case in most studies. In addition, it is widely accepted that a useful set of features should not only be individually relevant to class label, but also should not be redundant with respect to each other, that is features should not be highly correlated in the selected subset. To consider both relevancy and redundancy, a number of approaches have been proposed. For example, Battiti [27] proposed the Mutual Information Feature Selection (MIFS) criterion, which introduces an inter-feature correlation term into the MIM criterion. A penalty parameter β is employed to control the tradeoff between relevancy and redundancy. If the penalty parameter β is set to 0, it is equivalent to the MIM criterion. Peng et. al. [28] presented the Maximum-Relevance-Minimum-Redundancy (MRMR) criterion, which is in principle equivalent to MIFS with the $\beta = 1/(n-1)$, where n is the number of selected features in the current subset. Yang and Moody [29] used Joint Mutual Information (JMI) to focus on increasing complementary information between features. In particular, the mutual information between the class label and a joint random variable X_kX_j is calculated. By pairing a candidate X_k with each previously selected feature. The principle idea is that if the candidate feature is 'complementary' with the existing features, it should be included in the feature subset. Fleuret proposed the Conditional Mutual Information Maximization (CMIM) criterion [30], which examines the information between a feature and the class label, conditioned on each current feature. Instead of taking the mean of the redundancy term, CMIM takes the maximum value in the redundancy term and thus penalize more on feature redundancy.

Although mutual information-based feature selection approaches gained wide popularity in the literature, there are still some significant issues unsolved. First, all these criteria rely on highly restrictive assumptions on the underlying data distributions. In particular, due to the computational difficulties in high-dimensional mutual information estimation, most approaches only consider pairwise and conditional pairwise interactions, and omit the higher-order interactions. Second, most current MI-based approaches perform feature selection sequentially starting from high-ranked features. As a result, by excluding low MI ranking features, such approaches deny the possibility that a set of low-ranked features combined together may generate strong predictive power (e.g., in the famous XOR problem [31]). We have the risk of missing that strong signal by only working on the preselected candidate set [32, 33, 28].

2.2.2 *Feature selection with regularization*

In medical research, due to high cost of data acquisition, researchers often run into the issue of insufficient samples to train and validate developed models. Instead of heuristic selection schemes (such as many MI-based approaches), objective optimization methods have received more attention since they can be conveniently formulated as convex optimization problems with global optimal solutions. A typical objective function consists of an error term and a regularization term. One of the most widely used such feature selection algorithms is the least absolute shrinkage and selection operator (Lasso) [34]), which allows computationally efficient feature selection based on linear dependencies between input features and output values. The Lasso method as a shrinkage and selection method for linear regression gradually receives high recognition and a fast coordinate descent algorithm has been devised to solve the optimization problem. The optimization framework of lasso to minimize the sum of squared errors with a l_1-norm penalty (bound on the sum of the absolute values of the coefficients) is formulated as follows:

$$\sum_{i=1}^{n}(y_i - \beta x_i)^2 + \lambda||\beta||_1.$$

By penalizing and forcing some variables to be zero, lasso can effectively select a sparse model. However, it sacrifices unbiasedness to reduce the variance of the predicted value [35].

There are still some challenges for application of Lasso method in feature selection. The Lasso result is often subject to the scaling of features. Inappropriate scaling may cause imbalanced penalty on linear coefficients. The true underlining features with high coefficients may be suppressed to have smaller coefficients. As a result, the total explained variance is limited. Instead of rescaling all features, more generally one can employ adaptive Lasso [36] with penalty term $\lambda \sum w_i||\beta_i||_1$. Even so, effects of strong signal will be diminished due to shrinkage.

3 New integrated feature ranking and selection model

3.1 Model formulation and solution

The proposed integrated feature ranking and selection framework is performed in two stages: mutual information-based feature ranking and Lasso-based feature selection. In the feature ranking step, all features are ranked by their MI scores, and a subset of high-ranked features are selected and considered to have the best informative power. Among those features, a redundancy removal step is performed by checking pairwise correlation between the features. For a highly correlated feature pair (higher than a threshold), the feature with lower MI score is considered redundant and removed from the feature subset to prevent multicollinearity. In the feature selection step, we set the best informative features penalty-free in the generalized lasso method. We use Lasso to select additional features from the full-feature space, not restricted to the subset of high MI features. The additional features selected, although have lower MI scores individually, can improve model classification performance when combined together. Within the subset of high MI features, we start with setting the single top-ranked feature penalty-free, then all combinations of two top features, then all combinations of three top features, iteratively. The feature selection and classification model was validated by leave-one-out cross-validation (LOOCV). The search process stops when validation accuracy cannot be further improved. The resulting model will be the best model for class prediction. Comparing with other MI-ranking-based methods, the proposed framework can select from the full-feature space while still creating a sparse model. Comparing with standard regression approaches with regularization, the proposed framework integrates the information theoretic criteria in the generalized Lasso model, and sets the most informative features penalty-free to improve prediction accuracy and enhance model interpretability. The flowchart of the proposed integrated feature ranking and selection framework is shown in Fig. 1.

Mathematically, our framework can be formulated as an optimization problem. Let M be the set of indexes of top MI features selected from the MI ranking step. We set indexes in S penalty-free, where S is a subset of M. For each S, we want to solve the following problem.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

$$\text{s.t. } \beta_{j \in S} = 0. \tag{2}$$

The optimization model in our Integrated Feature Ranking and Selection Framework can be solved under generalized lasso framework [37], which is more flexible than lasso and is better in representing the intention to set coefficients of certain informative features penalty-free. Basically, it introduces an arbitrary matrix $D \in \mathbb{R}^{m \times p}, m \leq p$ to define the weights and relations of each element in β.

$$\min_{\beta \in \mathbb{R}^p} \quad ||y - X\beta||_2^2 + \lambda ||D\beta||_1.$$

We can construct a proper D in the generalized lasso framework to adjust penalty levels for each feature. To find such a D, we propose and prove the following two propositions.

Proposition 1

$$\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 + ||(\lambda_1\beta_1, \lambda_2\beta_2, \ldots, \lambda_p\beta_p)||_1.$$

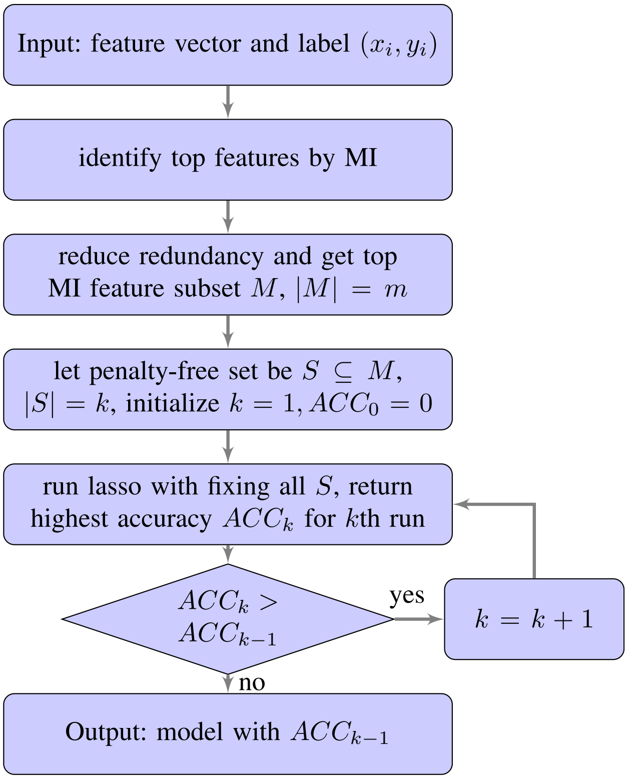
Input: feature vector and label (x_i, y_i)
identify top features by MI
reduce redundancy and get top MI feature subset M, $|M| = m$
let penalty-free set be $S \subseteq M$, $|S| = k$, initialize $k = 1, ACC_0 = 0$
run lasso with fixing all S, return highest accuracy ACC_k for kth run
$ACC_k > ACC_{k-1}$
yes
$k = k + 1$
no
Output: model with ACC_{k-1}

Fig. 1 Flowchart of Integrated Feature Ranking and Selection Model

The above problem of assigning weights λ_k for each feature is equivalent to the generalized lasso with diagonal matrix D and $\lambda_k = d_k\lambda$. (The above formula has also been previously presented as adaptive lasso [36].)

Proof Let D be diagonal matrix $\text{diag}(d_1, d_2, \ldots, d_p)$, we have

$$\begin{aligned}\lambda||D\beta||_1 &= \lambda||(d_1\beta_1, d_2\beta_2, \ldots, d_p\beta_p)^T||_1 \\ &= ||(\lambda_1\beta_1, \ldots, \lambda_p\beta_p)||_1.\end{aligned}$$

If D is $p \times p$ and invertible, β can be transformed into $\theta = D\beta$. The generalized form can be reduced to the standard lasso:

$$\min_{\theta \in \mathbb{R}^p} \quad ||y - XD^{-1}\theta||_2^2 + \lambda||\theta||_1.$$

□

Proposition 2 Without loss of generality, to keeping features $X_{p-k+1}, X_{p-k+2}, \ldots, X_p$ penalty-free is equivalent to setting $d_{p-k+1} = 0, d_{p-k+2} = 0, \ldots, d_p = 0$.

Proof In this case, D is a rank-deficient matrix

$$\text{diag}(d_1, d_2, \ldots, d_{p-k}, 0, \ldots, 0).$$

$$\lambda||D\beta||_1 = ||(\lambda_1\beta_1, \lambda_2\beta_2, \ldots, \lambda_{p-k}\beta_{p-k})||_1.$$

Following the construction procedures in [37], we can transform and reduce the problem to a standard lasso problem. First, we create a full rank matrix $\tilde{D}$ by removing the last k rows from D and adding $k \times p$ matrix A to the bottom, where $m = p - k < p$.

$$\tilde{D} = \begin{bmatrix} d_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & & \\ 0 & \cdots & d_m & & & \vdots \\ 0 & & & 1 & \cdots & 0 \\ \vdots & & & \vdots & \ddots & \vdots \\ 0 & & \cdots & 0 & \cdots & 1 \end{bmatrix}_{p \times p}.$$

In the above matrix $\tilde{D}$, A's rows are clearly orthogonal to those in D. Let $\theta = \tilde{D}\beta = (\theta_a, \beta_b)^T$, where θ_a is related to the coefficient vector β_a of the first m features that are not in the desired set. Now the objective function is

$$\min_{\theta \in \mathbb{R}^p} ||y - X_a\theta_a - X_b\beta_b||_2^2 + \lambda||\theta_a||_1,$$

where X_a is the rescaled first m columns of X, X_b is the original last k columns.

We optimize β_b, θ_a in a sequential way. First, fixing θ_a, the problem regarding β_b is a standard linear regression. The new objective function is to

$$\min_{\theta_a\in\mathbb{R}^m} \left|\left|(1-P)y-(1-P)X_a\theta_a\right|\right|_2^2+\lambda\left|\left|\theta_a\right|\right|_1,$$

where $P = X_b(X_b^TX_b)^{-1}X_b^T$. We get a standard lasso problem regarding θ_a. After solving θ_a, we can in turn determine β_b by $\hat{\beta_b} = (X_b^TX_b)^{-1}X_b^T(y - X_a\hat{\theta_a})$ from the result of linear regression. The solution of the original generalized lasso solution is $\hat{\beta} = \tilde{D}^{-1}\hat{\theta} = \tilde{D}^{-1}\left[\hat{\theta_a}, \hat{\beta_b}\right]^T$. □

Despite the formulation similarity between our model and adaptive lasso [36], adaptive lasso was previously proposed to include a data-dependent weight vector w. The weight vector is estimated as $\hat{w} = 1/|\hat{\beta}|^{\gamma}$ and no element is intended to be zero. From the formulation perspective, adaptive lasso is a special case of generalized lasso with a full-rank diagonal matrix. In our case, we construct D as a (0,1)-matrix that has exact one non-zero element in each row (i.e., $\sum_j d_{ij} = 1$) and at most one non-zero element in each column (i.e., $\sum_i d_{ij} \leq 1$). The column indices of non-zero elements are the features subject to l_1 penalty. The complement set of $p - m$ features are those, we believe, that are information rich and thus set penalty-free.

3.2 Performance evaluation using simulated dataset

To evaluate the performance of the proposed feature selection framework, we used a simulated dataset with binary response and contain $p = 45$ predictors and $n = 50$ samples. The dataset was generated in such a way that only two predictors were related to the response. Using LOOCV, the proposed framework achieved a validation accuracy of 0.92 with five features selected. As a comparison, we also tested the logistic regression (LR) with lasso, which generated a validation accuracy of 0.86 with 8 features selected. The detailed comparison results are summarized in Table 1 as well as Figs. 1 and 2. From those results, one can see clearly that the proposed framework is capable of selecting a model with higher validation accuracy while with less selected features compared to lasso (Fig. 3).

4 Application in the diagnosis of ADHD

4.1 Dataset

This study used a dataset that was collected as part of a larger study from the University of Texas at Austin and the University of Texas Health Science Center in San Antonio by Dr. Margaret Semrud-Clikeman.

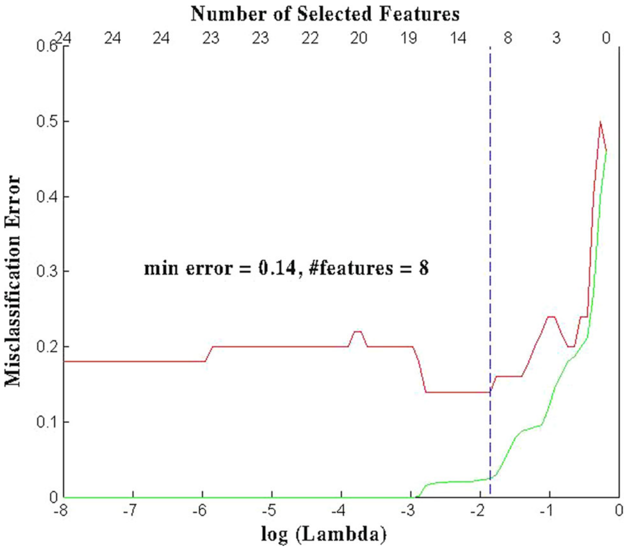

Fig. 2 Best prediction error using LR + lasso (*green curve* as training error, *red curve* as testing error, *dashed line* cuts at min testing error)

A total of 47 subjects matched on gender, SES, and ethnicity participated in the study. All subjects were right handed. There were two groups: 32 ADHD-Combined participants and 15 healthy subjects in a control group. All ADHD participants had less than 15 standard score point differences between general conceptual ability (DAS-GCA) and all achievement measures. The ADHD subjects were matched on severity of symptoms as measured by Conners' Ratings Scale (Conners, 1998a). All ADHD subjects met DSM IV-TR criterion for ADHD Combined-type and no other psychiatric or psychological disorder including Learning Disorders, Anxiety Disorders, Mood Disorder, or Oppositional Defiant Disorder. Control participants did not meet any criteria for a psychiatric or learning diagnosis nor have a history of medication treatment. All participants were recruited from a diversity of socioeconomic and ethnic backgrounds in order to control for potential group differences.

MRI images are acquired at the University of Texas Health Science Center at San Antonio using three-dimensional gradient recalled acquisitions in the study state (3D-GRASS) with a repetition time (TR) = 33 msec, echo time (TE) = 12 msec, and a flip angle of 60 degrees to obtain a 256 × 192 × 192 volume of data with a spatial resolution

Table 1 Performance comparison on simulated dataset

Method	Validation accuracy	Training accuracy	Features selected
Our Model	0.92	0.94	5
LR + lasso	0.86	0.97	8

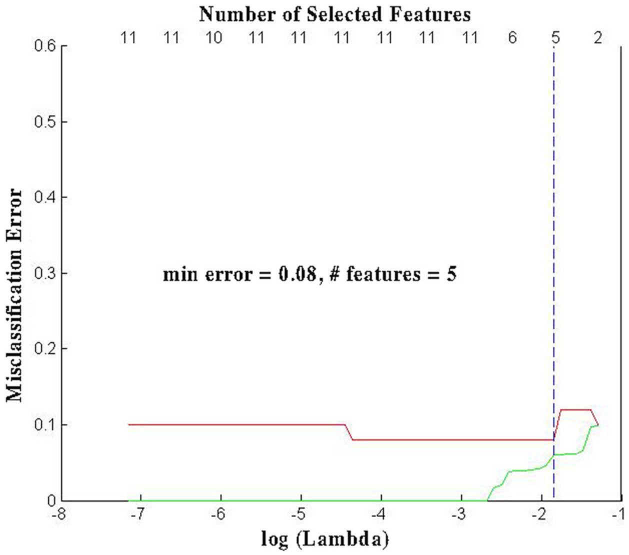

Fig. 3 Best prediction error using our framework (*green curve* as training error, *red curve* as testing error, *dashed line* cuts at min testing error)

of 1mm × 1mm × 1mm. Then all MRI images were processed and normalized using the FreeSurfer image analysis suite [38, 39] by Dr. Jesse Bledsoe on a Linux platform at MSU. All regions of interest (ROI) in the FreeSurfer suite (45 cortical ROIs) were developed using an automated labeling system based on gyral regions of the Desikan-Killiany Atlas [40]. We employed the brain cortical thickness of those ROIs as possible features for ADHD feature characterization and selection in this study.

4.2 Results of feature ranking and decorrelation step

The first step in our framework is to perform feature ranking using mutual information. The top ten features of cortical thickness with highest MI were picked first for further analysis. They are right rostral anterior cingulate (MI = 0.124), total rostral anterior cingulate (0.122), left rostral anterior cingulate (0.078), left caudal middle frontal (0.071), right frontal pole (0.068), right lateral orbito frontal (0.063), left caudal anterior cingulate (0.062), total caudal middle frontal (0.051), left inferior parietal (0.051), and left pars orbitalis (0.05). In the next step, we calculated the correlation between each pair of the high-ranked 10 features. If the correlation of a pair of features is 0.6 or higher, we consider one feature in the pair to be redundant, and remove the feature with a lower mutual information value. In this way, the following two features were removed: total rostral anterior cingulate and total caudal middle frontal. The remaining eight features were used as feature candidates in the Lasso-based feature selection step.

4.3 Results of feature selection step

4.3.1 Comparison of testing accuracy

In feature selection step, within top eight highest MI and uncorrelated feature set, we started with fixing the single top feature penalty-free, then all combinations of two features penalty-free, then all combinations of three features, iteratively. We evaluated the selection and prediction model using the validation accuracies in a LOOCV procedure. The model search process stops at fixing four features penalty-free, as when fixing more features, the validation accuracy started to decrease. The resulting model is the best prediction model with the highest LOOCV validation accuracy. As shown in Table 2, the proposed framework achieved a testing accuracy of 0.81 with a sensitivity of 0.81 and a specificity of 0.80.

In addition, we also tested and compared the performance of the state-of-the-art feature selection algorithms, including the aforementioned information theoretic methods MRMR [28], MIFS [27], JMI [29], CMIM [30], MIM [26], as well as the popular Pudil's floating search method [41], and the principle component analysis (PCA)-based approach, for which we took the components that account for 95 % of data variance as the selected features in prediction. The prediction results of these approaches are also summarized in Table 2. One can observe that the proposed method achieved higher validation accuracy (0.81) than all other compared feature selection approaches, while using the lowest number of features in the final prediction model. These experimental results confirmed that our model is efficient to select the most predictive features of ADHD given a small sample size.

4.3.2 Analysis of features in best models

To investigate the model interpretability, we also checked the locations of the selected cortical thickness. All the features (regions of interest) selected by the best models were located in prefrontal cortex (PFC), anterior cingulate cortex, and parietal cortex. Structural and functional impairments are in accordance with current understanding of brain–behavior relationships in ADHD.

The prefrontal cortex (PFC) is connected with nearly every cortical structure of the central nervous system [42] and is involved in nearly all aspects of human personality and cognition. The PFC has received much attention in the ADHD literature given a large body of research on impairments in tests thought to tap PFC functioning [43, 44]. For example, the PFC has been implicated in complex behavior relevant to central impairments in ADHD such as inhibitory control [45, 46], attention, working memory, and planning [42, 47]. Furthermore, specific differences within

Table 2 Comparison of testing results (leave-one-out cross-validation)

Selected features	Testing accuracy	Training accuracy	Sensitivity	Specificity	Selection method
4	0.81	0.87	0.81	0.80	Proposed method
5	0.76	0.78	0.75	0.80	MRMR [28]
7	0.66	0.76	0.66	0.67	Pudil's floating search [41]
14	0.70	0.74	0.72	0.67	PCA
5	0.74	0.75	0.81	0.60	MIM [33]
5	0.70	0.76	0.69	0.73	MIFS [27]
5	0.72	0.78	0.72	0.73	JMI [29]
5	0.74	0.76	0.75	0.73	CMIM [30]

the frontal pole and orbital frontal cortex observed here may provide further evidence for impairments in frontal limbic structures and emotional disorders which often co-occur in children with ADHD [48].

The anterior cingulate cortex is a key structure implicated in attentional control [47]. It is implicated in a wide variety of cognitive operations including response inhibition, reward processing, behavioral motivation, target detection, and decision making [49]. Functional neuroimaging studies suggest hypoactivation of areas of the anterior cingulate in children and adults with ADHD [50–52]. Studies observed decreased activation of the anterior cingulate in tasks thought to require behavioral inhibition (e.g., counting Stroop task) in children with ADHD compared to controls [50, 52] also reported reduced activation of the anterior cingulate during tasks of behavioral inhibition (e.g., stop signal task) in children with ADHD-C. Further, cortical thinning of the anterior cingulate cortex has been demonstrated in adults with ADHD [53]. Moreover, the right rostral anterior cingulate cortex (ACC) contributed the most predictive variance in classifying those with ADHD from typically developing controls. This finding supports the hypothesis that abnormal development of the the right ACC, in particular, may be considered a biomarker for ADHD and inhibitory control [54]. The ACC is likely implicated in ADHD due to its involvement in complex behavior. However, the ACC, itself, is unlikely to contribute to impaired attention. Rather, future work will need to address the complex networks and systems that involve the ACC in order to provide valid causal pathways for ADHD.

The left inferior parietal cortex also contributed to the classification of ADHD versus healthy children. This was a particularly interesting finding given recent work that has implicated abnormalities in parietal cortex during resting-state functional MRI [55]. Prior to this work, the posterior cortex was proposed to underlie the basis for arousal and vigilance which were considered precursors for targeted attention [47, 56]. And, more recent work has found the posterior parietal lobe to be important for shifting attention during dynamic attention tasks [57]. Structurally, reduced cortical thinning of the right-parietal cortex has also been observed in adults with ADHD [53]. Taken together, the parietal cortex, likely due to its frontal projections, is another important area in the attention network that may undergo abnormal development in those with ADHD.

The prefrontal cortex, anterior cingulate cortex, and parietal cortex have all been implicated in attentional control and ADHD. Given these regions provided the best classification of ADHD from controls, the proposed model would appear to be theoretically valid. A significant advantage of the proposed approach is that we novelly integrate the information theoretic feature selection framework with the generalized lasso framework. Through adaptively manipulating penalty weights of each feature in regularization term, we are able to preserve the most informative features in the final model and eliminate less informative and redundant features.

5 Conclusion

ADHD feature characterization and selection has never been an easy task. In this paper, the proposed integrated feature ranking and selection framework provides a sparse, accurate, and highly interpretable model to assist ADHD feature characterization. With the proposed two-step formulation, one can integrate information theory conveniently to supervise the feature selection process while the optimal solutions can be guaranteed due to the convex optimization formulations in a generalized lasso framework. The information-guided selection structure enforces the most useful discriminative predictors to be included in the final prediction model while eliminating less-informative and redundant variables to create an accurate sparse prediction model. In addition to mutual information, due to the flexible structure of the proposed framework, one can also conveniently integrate clinical prior knowledge into the feature selection model. For example, one can set clinician-identified potentially important features penalty-free and encourage them to be included in the final prediction model. The information theory-guided and clinical prior knowledge-guided feature selection framework will be greatly useful to construct prediction models that are more transparent and interpretable by medical and healthcare professionals. Such a supervised feature selection framework is highly demanded in making clinical decisions compared to the 'black box' predictive models generated by traditional machine learning algorithms. As this is a general feature selection approach, the proposed technique can also be applied to other decision-making problems that require interpretable prediction models. The research in this study also suggest that machine learning techniques can be useful tools for understanding and measuring brain abnormalities associated with ADHD.

References

1. Swanson JM, Sunohara GA, Kennedy JL, Regino R, Fineberg E, Wigal T, Lerner M, Williams L et al (1998) Association of the dopamine receptor d4 (drd4) gene with a refined phenotype of attention deficit hyperactivity disorder (adhd): a family-based approach. Mol Psychiatr 3:38–41
2. Cheng W, Ji X, Feng J (2012) Individual classification of adhd patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. Front Syst Neurosci 6(58)
3. Ahmadlou M, Adeli H (2010) Wavelet-synchronization methodology: a new approach for eeg-based diagnosis of adhd. Clin EEG Neurosci 41(1):1–10
4. Holtmann M, Becker K, Kentner-Figura B, Schmidt MH (2003) Increased frequency of rolandic spikes in adhd children. Epilepsia 44:1241–1244
5. Hynd GW, Semrud-Clikeman M, Lorys AR, Novey ES, Eliopulos D, Lyytinen H (1991) Corpus callosum morphology in attention deficit-hyperactivity disorder: morphometric analysis of mri. J Learn Disabil 24(3):141–146
6. Linden DEJ (2012) The challenges and promise of neuroimaging in psychiatry. Neuron 73:8–22
7. Tenev A, Markovska-Simoska S, Kocarev L, Pop-Jordanov J, Muller A, Candrian G (2014) Machine learning approach for classification of adhd adults. Int J Psychophysiol 93(1):162–166
8. Peng X, Lin P, Zhang T, Wang J (2013) Extreme learning machine-based classification of adhd using brain structural mri data. PloS One 8(11):e79476
9. Zhao P, Bin Y (2006) On model selection consistency of lasso. J Mach Learn Res 7:2541–2563
10. Weaver W (1949) Recent contributions to the mathematical theory of communication. The mathematical theory of communication. University of Illinois Press, Urbana
11. Hastie T, Tibshirani R (1996) Discriminant analysis by gaussian mixtures. J R Stat Soc Ser B 58:155–176
12. Posner MI, Fan J (2008) Attention as an organ system. In: Pomerantz JR (ed) Topics in integrative neuroscience: from cells to cognition. Cambridge University Press, Cambridge
13. Barkley RA (1997) Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of adhd. Psychol Bull 121(1):65–94
14. Dougherty DD, Bonab AA, Spencer TJ, Rauch SL, Madras BK, Fischman AJ (1999) Dopamine transporter density in patients with attention deficit hyperactivity disorder. The Lancet 354(9196):2132–2133
15. Nigg J, Willcutt EG, Doyle AE, Sonuga-Barke EJS (2005) Casual heterogeneity in attention-deficit hyperactivity disorder: do we need neuropsychologically impaired subtypes. Biol Psychiatr 57:1224–1230
16. Sagvolden T, Sergeant JA (1998) Attention deficit hyperactivity disorder: from brain dysfunctions to behaviour. Behav Brain Res 94(1):1–10
17. Sonuga-Barke EJS (2002) Psychological heterogeneity in adhd—a dual pathway model of behaviour and cognition. Behav Brain Res 130(1):29–36
18. Swanson JM, Elliott GR, Greenhill LL, Wigal T, Arnold E, Vitiello B et al (2007) Effects of stimulant medication on growth rates across 3 years in the mta follow-up. J Am Acad Child Adolesc Psychiatr 46(8):1015–1027
19. Castellanos FX, Giedd JN, Berquin PC, Walter JM, Sharp W, Tran T (2001) Quantification brain magnetic resonance imaging in girls with attention-deficit/hyperactivity disorder. Arch Gen Psychiatr 58:289–295
20. Castellanos FX, Giedd JN, Eckburg WL, Marsh AC, Kaysen D, Hamburger SD (1994) Quantitative morphology of the caudate nucleus in attention deficit hyperactivity disorder. Am J Psychiatr 151(1212):1791–1796
21. Castellanos FX, Giedd JN, Marsh AC, Hamburger SD, Vaiturzis AC, Dickstein DP (1996) Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. Arch Gen Psychiatr 53(7):607–616
22. Castellanos FX, Lee PP, Sharp W, Jeffries NO, Greenstein DK, Clasen LS et al (2002) Developmental trajectories of brain volume abnormalities in children with adolescents with attention-deficit/hyperactivity disorder. J Am Med Assoc 28(4):1740–1749
23. Durston S, Hulshoff HG, Schnack JK, Buitelaar MP Steenhuis, Minderaa RB et al (2004) Magnetic resonance imaging of boys with attention-deficit/hyperactivity disorder and their unaffected siblings. J Am Acad Child Adolesc Psychiatr 43(3):332–340
24. Cover TM, Thomas JA (2001) Elements of information theory, 2nd edn., Wiley Series in Telecommunications and Signal ProcessingWiley-Interscience, Hoboken
25. Tourassi GD, Frederick ED, Markey MK, Floyd EC Jr (2001) Application of the mutual information criterion for feature selection in computer-aided diagnosis. Am Assoc Phys Med
26. Lewis D (1992) Feature selection and feature extraction for text categorization. In: Proceedings of Speech and Natural Language Workshop. Morgan Kaufmann, pp 212–217
27. Battiti R (1994) Using the mutual information for selecting in supervised neural net learning. IEEE Trans Neural Netw 5:537–550
28. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226
29. Yang H, Moody J (1999) Feature selection based on joint mutual information. In: Proceedings of international ICSC symposium on advances in intelligent data analysis
30. Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555

31. Guyon I (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
32. Kohavi Ron, John George H (1997) Wrapper for feature subset selection. Artif Intell 97:273
33. Kwak N, Choi C-H (2002) Input feature selection by mutual information based on parzen window. IEEE Trans Pattern Anal Mach Intell 24:1667
34. Tibshirani R (1994) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
35. Kim Y, Kwon S, Choi H, Shen X (2012) Consistent model selection criteria on high dimensions. J Mach Learn Res 13(1):1037–1057
36. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
37. Tibshirani RJ, Taylor J (2011) The solution path of the generalized lasso. Ann Stat 39(3):1335–1371
38. Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. i. segmentation and surface reconstruction. NeuroImage 9:179–194
39. Dale AM, Sereno MI (1993) Improved localization of cortical activity by combining eeg and meg with mri cortical reconstruction: A linear approach. J Cognit Neurosci 5:162–176
40. Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D et al (2006) An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. NeuroImage 31:968–980
41. Pudil P (1994) Floating search methods in feature selection. Pattern Recognit Lett 15(11):1119–1125
42. Fuster JM (2008) The Prefrontal Cortex, 4th edn. Academic Press, London
43. Martinussen R, Hayden J, Hogg-Johnson S, Tannock R (2005) A meta-analysis of working memory impairments in children with attention-deficit hyperactivity disorder. J Am Acad Child Adolesc Psychiatr 44:377–384
44. Nigg J, Blaskey LG, Huang-Pollock CL, Rappley MD (2002) Neuropsychological executive functions and dsm-iv adhd subtypes. J Am Acad Child Adolesc Psychiatr 41:59–66
45. Aron AR, Dowson JH, Sahakian BJ, Robbins TW (2003) Methylphenidate improves response inhibition in adults with attention-deficit hyperactivity disorder. Biol Psychiatr 54(12):1465–1468
46. Rubia K et al (1999) Hypofrontality in attention deficit hyperactivity disorder during higher-order motor control: a study with functional mri. Am J Psychiatr 156(6):891–896
47. Posner MI, Petersen SE (1990) The attention system of the human brain. Ann Rev Neurosci 13:25–42
48. Proal E, Castellanos FX (2012) Large-scale brain systems in adhd: beyond the prefrontal striatal model. Trends Cognit Sci 16(1):17–26
49. Bush G, Luu P, Posner MI (2000) Cognitive and emotional influences in anterior cingulate cortex. Trends Cognit Sci 4(6):215–222
50. Bush G, Frazier JA, Rauch SL, Seidman LJ, Whalen PJ, Jenike MA, Rosen BR, Biederman J (1999) Anterior cingulate cortex dysfunction in attention-deficit hyperactivity disorder revealed by fmri and the counting stroop. Biol Psychiatr 45:1542–1552
51. Ernst M et al (2003) Neural substrates of decision making in adults with attention deficit hyperactivity disorder. Am J Psychiatr 160:1061–1070
52. Pliszka SR et al (1999) Neuroimaging of inhibitory control areas in children with attention deficit hyperactivity disorder who were treatment naive or in long-term treatment. Am J Psychiatr 163(6):1052–1060
53. Makris N, Biederman J, Valera EM, Bush G, Kaiser J, Kennedy DN, Caviness VS, Faraone SV, Seidman LJ (2007) Cortical thinning of the attention and executive networks in adults with attention-deficit hyperactivity disorder. Cereb Cortex 17(6):1364–1375
54. Semrud-Clikeman M, Pliszka SR, Bledsoe J (2013) Anterior cingulate cortex and symptom severity in attention-deficit/hyperactivity disorder. J Abnorm Psychol 122(2):558
55. Bush G (2011) Cingulate, frontal, and parietal cortical dysfunction in attention-deficit/hyperactivity disorder. Biol Psychiatr 69(12):1160–1167
56. Sturm W et al (1999) Functional anatomy of intrinsic alertness: evidence for a fronto-parietal-thalamic-brainstem network in the right hemisphere. Neuropsychologia 37(7):797–805
57. Sapir A, Hayes A, Henik A, Danziger S, Rafal R (2004) Parietal lobe lesions disrupt saccadic remapping of inhibitory location tagging. J Cognit Neurosci 16:503–509

Cao Xiao is a Ph.D. candidate in industrial and systems engineering at the University of Washington, Seattle, WA, USA. Previously, she received her bachelor's degree and master's degree in mathematics from Peking University, Beijing, China and Boston College, Boston, MA, USA, respectively. Her current research interests include data mining and optimization, machine learning, intelligent decision-making systems, and big data applications. Particularly, she enjoys developing efficient data mining methods to enable effective problem solving for bioinformatics and social network applications.

Jesse Bledsoe is a Ph.D. candidate, and he is an acting Assistant Professor in the Department of Psychiatry and Behavioral Sciences at the University of Washington School of Medicine and clinical neuropsychologist. His current research focuses on behavioral measurement and brain–behavior relationships in psychiatric and medical disease. Current research includes diagnostic classification of psychiatric diseases as well as neuroanatomical and neurocognitive late effects of radiation in pediatric brain tumors.

Shouyi Wang is an Assistant Professor in the Department of Industrial, Manufacturing, and Systems Engineering and the COSMOS at the University of Texas at Arlington (UTA). He received his B.S. degree in control science and engineering from the Harbin Institute of Technology, China, in 2003; M.S. degree in systems and control engineering from the Delft University of Technology, the Netherlands, in 2005; and Ph.D. degree in Industrial and Systems Engineering from the Rutgers University in 2012. His research interests include data mining, pattern recognition, multivariate time series, and applied operations research with applications to healthcare decision-making systems, computational neuroscience, and medical informatics.

Wanpracha Art Chaovalitwongse is a Professor in the Departments of Industrial & Systems Engineering, Radiology (joint), and Bioengineering (adjunct) at the University of Washington, Seattle (UW). He also serves as an Associate Director of the Integrated Brain Imaging Center (IBIC) at UW Medical Center. He received his M.S. and Ph.D. degrees in Industrial & Systems Engineering from the University of Florida in 2000 and 2003, respectively. His research group conducts basic computational science, applied, and translational research at the interface of engineering, medicine, and other emerging disciplines.

Sonya Mehta is a Ph.D student in the Department of Psychology at University of Washington. Her research interests mainly focus on Cognition and Perception, particularly, investigating the organization of the brain and the basis of cognition using MRI imaging approaches.

Margaret Semrud-Clikeman received her Ph.D. degree from the University of GA in pediatric neuropsychology/school psychology, and she completed her internship and post-doctoral fellowship at the Massachusetts General Hospital/Harvard Medical School, and she holds a Diplomate from the American Board of Pediatric Neuropsychology. She has worked as a neuropsychologist/psychologist for the past 42 years. In addition, her current research involves neuroimaging in neurodevelopmental disorders, neurocognitive functioning following recovery from cerebral malaria, and concussions. These continue to be areas of current research interest.

Thomas Grabowski is a UW Professor of radiology and joint professor of neurology. He is a neurologist and an expert on the diagnosis and treatment of conditions affecting memory, language, and cognition. He is the director of the UW Medicine Memory and Brain Wellness Center. He is also a cognitive neuroscientist who directs the UW Integrated Brain Imaging Center, dedicated to investigating the organization of the brain and the basis of cognition using MRI imaging approaches.