The Predictive Power of Structural MRI in Autism Diagnosis

Gajendra J. Katuwal[1, 2], Nathan D. Cahill[3], Stefi A. Baum[1,4], Andrew M. Michael[1, 2]

[1]*Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, USA*
[2]*Autism and Developmental Medicine Institute, Geisinger Health System, Danville, PA, USA*
[3]*School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, USA*
[4]*Faculty of Science, University of Manitoba, Winnipeg, Canada*

***Abstract*— Diagnosis of Autism Spectrum Disorder (ASD) using structural magnetic resonance imaging (sMRI) of the brain has been a topic of significant research interest. Previous studies using small datasets with well-matched Typically Developing Controls (TDC) report high classification accuracies (80-96%) but studies using the large heterogeneous ABIDE dataset report accuracies less than 60%. In this study we investigate the predictive power of sMRI in ASD using 373 ASD and 361 TDC male subjects from the ABIDE. Brain morphometric features were derived using FreeSurfer and classification was performed using three different techniques: Random Forest (RF), Support Vector Machine (SVM) and Gradient Boosting Machine (GBM). Although high classification accuracies were possible in individual sites (with a maximum of 97% in Caltech), the highest classification accuracy across all sites was only 60% (sensitivity = 57%, specificity = 64%). However, the accuracy across all sites improved to 67% when IQ and age information were added to morphometric features. Across all three classifiers, volume and surface area had more discriminative power. In general, important features for classification were present in the frontal and temporal regions and these regions have been implicated in ASD. This study also explores the effect of demographics and behavioral measures on the predictive power of sMRI. Results show that classification accuracy increases with autism severity and that ASD detection with sMRI is easier before the age of 10 years.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a polygenetic neurodevelopmental disorder characterized by repetitive behavior, intellectual disability, and impaired language and social skills. The behavioral phenotype of ASD is well characterized, but its etiology and pathogenesis remains elusive. However, structural magnetic resonance imaging (sMRI) studies have reported subtle anatomical differences in brain structures of ASD subjects versus Typically Developing Control (TDC) subjects, including differences in the structure of the frontal lobe, parietal lobe, temporal lobe, limbic system and cerebellum [1].

At present, ASD diagnosis is primarily based on behavioral criteria. This approach is subjective, time consuming and does not help understanding the underlying etiology. This makes ASD diagnosis based on imaging data highly desirable. Brain biomarkers derived from sMRI can help identify the neuroanatomical basis of heterogeneity in ASD and can be a powerful tool for early diagnosis and intervention. Mass-univariate techniques such as Voxel Based Morphometry (VBM) are commonly employed in brain imaging studies to detect brain anatomical differences. Although VBM has high exploratory power, it lacks the statistical power required to detect subtle multivariate structural differences. Multivariate pattern recognition techniques (MVPT) are capable of detecting subtle and spatially distributed differences in data and thus hold promise for extracting higher predictive power.

In recent years, there have been a number of studies applying MVPT for ASD vs. TDC classification. These studies can be broadly categorized into two groups based on the data used: 1) data matched for demographics and behavioral (DB) measures such as age, sex and IQs [2]–[5] and 2) large heterogeneous data [6], [7]. The following classification methods and accuracies have been reported by the first group of studies: Support Vector Machine (SVM) on gray matter scans (81%) [2], Logistic Model Trees (LMT) on regional cortical thickness (87%) [3], SVM on gray matter in default mode network regions (90%) [4], and SVM on regional and inter-regional cortical and subcortical features (96%) [5]. The second group of studies uses the large heterogeneous ABIDE dataset and reports classification accuracies less than 60% [6], [7].

In this study, we investigate the predictive power of sMRI in ASD utilizing three heterogeneous classifiers– Random Forest (RF) [8], SVM [9] and Gradient Boosting Machine (GBM) [10]. We perform classification within each and across all sites of the ABIDE dataset. We investigate the effects of DB measures on the predictive power of sMRI in ASD and the incremental power that can be gained from them. This has not been addressed by previous studies. In addition, we investigate the relationship between Autism Diagnostic Observation Score (ADOS) and autism class probability, an autism score produced by a classifier. In addition we compare the discriminative power of morphometric properties such as volume, area, thickness, curvature and folding index with three classifiers. We discuss issues related to over-fitting due to small sample size and feature selection, using results from individual sites. Finally, we conclude by discussing the challenges and future directions on predicting ASD using sMRI.

II. METHODS

A. Data

sMRI of 373 male TDC and 361 male ASD from 15 scanning sites of the ABIDE dataset [11] were used in this study.

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

B. Image Processing

Seven different morphometric properties: surface area, volume, Gaussian curvature (*gauscurv*), mean curvature (*meancurv*), folding index (*foldind*), thickness, thickness standard deviation (*thicknessstd*) of cortical and subcortical structures were derived using the *recon-all* workflow of FreeSurfer [12]. This workflow includes motion correction, brain extraction, Talairach transformation, segmentation of cortical and subcortical structures, intensity normalization, gray matter-white matter boundary tessellation, and topology correction.

C. Classification

The predictive power of sMRI was investigated in three different experiments. In Experiments 1 & 2, subjects across all sites were used and classification was performed without selection of optimal features. In Experiment 3 classification was performed within each individual site after selecting optimal features. Experiment 2 was performed with three heterogeneous classifiers: RF, SVM and GBM and, for Experiments 1 and 3, RF was used since its performance was marginally better than that of the other two classifiers.

1) Experiment 1

This experiment explores the effects of DB measures on classification accuracy. To minimize the model bias, 29 different random forests with different parameters were built for ASD vs. TDC classification. The *mtry* parameter of RF, which represents the number of predictors used for splitting at each node during the construction of decision trees, was varied from 10 to 38. All the brain morphometric features derived from FreeSurfer were used to train RF models under a leave-one-out cross validation (LOOCV) framework. For each test subject, mean accuracy and mean probability of autism (POA) were calculated by averaging test accuracy and POA across the RFs. POA of a test subject is the proportion of decision trees that classify it as class "ASD." The results of this classification framework are presented in Fig. 1.

2) Experiment 2

This experiment investigates the discriminative power of different morphometric properties of the brain in ASD vs. TDC classification. RF, SVM and GBM predictive models were built for each morphometric property using all subjects. Parameter search was performed under a repeated 10-fold cross validation framework with 20 repetitions (CV10-20) using the *caret* package of R software. Optimum parameters based on classification accuracy metric were chosen to train the final models in 10-fold cross validation framework (CV10). Results across the 10 test folds were collected to calculate the area under the ROC curve (AUC) as a measure of discriminative power. The results of this experiment are in Table 1.

3) Experiment 3

This experiment investigates how the highest classification accuracies and the optimal features change with individual sites. Unlike experiment 1 and 2, feature selection was performed and predictive models were built for each site. For each morphometric property, an RF model was built in each of the 15 individual sites. Recursive Feature Elimination (RFE) was performed under a CV10-20 framework to select the optimal features. In RFE, a model is fitted using all the predictors, and then each predictor is ranked according to its importance to the model. At each iteration of RFE, the model is refitted using the top n predictors, and the optimum value of $n=n^*$ is chosen to give the best model performance. Finally, the top n^* predictors are used to train the final model under a CV10-20 framework. The results of this experiment are presented in Table 2.

III. RESULTS

A. Experiment 1: Effect of DB measures

The highest RF classification accuracy of 60% (sensitivity = 57%, specificity = 64) was achieved when *mtry* = 25 under LOOCV. Similar classification accuracies were achieved from the ensemble of 29 forests with both majority voting scheme as well as the weighted average of class probabilities. However, when RF class probabilities were combined by a logistic model classification accuracy dropped below 60%.

1) Mean Classification Accuracy Increases with ADOS

Results presented in Sections 1-2 are from ASDs only since ADOS were not available for the majority of TDCs. A significant positive correlation (r = 0.533, p-value = 0.041*) was observed between mean classification accuracy and ADOS; see Fig 1(b). To our knowledge, no previous studies have reported this relationship.

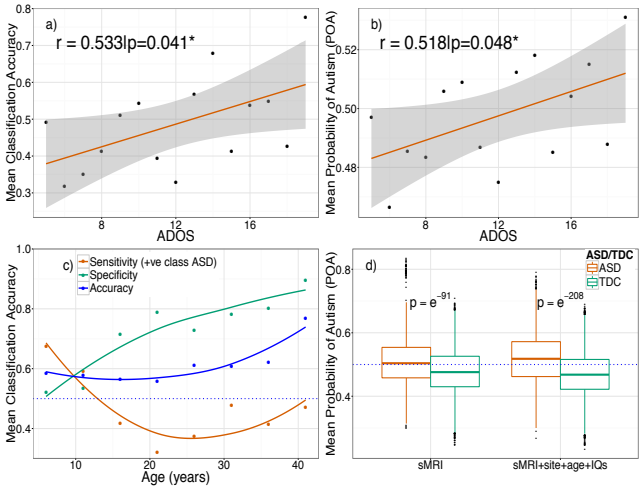

Fig. 1: Effect of demographics & behavioral measures on classification accuracy. 1(a, b) Mean accuracy across ASD patients were calculated for each ADOS. Shaded region represents 95% confidence interval.

2) Relationship between POA & ADOS

A significant positive correlation (r = 0.518, p-value = 0.048*) was observed between mean POA and ADOS for ASD subjects; see Fig 1(b). Ecker et al. [2] also reported a significant positive correlation between the distance from the optimal hyper plane of a SVM (comparable to POA of a RF model) and ADOS. Jiao et al. [3] reported positive correlation (not significant) between autism severity and the output weight of a LMT (comparable to POA of a RF model).

The above findings demonstrate that the severity of autism maps to a spectrum of brain anatomical features. In other words, behavioral features of ASD can be captured through multivariate patterns of sMRI features. In addition,

this approach provides a framework to find sub-groups of ASD with distinct brain structure–behavior associations.

3) Mean Accuracy vs. Age

Classification accuracy was found to be dependent on the DB measures age, site, and IQ. Fig. 1(c) presents the effect of age on classification accuracy. Sensitivity (positive class is ASD), specificity and mean accuracy were calculated over 5-year age intervals. In Fig. 1(c), sensitivity (red) is higher than specificity (green) before the age of 10 years, suggesting that it is easier to detect ASD than TDC. The sensitivity decreases till the age of 20 years and again increases after the age of 30 years, suggesting that the ASD subjects in the 20-30 years age group are the most difficult to recall. Although the results across 29 RFs were aggregated to minimize the model bias, this particular trend observed can be dependent on the model choice. Nonetheless, this result demonstrates the dependence of sensitivity, specificity and accuracy on age, and DB measures in general. This suggests that brain morphometric differences between ASD & TDC groups change with respect to the DB measures. This motivated us to incorporate DB measures in classification.

4) Incremental Predictive Power from Demographic and Behavioral Measures

P-value of POA_{ASD} versus POA_{TDC} two-sample t-test was used as a measure of the discriminative power in ASD vs. TDC classification. The discriminative power of sMRI (p = 1E-91) and the total discriminative power from sMRI features + age + IQs (p = 1E-208) are presented in Fig. 1(d). This result demonstrates that brain structure and DB measures contain non-overlapping information helpful in diagnosing ASD. Through this result, we show that augmenting DB measures to brain imaging features in a prediction model would be beneficial.

B. Experiment 2: Predictive power of individual morphometric property

1) Discriminative Power of Morphometric properties

The discriminative powers of the morphometric properties have been compared based on the AUC metric and its *p*-value. In Table 1, volume and area have the highest AUC or discriminative power according to all the three classification models. AUC is slightly increased to 0.6 when features from all morphometric properties were utilized. Similarly, thickness has the lowest AUC except in SVM. This finding is opposite to the finding of Jiao et al. [3], where it was reported that thickness-based diagnostic models are superior to volume-based models. Their result was based on LMT and used a smaller sample size of 38.

2) Features of Importance in Classification

For four morphometric properties, the top 10 most important features for ASD vs. TDC classification using RF, SVM and GBM are presented in Fig. 2. For each morphometric property, most of the important features or brain structures are common across classification models. Most of the brain structures are from frontal and temporal regions, particularly from social and language regions, which are consistently implicated in ASD. In classification using volume, the common structures across methods such as bank of superior temporal sulcus (*bankssts*), amygdala, parahippocampus and ventricles have been frequently implicated in ASD [13]. Similarly for thickness and mean curvature, common regions across methods such as parahippocampal, frontal-pole, pericalcrine, superior-temporal are frequently implicated in ASD [1]. Moreover, the same set of structures such as third ventricle, amgydala and parahippocampal regions show up as important features while performing classification using all the morphometric properties.

Table 1: Discriminative power of morphometric properties based on AUC metric.

	AUC | *p*-value		
	RF	SVM	GBM
Volume	0.59 | 3E-5	0.58 | 8.1E-5	0.58 | 3E-4
Area	0.59 | 4.3E-5	0.59 | 1.1E-5	0.57 | 8.3E-4
Thickness	0.51 | 0.64	0.57 | 4.5 E-4	0.54 | 0.046
Folding Index	0.57 | 2E-3	0.52 | 0.43	0.57 | 1.6E-3
Mean Curvature	0.56 | 2.6E-3	0.56 | 3.6E-3	0.57 | 1.6E-3
Gauss Curvature	0.57 | 6.4E-4	0.53 | 0.12	0.57 | 2E-3
All	0.60 | 8.3E-6	0.59 | E-5	0.60 | 6.6E-6

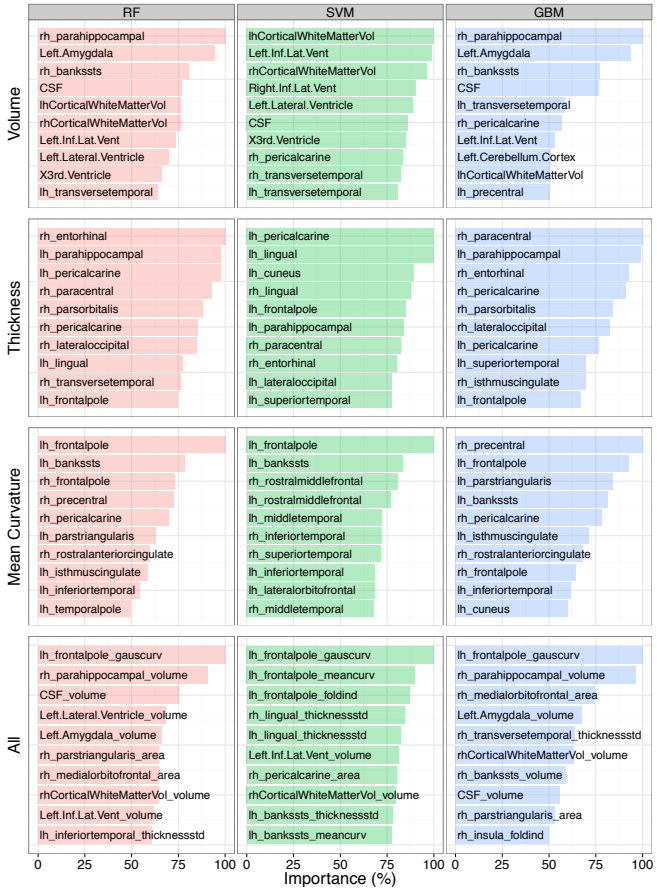

Fig. 2: Top 10 features for each morphometric property for ASD vs. TDC classification; many features are common across three different classification models.

C. Experiment 3: Classification within sites with feature selection

1) High Classification Accuracies in Individual Sites

The highest accuracy achieved and the top two predictors

in each site are presented in Table 2. The classification accuracies range from 66% in NYU utilizing thickness standard deviation of 11 brain structures to 97% in CAL utilizing area of just 5 brain structures. In general, classification accuracies increased in individual sites after RFE and were similar to that of previous studies [2]–[5].

Table 2: Highest classification accuracies and the most important features in within-site classification.

Site	Size	A %	# F	Top features
CAL	15/15	97	5	bankssts_area, parsopercularis_area
CMU	9/9	94	1	pericalcarine gauscurv
KKI	22/15	84	15	frontalpole foldind, inferiortemporal_foldind
KUL	27/25	77	64	rostralmidfrontal_gauscurv, lateralorbitofrontal_gauscurv
MPG	26/12	82	2	parstriangularis_gauscurv, lateralorbitofrontal_gauscurv
NYU	71/58	66	11	precuneus_thicknessstd, lingual_thicknessstd
OHSU	14/12	77	12	fusiform_area, Accumbens_volume
OLIN	6/8	86	1	rostralanteriorcingulate_thicknessstd
PITT	13/15	82	16	CC.Mid.Anterior_volume, superiortemporal_foldind
SBL	14/14	86	1	temporalpole_thickness
SJH	24/23	70	14	middletemporal_gauscurv, isthmuscingulate_gauscurv
UCLA	30/36	64	2	inferiorparietal_area, X4th.Ventricle_volume
UM	47/50	72	3	entorhinal_thicknessstd, superiorfrontal_thicknessstd
USM	37/51	74	61	parsorbitalis_area, X3rd.Ventricle_volume
YALE	18/18	75	2	superiorparietal_thickness, medialorbitofrontal_thickness

A-Accuracy, #F-Number of optimal features

2) Overfitting due to Small Sample Size and Feature Selection

High classification accuracies achieved in individual sites and low accuracies achieved across sites imply that the prediction model fitted for individual sites suffer from overfitting. For classification across sites, the important features were common even across three different classification models. However, the important features of individual sites varied between sites even when only one classifier (i.e. RF) was utilized. This demonstrates the heterogeneity in ASD as well as the issue of over fitting. The overfitting arises from both small sample size and feature selection. Similarly, a significant negative correlation (r = -0.707, p-value = 0.003*) between the sample size of a site and the highest accuracy achieved in that site was noted. This was caused mainly by the fact that increase in sample size results in increased heterogeneity. As the data becomes more heterogeneous, the training, validation and testing partitions of the dataset become more dissimilar and hence the overall accuracy across the test partitions decreases. We present this result to suggest that high accuracy rates reported in previous studies using small sample sizes can be due to over-fitting and their results cannot be generalized with confidence. Therefore, it is very important to evaluate classification models with a large number of samples to estimate their generalizability with high confidence.

IV. CONCLUSION

The findings of this study suggest that high-performance diagnostic models can be built with very few features from sMRI data in a small well-matched dataset at the cost of overfitting. The overfitting stems from both small sample size and feature selection, and is responsible for poor generalizability of the diagnostic model. In contrast, all studies using large heterogeneous autism data (including this study), have reported low classification accuracies. Thus, diagnosis of a highly heterogeneous disorder such as ASD via sMRI is a very difficult, but likely not impossible. Constructing a predictive model utilizing different structural properties of the brain on multiple scales (from voxel and vertex level to global features such as volume) in an ensemble framework would be one of the possible alternatives to explore. If a very large number of samples were available, another very promising technique would be to develop a predictive model from whole brain data utilizing deep learning algorithm. In addition, demographics and behavioral information can be added to augment the predictive power of a diagnosis model for autism.

REFERENCES

[1] E. D. Bigler, S. Mortensen, E. S. Neeley, S. Ozonoff, L. Krasny, M. Johnson, J. Lu, S. L. Provencal, W. McMahon, and J. E. Lainhart, "Superior temporal gyrus, language function, and autism.," *Dev. Neuropsychol.*, vol. 31, no. 2, pp. 217–38, Jan. 2007.

[2] C. Ecker, V. Rocha-Rego, P. Johnston, J. Mourao-Miranda, A. Marquand, E. M. Daly, M. J. Brammer, C. Murphy, and D. G. Murphy, "Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach," *Neuroimage*, vol. 49, no. 1, pp. 44–56, Jan. 2010.

[3] Y. Jiao, R. Chen, X. Ke, K. Chu, Z. Lu, and E. H. Herskovits, "Predictive models of autism spectrum disorder based on brain regional cortical thickness," *Neuroimage*, vol. 50, no. 2, pp. 589–599, 2010.

[4] L. Q. Uddin, V. Menon, C. B. Young, S. Ryali, T. Chen, A. Khouzam, N. J. Minshew, and A. Y. Hardan, "Multivariate searchlight classification of structural magnetic resonance imaging in children and adolescents with autism," *Biol. Psychiatry*, vol. 70, pp. 833–841, 2011.

[5] C.-Y. Wee, L. Wang, F. Shi, P.-T. Yap, and D. Shen, "Diagnosis of autism spectrum disorders using regional and interregional morphological features.," *Hum. Brain Mapp.*, vol. 35, no. 7, pp. 3414–30, 2014.

[6] S. Haar, S. Berman, M. Behrmann, and I. Dinstein, "Anatomical Abnormalities in Autism?," *Cereb. Cortex*, p. bhu242–, Oct. 2014.

[7] M. R. Sabuncu and E. Konukoglu, "Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study," *Neuroinformatics*, vol. 13, no. 1, pp. 31–46, Jul. 2014.

[8] L. Breiman, "Random Forrest," *Mach. Learn.*, pp. 1–33, 2001.

[9] C. J. C. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[10] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, pp. 1189–1232, 2000.

[11] "http://fcon_1000.projects.nitrc.org/indi/abide/)." .

[12] B. Fischl, "FreeSurfer.," *Neuroimage*, vol. 62, no. 2, pp. 774–81, Aug. 2012.

[13] S. Baron-Cohen, H. a. Ring, E. T. Bullmore, S. Wheelwright, C. Ashwin, and S. C. R. Williams, "The amygdala theory of autism," *Neurosci. Biobehav. Rev.*, vol. 24, no. 3, pp. 355–364, 2000.