


RESEARCH REPORT

Multivariate techniques enable a biochemical classification of children with autism spectrum disorder versus typically-developing peers: A comparison and validation study

Daniel P. Howsmon^{1,2} | Troy Vargason^{2,3} | Robert A. Rubin⁴ | Leanna Delhey^{5,6} | Marie Tippett^{5,6} | Shannon Rose^{5,6} | Sirish C. Bennuri^{5,6} | John C. Slattery⁶ | Stepan Melnyk⁶ | S. Jill James⁶ | Richard E. Frye⁷ | Juergen Hahn^{1,2,3} 

¹Dept. of Chemical & Biological Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180

²Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180

³Dept. of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180

⁴Dept. of Mathematics, Whittier College, Whittier, CA 90602

⁵Arkansas Children's Research Institute, Little Rock, AR 72202

⁶Dept. of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR 72205

⁷Barrow Neurological Institute at Phoenix Children's Hospital, Phoenix, AZ 85013 and University of Arizona College of Medicine, Phoenix, AZ 85004

Correspondence

Juergen Hahn, 110 Eighth St., Rensselaer Polytechnic Institute, CBIS # 4213, Troy, NY 12180. Email:hahnj@rpi.edu.

Present address

Daniel P. Howsmon, Willerson Center for Cardiovascular Modeling and Simulation, The University of Texas at Austin, Austin, TX.

Funding information

National Institutes of Health, Grant/Award Number: 1R01AI110642

Abstract

Autism spectrum disorder (ASD) is a developmental disorder which is currently only diagnosed through behavioral testing. Impaired folate-dependent one carbon metabolism (FOCM) and trans-sulfuration (TS) pathways have been implicated in ASD, and recently a study involving multivariate analysis based upon Fisher Discriminant Analysis returned very promising results for predicting an ASD diagnosis. This article takes another step toward the goal of developing a biochemical diagnostic for ASD by comparing five classification algorithms on existing data of FOCM/TS metabolites, and also validating the classification results with new data from an ASD cohort. The comparison results indicate a high sensitivity and specificity for the original data set and up to a 88% correct classification of the ASD cohort at an expected 5% misclassification rate for typically-developing controls. These results form the foundation for the development of a biochemical test for ASD which promises to aid diagnosis of ASD and provide biochemical understanding of the disease, applicable to at least a subset of the ASD population.

KEYWORDS

autism spectrum disorder, biomarkers, multivariate statistical analysis

1 | INTRODUCTION

Autism Spectrum Disorder (ASD) encompasses a large group of early-onset developmental disorders that are collectively characterized by deficits in social interaction and communication as well as the

expression of restricted, repetitive behaviors and interests.¹ ASD is currently estimated to affect 1 in 68 children,² incurring an annual expenditure of \$268 billion in the United States,³ and this prevalence continues to rise.^{2,4} The median age of diagnosis is still 50 months in the United States,² with some evidence that this age is lowering.⁵ In

© 2018 The Authors. Bioengineering & Translational Medicine is published by Wiley Periodicals, Inc. on behalf of The American Institute of Chemical Engineers. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

the United Kingdom, the average age of diagnosis is estimated to be 55 months with no evidence of decreasing.⁶ Part of this discrepancy can be attributed to the difference in positions on national screening between the two countries: in 2007, the American Academy of Pediatrics in the U.S. called for general screening followed by a comprehensive evaluation for ASD by 24 months^{7,8} while the National Health Service in the U.K. advocates against universal screening.⁹ Camarata¹⁰ indicates that this difference in recommendations largely lies in the reliability of ASD diagnosis at 24 months^{11,12} and that the stimulus behind the case for universal screening lies in early intervention. Numerous studies (e.g., Refs. 13–17) have related improved clinical outcomes to early intervention, providing motivation for diagnosing individuals as early as accurate diagnosis is possible.

The “spectrum” nature of ASD coupled with the rapid, highly variable development processes present early in life elevates the challenges of early diagnosis of ASD. Miller et al.¹⁸ illustrate this point by comparing two hypothetical children:

“An active, verbose child who speaks primarily in stereotyped phrases and is preoccupied with train schedules might be immediately recognized as autistic. Likewise, a child who is nonverbal, does not respond to his name despite normal hearing, and who spins things repetitively might also be immediately recognized as autistic. Both children have underlying impairments in social communication and restricted interests, but the surface presentation is quite different.”

A wealth of psychometric tools available to healthcare professionals aid in diagnosing ASD; however, a biological signature of the disorder promises to lower the age of diagnosis without the challenges associated with a behavioral diagnosis of a developmental disorder. Heterogeneity in typical development patterns limits the earliest age at which ASD is reliably diagnosed; prospective studies on social behaviors such as gaze to faces, shared smiles, and vocalizations to others found that these behaviors were not different at 6 months of age, but group differences began to appear at 12 months.¹⁹ However, biomarker signatures of ASD, such as imaging of white matter tract organization²⁰ and EEG complexity,²¹ have been observed as early as 6 months of age. Recent reports from NeuroPointDX suggest amino acid panels are predictive of ASD status in children aged 4–6 years²² and 18–48 months^{23,24} and these promising results suggest extensions down to even younger participants to evaluate the earliest age at which these signatures appear. Such biomarker-based metrics have been shown to aid in the diagnosis of other disorders traditionally solely diagnosed by behavioral observations such as major depressive disorder.²⁵

Biomarkers come with their own set of challenges before they reach clinical translation: less than 0.1% of cancer biomarkers reported in the literature ever enter clinical practice²⁶ and scores of genome wide association studies seeking predictive ASD biomarkers have found few significant findings, most of which are specific to individual studies.²⁷ Reconciling the “holy grail” potential of successful biomarkers

with the poor predictive power among those reported in the literature draws into question the manner in which biomarkers are identified. A better framework is clearly needed for identifying predictive biomarkers that can accurately and differentially diagnose ASD.

Classic biomarker development measures a plethora of candidate biomarkers but evaluates this panel by a series of univariate tests that considers each measurement as independent from all others. Furthermore, the population-level hypothesis tests that are almost synonymous with this approach are ill-suited for quantifying the separation of two or more groups.²⁸ Multivariate biomarkers evaluated by separating individuals (e.g., C-statistic, confusion matrix, etc.) have become increasingly popular since they can incorporate many pieces of information to arrive at a diagnosis. However, since they require more parameters than their univariate counterparts, special attention has to be paid to avoid overfitting when investigating multivariate biomarkers.

The folate-dependent one-carbon metabolism (FOCM) and transsulfuration (TS) pathways comprise a promising source for a multivariate biomarker for ASD. These pathways incorporate both genetic and environmental factors linked to ASD liability.²⁹ The authors have previously developed a multivariate biomarker, based upon Fisher Discriminant Analysis (FDA) for ASD that achieved a positive predictive value and negative predictive value of 97.6 and 96.1%, respectively.²⁹ This investigation will be extended in this article in two different directions: (a) By comparing univariate analysis with four different multivariate methods on FOCM/TS data for ASD biomarker development to ensure that the identified results are not restricted to FDA and (b) to test and validate multivariate FOCM/TS biomarkers on data collected from a new cohort of ASD participants.

2 | MATERIALS AND METHODS

2.1 | Training data

The training data used in this study have been published previously.²⁹ Briefly, data come from the Arkansas Children’s Hospital Research Institute’s autism IMAGE study and detailed study design, inclusion/exclusion criteria, and demographic information have been published elsewhere.³⁰ Children between the ages of 3 and 10 years were recruited locally and enrolled to assess levels of oxidative stress. ASD was assessed by a diagnosis of “Autistic Disorder” as defined in the *Diagnostic and Statistical Manual for Mental Disorders, Fourth Edition*, the Autism Diagnostic Observation Schedule (ADOS), and/or the Childhood Autism Rating Scales (CARS; score > 30); children with other diagnoses on the autism spectrum or rare genetic diseases with similar symptoms to ASD (e.g., fragile X syndrome) were not eligible for participation. TD participants had no medical history of behavioral or neurologic abnormalities by parent report. FOCM/TS metabolites from 83 and 76 case (ASD) and age-matched typically-developing (TD) control children, respectively, were used for classification. The protocol was approved by the Institutional Review Board (IRB) at the University of Arkansas for Medical Sciences and all parents signed informed consent.

2.2 | Validation data

The validation data are taken at baseline from three previously published studies investigating pharmaceutical interventions to normalize metabolic abnormalities of children with ASD³¹: (1) a combination of methylcobalamin and low dose folinic acid^{32,33} (2) high dose folinic acid,³⁴ and (3) sapropterin.³⁵ Given that these studies all focused on evaluating treatment strategies for ASD, all participants had a confirmed diagnosis of ASD. FOCM/TS metabolites were available for 154 (76% male) participants with ASD with a mean age of 8.8 years (range 2–17 years). These ages are different than reported by Delhey et al.³⁴ because this study only required that measurements be available at baseline, rather than both at baseline as well as the conclusion of the treatment phase. Furthermore, stratifying patients by age or gender did not reveal any differences in the univariate metabolite distributions. The first two studies were approved by the IRB at the University of Arkansas for Medical Sciences and the third study was approved by the IRB at the University of Texas Health Science Center at Houston. All parents gave written, signed consent and patients provided assent when appropriate.

2.3 | Metabolites

The metabolites under investigation are presented in Table 1 and additional details of these measurements and derivations are presented in Melnyk, et al.³⁰ This is only a subset of the measurements investigated previously²⁹ because “% DNA methylation” and “8-OHG” were absent from the validation data set and were therefore removed from this study to ensure that a consistent set of metabolites are used for training and testing.

2.4 | Kernel density estimation

Kernel density estimation (KDE) is a nonparametric density estimation technique that overcomes many shortcomings of the common histogram, including discontinuities at bin boundaries, sensitivity with respect to the origin, and zero-valued outside of a certain range.³⁶ In

TABLE 1 Variable identifiers (ID) and names

ID	Variable name	ID	Variable name
x ₁	Methionine	x ₁₂	GSSG
x ₂	SAM	x ₁₃	fGSH/GSSG
x ₃	SAH	x ₁₄	tGSH/GSSG
x ₄	SAM/SAH	x ₁₅	3-CIT
x ₅	Adenosine	x ₁₆	3-NT
x ₆	Homocysteine	x ₁₇	Tyrosine
x ₇	tCysteine	x ₁₈	Tryptophane
x ₈	Glu-Cys	x ₁₉	fCystine
x ₉	Cys-Gly	x ₂₀	fCysteine
x ₁₀	tGSH	x ₂₁	fCystine/fCysteine
x ₁₁	fGSH	x ₂₂	% oxidized glutathione

this work, all KDE procedures use Gaussian kernels. The probability density function (PDF) estimates provided by KDE are then used to evaluate both the C-statistic and misclassification errors at specific one-sided thresholds on the *p* value for membership in the ASD class to characterize the various statistical models described below.

2.5 | Statistical techniques

Multivariate classification for ASD diagnostic status was explored through classification and regression trees, principal component analysis, fisher discriminant analysis, and logistic regression. The presented techniques can be extended to classification tasks with more than two classes; however, only binary classification (i.e., classification into two different groups) will be discussed below. Sample *x* can belong to one of two classes Π_1 and Π_2 .

2.5.1 | Univariate classification

Perhaps the simplest way to develop a classifier for a diagnostic biomarker is to place a simple threshold on a single measurement. For multivariate data, the modeler would then evaluate each measurement independently and choose the measurement with the best discriminating power. In this work, single measurements are mean-centered and normalized to unit variance before estimating the PDFs of the ASD and TD groups.

2.5.2 | Classification and regression trees

When univariate techniques fall short, the modeler must turn to multivariate techniques (i.e., techniques that incorporate multiple features to determine the classification). One intuitive extension from the simple univariate classification scheme is to sequentially place thresholds on many variables in the data set. The most common application of this principle is through recursive partitioning via the classification and regression tree (CART) methodology.³⁷ Since sequential thresholds are placed on variables and multiple thresholds on the same variable are permitted, CART-based classifiers are generally nonlinear.

The tree-growing process begins with a node τ and a node impurity function $i(\tau)$. A proposed split *s* generates two daughter nodes τ_L and τ_R that contain p_L and p_R proportions of the samples in τ . Defining the node impurity function $i(\tau)$ to be the conditional probability that a sample is in Π_1 , the change in impurity is given by

$$\Delta i(s, \tau) = i(\tau) - p_L i(\tau_L) - p_R i(\tau_R) \quad (1)$$

and the split with the greatest reduction in impurity over all variables and all thresholds is chosen. This procedure is repeated for each node until each node contains fewer than some minimum splitting threshold. Each terminal node (i.e., a node with no daughter nodes) is associated with either the ASD or TD class. Next, the tree is pruned upwards by estimating the misclassification rate or risk $R(\tau)$ of the entire tree versus subtrees with one terminal node removed, regularized by the number of terminal nodes T_τ via parameter α :

$$R_\alpha(\tau) = R(\tau) + \alpha |T_\tau| \quad (2)$$

With α chosen via 10-fold cross-validation (see section “Avoiding Overfitting: Cross-validation”), the tree that minimizes $R_\alpha(\tau)$ is chosen

as the final tree. In other words, the tree that balances the misclassification rate with the number of terminal nodes/splits is chosen as the final tree. The classification trees were carried out using the R package "rpart"³⁸ with $i(\tau)$ defined by the Gini index and the minimum splitting threshold set to eight.

2.5.3 | Principal component analysis

Rather than making many sequential decisions through CART methodology, the original data can be projected onto a line and a single binary threshold can be applied to the resulting univariate score. Under the naïve assumption that the most favorable projection for separating the two classes coincides with the projection with maximum variance, principal component analysis (PCA) can be used to establish the projection direction. Principal components are orthogonal and ordered such that the first principal component explains the majority of the total variance.³⁹ In this work, PCA is applied to the panel of metabolites with unknown class membership. Then, the density of the first principal component for each class is estimated via KDE to construct a binary classifier. A threshold is applied to the first principal component at an accepted misclassification rate of the ASD class to develop a simple binary classifier. Although multiple principal components could be used in a classifier, this work focuses on the first principal component to better compare with the other projection-based methods discussed below and to avoid overfitting by introducing more variables. The PCA analysis was conducted in MATLAB.

2.5.4 | Fisher discriminant analysis

Since the data contains class labels (i.e., ASD or TD) for each panel of measurements, a potentially better way to determine the projection direction would directly use the class membership information to maximally separate the distance between the two classes of data. FDA⁴⁰ achieves this by maximizing the ratio of the between-class scatter to the within-class scatter of the two classes. In other words, the mean of the ASD group and that of the TD group are separated as far as possible, while simultaneously minimizing the variance within each group. As in the PCA analysis, the densities of the FDA scores of the two classes are estimated via KDE and a threshold is applied at an accepted misclassification rate of the ASD class to develop a simple binary classifier. The FDA analysis was conducted through routines developed in-house in MATLAB.

2.5.5 | Logistic regression

Using a probabilistic approach, the conditional probability of membership in class i given the data point is given as $p(\Pi_i|x)$. The odds ratio that Π_1 is the correct class is then

$$\frac{p(\Pi_1|x)}{p(\Pi_2|x)} = \frac{p(\Pi_1|x)}{1-p(\Pi_1|x)} \quad (3)$$

Logistic regression (LR) then assumes that the logarithm of this odds ratio can be modeled as a linear function of x

$$L(x) = \log \left(\frac{p(\Pi_1|x)}{1-p(\Pi_1|x)} \right) = w_0 + w^T x \quad (4)$$

where w and w_0 are estimated through maximum likelihood estimation (MLE). Then, the probability distributions can be directly determined as $p(\Pi_1|x) = e^{L(x)} / (1 + e^{L(x)})$ without the need for KDE.

There are many theoretical and experimental studies comparing FDA and LR, resulting in the following outcomes: (a) LR performs better than FDA for non-normal data,⁴¹ (b) LR requires more data to achieve the same asymptotic error rate as achieved by FDA,⁴² though it is possible for LR to achieve its asymptotic error rate with less training data than FDA,⁴³ and (c) MLE of parameters in LR is unstable for separable data, requiring regularization approaches or alternatives to MLE. In practice, these algorithms can be compared on specific data sets to determine the best algorithm for each scenario. LR analysis was conducted using the "glm" function in R.

2.6 | Avoiding overfitting

Multivariate techniques promise more accurate biomarkers as they incorporate many measurements into a diagnosis; however, these techniques can also suffer from overfitting where the model fits the training data exceptionally well but extends poorly to new data sets. Overfitting can occur when some of the included variables are uninformative or correlated with other variables. Therefore, variable selection is employed to choose a minimal set of measurements for use in the classification procedure. In this work, all combinations of variables were evaluated and the combination with the highest C-statistic was chosen for a fixed number of variables. While this presents a large number of variable combinations, this task was completed on a standard laptop computer and runtime was no more than an hour for a fixed number of variables. As the number of variables and/or complexity of the classifier increases, filter methods, such as the report by Li et al.⁴⁴ that used this training data, can be used to separate the variable selection and classification problems for improved computational efficiency. Furthermore, validation strategies such as testing on separate validation data sets and cross-validation are needed to mitigate overfitting and estimate the model's predictive power.

2.6.1 | Validation set

A validation set directly estimates the model's predictive power by evaluating the model on data which it has not yet seen (i.e., was not used for training). A validation set can either be collected after the initial model was trained, or more commonly, a single data set is split into training and validation sets at the start of the modeling procedure. Here, a new data set was incorporated to validate previously developed algorithms.

2.6.2 | Cross-validation

Since initial clinical investigations aiming to uncover diagnostic biomarkers usually obtain measurements on a small number of individuals (e.g., less than 100 participants per group), setting aside too large a portion of data for validation does not leave sufficient samples for proper model identification. Therefore, the estimated predictive power is artificially low. To retain as many samples as possible in the training set, k -fold cross-validation can be used where the samples are divided into k groups and one group is left out at a time, leaving the remaining $k-1$

groups available for model training. Then the model is validated on the group that is left out and the procedure is repeated such that every group is successively left out, so every sample is validated in a statistically independent manner. Here, a variation of k -fold cross-validation known as leave-one-out cross-validation is used such that k is equal to the number of samples, and this cross-validation strategy is used to estimate the model's predictive power within the training data set.

2.6.3 | Binary classification for projection-based methods

The PCA, FDA, and LR models all define a projection direction, but classification requires transforming the continuous-valued scores into a single ASD or TD classification. Throughout this work the threshold β is chosen to fix the estimated probability that a TD participant will be misclassified as ASD.

3 | RESULTS

The different classification methods are first compared with regard to their performance on the training data. Variable selection is used in some cases to determine final FDA and LR models. Once these final models are identified, the models with the highest classification accuracy are evaluated on the validation data.

3.1 | Univariate modeling

Since most biomarker studies focus solely on univariate methods, each variable is first explored individually on the training data for their potential predictive power. Univariate classifiers were evaluated via the C-statistic and cross-validated misclassification errors at thresholds on the p value for membership in the ASD class. These results point to “% oxidized glutathione” as the variable with the greatest discriminatory power (Table 1; Figure 1a). At a threshold of $\beta=0.10$, this univariate classifier achieves a cross-validated misclassification of 6/76 and 18/83 for the TD and ASD participants, respectively. While not sufficient for a diagnostic biomarker, this univariate classifier was used as a baseline for comparison in the multivariate analyses.

3.2 | Performance of projection methods on the training data

Next, multiple measurements were combined in linear, projection-based classifiers. PCA was employed first as it does not make use of class labels in determining the projection-direction. Using all variables, PCA obtained a C-statistic of 0.9706 on the training data. Furthermore, a binary threshold at $\beta \geq 0.10$ membership in the TD class to be classified as TD results in misclassifying only 6/76 and 3/83 of the TD and ASD participants, respectively (Table 1; Figure 1b). This result indicates that the direction of highest variance in the training data also separates the data fairly well for this problem.

Using the group membership in developing a linear, multivariate classifier through FDA or LR promises to further enhance the separation and these methods provide a more solid statistical background for choosing the projection direction than using the direction obtained

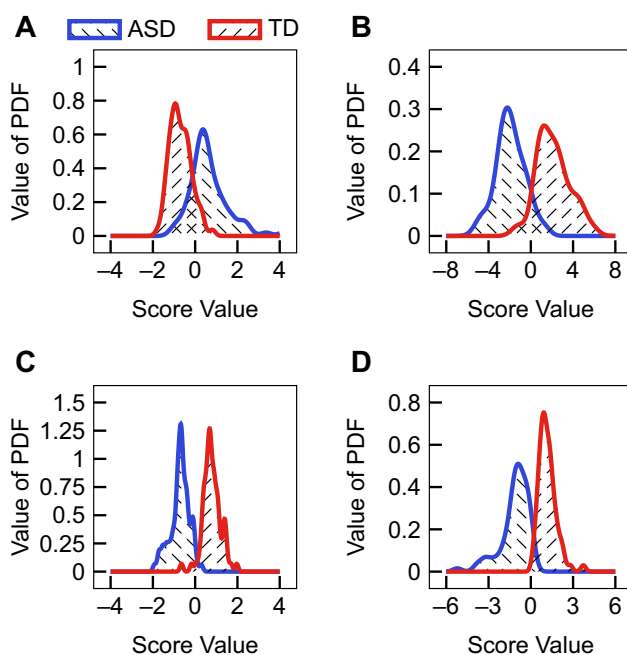


FIGURE 1 Comparison of fitted PDFs for (a) univariate, (b) PCA-all, (c) FDA-all, and (d) LR-all

from PCA for classification. Using all variables, both the FDA and LR models achieve a fitted C-statistic of >0.99 (Table 1; Figure 1c,d). These results suggest that including the group membership in determining the separation direction improves the classification performance, as expected, which is also reflected by the low misclassification numbers of 7/159 and 13/159 at a threshold of $\beta=0.05$ for the FDA-all and LR-all models, respectively.

3.3 | Classification trees

Instead of investigating a single binary threshold on single variables or scores, multiple thresholds on multiple variables were investigated

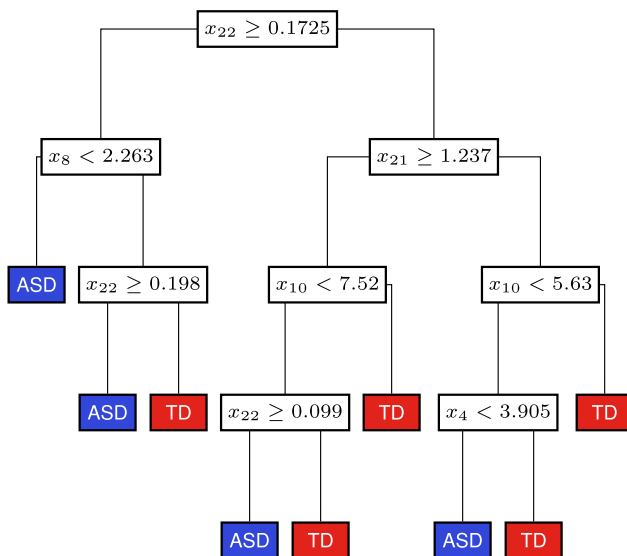


FIGURE 2 Final CART tree fitted to the training data

TABLE 2 Cross-validation comparison of univariate and projection-based multivariate classifiers

Classifier	Variables	Fitted C-statistic	Cross-validated confusion matrix				
			β	TP	FP	FN	TN
Univariate	x_{22}	0.9159	0.01	23	1	60	75
			0.05	47	3	36	73
			0.10	65	6	18	70
			0.20	71	13	12	63
PCA-all	All	0.9706	0.01	56	1	27	75
			0.05	77	3	6	73
			0.10	80	6	3	70
			0.20	80	16	3	60
FDA-all	All	0.9915	0.01	50	1	33	75
			0.05	81	5	2	71
			0.10	81	13	2	63
			0.20	82	19	1	57
FDA-sub	$x_4, x_8, x_{12}, x_{21}, x_{22}$	0.9711	0.01	38	1	45	75
			0.05	78	3	5	73
			0.10	81	8	2	68
			0.20	81	16	2	60
LR-all	All	0.9972	0.01	73	6	10	70
			0.05	78	8	5	68
			0.10	79	11	4	65
			0.20	81	19	2	57
LR-sub	$x_4, x_8, x_{10}, x_{17}, x_{21}, x_{22}$	0.9757	0.01	21	1	62	75
			0.05	79	6	4	70
			0.10	81	8	2	68
			0.20	81	15	2	61

The threshold β is the percent membership in the TD class. TP refers to True Positive, FP to False Positive, FN to False Negative, and TN to True Negative.

using the CART methodology. This simple nonlinear classifier was used to investigate the advantages of moving from linear to nonlinear descriptions of the data. After growing and pruning a classification tree, the resulting model fitted to the training data included five variables (x_{22} = % oxidized, x_8 = Glu-Cys, x_{21} = fCystine/fCysteine, x_{10} = tGSH, and x_4 = SAM/SAH) and this final tree is illustrated in Figure 2. This model achieves a fitted misclassification of 3/76 and 3/83 for TD and ASD participants, respectively. Since the CART methodology can change both the model structure and parameters upon cross-validation, only fitted results are provided for the classification tree.

3.4 | Variable selection for the FDA and LR models

Since previous analysis and the CART results suggest that only a subset of variables is needed to effectively classify the participants into ASD and TD cohorts, variable selection was performed for both the FDA

and LR models. All variables combinations were evaluated for the fitted C-statistic. The C-statistic began to saturate at five variables for the FDA model and at six variables for the LR model. The cross-validated performance of these models with only a subset of the variables (FDA-sub and LR-sub) is provided in Table 2. The fitted C-statistic and cross-validation results are similar for FDA-sub and LR-sub and the variables selected between these two models overlap substantially. These diagnostics indicate that these models are able to find similar patterns in the data that include the most important variables for classification. At a threshold of $\beta=0.05$, using five variables only slightly increased the cross-validated misclassification error from 7/159 in the FDA-all model to 8/159 in the FDA-sub model, indicating that many of the variables provide limited or redundant information for classification of the ASD and TD cohorts. For the LR models at a threshold of $\beta=0.05$, decreasing the number of variables decreased the cross-validated misclassification error from 13/159 to 10/159, providing further evidence that

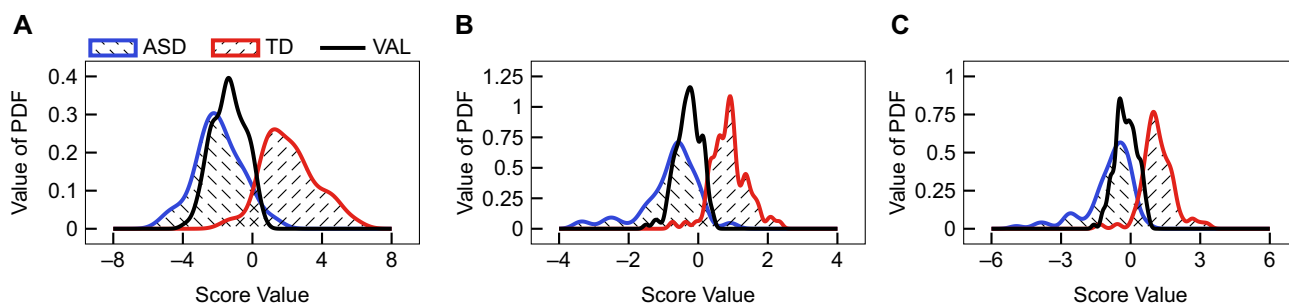


FIGURE 3 Predictions of the ASD validation data from the (a) PCA-all, (b) FDA-sub, and (c) LR-sub models. Note that the validation data, "VAL," only consists of data from a set of children with an ASD diagnosis and therefore significant overlap between VAL and ASD is expected and desired

including too many variables can lead to overfitting of these models to the training data.

3.5 | Validation performance

All previous sections only made use of the training data. Here, the final classifiers were then fitted to the training data set and evaluated on the validation set. It should be noted that the validation set only includes participants with an ASD diagnosis and so the validation can only test for true positive and false negative ASD classifications. Therefore, all classifiers are evaluated at a threshold of $\beta=0.05$ percent membership in the TD class to provide some measure of the expected false positive rate for a given classifier. The PDFs of the predicted ASD class in the validation set are provided in Figure 3. With a threshold of $\beta=0.05$ membership in the TD class to be classified as TD, the predicted binary classification is provided in Table 3. The univariate model produces a high false negative rate of 86/154 at $\beta=0.05$ which can be reduced to 42/154 at $\beta=0.10$. These results provide further evidence that univariate models are insufficient for classifying ASD and TD participants. In contrast, the FDA-sub model produces a false negative rate of only 19/154 at $\beta=0.05$, highlighting the advantages of using multiple FOCM/TS variables in classifying participants as ASD versus TD. All of the linear projection-based methods perform reasonably well

TABLE 3 Validation performance for the final biomarker models

Classifier	Variables	β	TP	FN
Univariate	x_{22}	0.05	68	86
		0.10	112	42
PCA-all	All	0.05	128	26
		0.10	149	5
FDA-sub	$x_4, x_8, x_{12}, x_{21}, x_{22}$	0.05	135	19
		0.10	150	4
LR-sub	$x_4, x_8, x_{10}, x_{17}, x_{21}, x_{22}$	0.05	122	32
		0.10	139	15
CART	$x_4, x_8, x_{10}, x_{21}, x_{22}$	–	116	38

TP refers to True Positive and FN to False Negative. False Positive and True Negative are not needed for the validation set as it only consists of data from a set of children with an ASD diagnosis.

on the validation set, but the FDA-sub model has the lowest validation error. Furthermore, the linear methods are sufficient for separating these two cohorts, supported by the 38/154 false negative rate produced by the CART model.

4 | DISCUSSION

This study compares univariate analysis to four different multivariate statistical techniques through a cross-validation study for the classification of ASD versus TD cohorts. The final models are then fitted to the training data and the resulting models are tested on new validation data, compiled from three previous studies. Since many biomarker studies evaluate each variable individually, the best univariate model was used for comparison. The univariate model using % oxidized indicated a modest separation between the ASD and TD cohorts with a very high false positive rate of 86/154 at an expected false negative rate of 5%. This false positive number reduces to 42/154 if the expected false negative rate is raised to 10%; however, this classification accuracy is not sufficient for developing a diagnostic test.

As an alternative to univariate classifiers, multivariate techniques were introduced to demonstrate the improved performance obtained from simultaneously using data from multiple variables. All multivariate classification techniques performed similarly well on the training data set and they far outperformed the univariate classifier by a wide margin for all four values of the expected false negative rate that were investigated ($\beta = 1\%, 5\%, 10\%, 20\%$). When these classifiers were applied to the validation set then the general trend was upheld, but some differences between the methods emerged: The first principal component of the data was able to sufficiently separate the data (PCA-all produced a false positive rate of 26/154 at an expected false negative rate of 5%), indicating that most of the variation present in the data can be attributed to the differences in the ASD and TD cohorts. The FDA-sub model had the best validation performance with a false positive rate of 19/154 at an expected false negative rate of 5%. As expected, these misclassification rates are higher than the false negative rate of 3.6% at a false positive rate of 2.6% found previously²⁹ due in part to the absence of "% DNA methylation" and "8-OHG" in the validation data, two of the most important variables for separating ASD and TD cohorts.^{29,44} CART and LR of a subset of metabolites produced results

that were slightly worse than PCA with false positive rates of 38/154 and 32/154, respectively; both of these results are at an expected false negative rate of 5%. These results also indicate that linear combinations of variables are sufficient for separation of ASD and TD cohorts using these FOCM/TS metabolites and nonlinear techniques are likely not warranted for these data sets.

When the expected false negative rate was increased to a level of 10% then the true positive predictions of FDA-sub, PCA-all, and LR-sub improved as expected. CART does not allow for a single parameter to tune the classification in a way analogous to the projection-based methods, so the number of true positive rates for CART remains the same. Among these methods FDA-sub performs slightly better than the other techniques on the validation set, reaching true positive predictions as high as 150/154. That being said, a false negative rate of 10% is quite large for a diagnostic test and as such the results for $\beta=0.05$ as discussed in the previous paragraph are more relevant than the results for $\beta=0.1$.

The large number of classification algorithms found in the literature prohibits an exhaustive comparison of all the available techniques. In particular, classification trees are not the only method for nonlinear classification, but they are generally easier to convey to the general scientific audience than the popular kernel methods that can extend PCA, FDA, and LR to nonlinear classification or various neural network strategies. Nonlinear classification techniques are usually applied in domains where data are easier to collect, and data sets are generally larger; therefore, these methods are likely not ideal for early-stage clinical data. Additionally, at least for the data sets under study in this work, we did not find advantages of using nonlinear classification techniques.

Aside from the comparison of the algorithms it is equally important to discuss the finding of which metabolites were identified as contributing the most to the predictive performance. Univariate analysis identified x_{22} , “% oxidized glutathione,” as the most important variable. This variable also appears in the FDA-sub, LR-sub, and CART models further highlighting the importance of the contribution of “% oxidized glutathione” even when multivariate analysis is used. In addition to this variable the subsets chosen for FDA-sub, LR-sub, and CART also have three other variables in common: x_4 : “SAM/SAH,” x_8 : “Glu-Cys,” and x_{21} : “fCystine/fCysteine.” Furthermore, “% oxidized glutathione” highlights the importance of oxidative stress for classification while the “SAM/SAH” ratio is directly linked to DNA methylation and epigenetic components. As suggested previously,^{29,44} it is important to include variables that account for both FOCM (DNA methylation) and TS (oxidative stress) pathways in separating ASD from TD cohorts and this is what the performed analysis returned regardless of which technique was used.

Although the results of this study are promising, there are several limitations that should be considered in future studies.

1. Including the same variables in the training and validation data. Previous analyses found that “% DNA methylation” and “8-OHG” were two of the most important variables for separation,^{29,44} but these data were not present in the validation set. Future studies should include these variables to allow for the highest possible

classification accuracy as these classifiers are considered for clinical translation into a diagnostic test.

2. Including both ASD and TD populations. The validation set included in this study provides a first attempt to validate previous findings with a new data set of similar size, but it only includes ASD participants. While it would be preferable to have a validation set that includes measurements from ASD and TD cohorts, as compared to only from an ASD cohort, these data do not currently exist aside from the one which was used for training here. Future studies should collect additional data and evaluate both ASD and TD populations to confirm separation of these two groups. Finally, the slight improvements in classification obtained by the FDA-sub classifier in comparison with the other methods tested herein should be reaffirmed after evaluating these classifiers on additional TD data.
3. Analyzing younger participants. The training and validation set comprise cohorts of 3–10 years and 2–17 years, respectively. However, a young cohort comprised mainly of participants younger than about 3 years would provide more compelling evidence toward using classifiers such as these to aid in the diagnosis of ASD.

Successfully addressing these limitations would help to solidify the ability of multivariate statistical tools based on FOCM/TS measurements to accurately separate ASD and TD participants and ultimately allow these classifiers to be translated into the clinic.

5 | CONCLUSIONS

This study compared several different classification techniques applied to FOCM/TS metabolites with the purpose of evaluating these metabolites and the analysis techniques as potential biomarkers for ASD. PCA, FDA, LR, and CART models were evaluated in a cross-validation approach on a training data set and then used to predict validation data comprised of 154 new ASD participants and all of these classifiers achieved satisfactory results. An FDA model using five variables was shown to slightly outperform the other models on this new validation data set. While these results look very promising and serve as a partial validation of our previous investigation, future studies should investigate larger cohorts of ASD and TD populations across multiple clinical sites to further support the indication for impaired DNA methylation and increased oxidative stress associated with ASD.

LITERATURE CITED

- [1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, DC: American Psychiatric Association; 2013.
- [2] Christensen DL, Baio J, Braun KVN, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years – autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveill Summ*. 2016;65(3):1, doi: 10.15585/mmwr.ss6503a1

- [3] Leigh JP, Du J. Brief report: forecasting the economic burden of autism in 2015 and 2025 in the United States. *J Autism Dev Disord*. 2015;45(12):4135–4139. doi:10.1007/s10803-015-2521-7
- [4] Braun KVN, Christensen D, Doernberg N, et al. Trends in the prevalence of autism spectrum disorder, cerebral palsy, hearing loss, intellectual disability, and vision impairment, metropolitan Atlanta, 1991–2010. *PLoS One*. 2015;10(4):e0124120. doi:10.1371/journal.pone.0124120
- [5] Christensen DL, Bilder DA, Zahorodny W, et al. Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network. *J Dev Behav Pediatr*. 2016;37(1):1. doi:10.1097/DBP.0000000000000235
- [6] Brett D, Warnell F, McConachie H, Parr JR. Factors affecting age at ASD diagnosis in UK: no evidence that diagnosis age has decreased between 2004 and 2014. *J Autism Dev Disord*. 2016;46(6):1974–1984. doi:10.1007/s10803-016-2716-6
- [7] Johnson CP, Myers SM, the Council on Children With Disabilities. Identification and evaluation of children with autism spectrum disorders. *Pediatrics*. 2007;120(5):1183–1215. doi:10.1542/peds.2007-2361
- [8] CDC | Screening and Diagnosis | Autism Spectrum Disorder (ASD) | NCBDDD. <https://www.cdc.gov/ncbddd/autism/screening.html>. Accessed November 3, 2017.
- [9] Allaby M, Sharma M. Screening for autism spectrum disorders in children below the age of 5 years. A Draft Report for the UK National Screening Committee. Solutions for Public Health, NHS; 2011.
- [10] Camarata S. Early identification and early intervention in autism spectrum disorders: Accurate and effective?. *Int J Speech Lang Pathol*. 2014;16(1):1–10. doi:10.3109/17549507.2013.858773
- [11] Lord C, Risi S, DiLavore PS, Shulman C, Thurm A, Pickles A. Autism from 2 to 9 years of age. *Arch Gen Psychiatry*. 2006;63(6):694–701. doi:10.1001/archpsyc.63.6.694
- [12] Steiner AM, Goldsmith TR, Snow AV, Chawarska K. Practitioner's guide to assessment of autism spectrum disorders in infants and toddlers. *J Autism Dev Disord*. 2012;42(6):1183–1196. doi:10.1007/s10803-011-1376-9
- [13] Zwaigenbaum L, Bryson S, Garon N. Early identification of autism spectrum disorders. *Behav Brain Res*. 2013;251:133–146. doi:10.1016/j.bbr.2013.04.004
- [14] Koegel LK, Singh AK, Koegel RL, Hollingsworth JR, Bradshaw J. Assessing and improving early social engagement in infants. *J Posit Behav Interv*. 2014;16(2):69–80. doi:10.1177/1098300713482977
- [15] MacDonald R, Parry-Cruwys D, Dupere S, Ahearn W. Assessing progress and outcome of early intensive behavioral intervention for toddlers with autism. *Res Dev Disabil*. 2014;35(12):3632–3644. doi:10.1016/j.ridd.2014.08.036
- [16] Estes A, Munson J, Rogers SJ, Greenson J, Winter J, Dawson G. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry*. 2015;54(7):580–587. doi:10.1016/j.jaac.2015.04.005
- [17] Cidav Z, Munson J, Estes A, Dawson G, Rogers S, Mandell D. Cost offset associated with Early Start Denver Model for children with autism. *J Am Acad Child Adolesc Psychiatry*. 2017;56(9):777–783.
- [18] Miller JS, Pandey J, Berry LN. Pediatric screening of autism spectrum disorders. In: Patel VB, Preedy VR, Martin CR, eds. *Comprehensive Guide to Autism*. New York, NY: Springer New York; 2014:311–326. doi:10.1007/978-1-4614-4788-7_13
- [19] Ozonoff S, Iosif A-M, Baguio F, et al. A Prospective study of the emergence of early behavioral signs of autism. *J Am Acad Child Adolesc Psychiatry*. 2010;49(3):256–266.
- [20] Wolff JJ, Gu H, Gerig G, et al. Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. *AJP*. 2012;169(6):589–600. doi:10.1176/appi.ajp.2011.11091447
- [21] Bosl W, Tierney A, Tager-Flusberg H, Nelson C. EEG complexity as a biomarker for autism spectrum disorder risk. *BMC Med*. 2011;9(1):18. doi:10.1186/1741-7015-9-18
- [22] West PR, Amaral DG, Bais P, et al. Metabolomics as a tool for discovery of biomarkers of autism spectrum disorder in the blood plasma of children. *PLoS One*. 2014;9(11):e112445. doi:10.1371/journal.pone.0112445
- [23] Donley E, Burrier R, King J, et al. Prospective validation of molecular subtype diagnostics for autism spectrum and neurodevelopmental disorders (P3.321). Vol. 90 (15 Supplement). Los Angeles, CA: Neurology; 2018.
- [24] West P, Burrier RE, Egnash L, Smith A, Bais P. Biomarkers of autism spectrum disorder. <https://patentimages.storage.googleapis.com/76/c6/60/0e3537d9673f87/US20160169915A1.pdf>. Accessed May 8, 2018.
- [25] Bilello JA, Thurmond LM, Smith KM, et al. MDDScore: confirmation of a blood test to aid in the diagnosis of major depressive disorder. *J Clin Psychiatry*. 2015;76(2):199–206. doi:10.4088/JCP.14m09029
- [26] Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res*. 2012;72(23):6097–6101. doi:10.1158/0008-5472.CAN-12-3232
- [27] Gaugler T, Klei L, Sanders SJ, et al. Most genetic risk for autism resides with common variation. *Nat Genet*. 2014;46(8):881–885. doi:10.1038/ng.3039
- [28] McPartland JC. Considerations in biomarker development for neurodevelopmental disorders. *Curr Opin Neurol*. 2016;29(2):118–122. doi:10.1097/WCO.0000000000000300
- [29] Howsmon DP, Kruger U, Melnyk S, James SJ, Hahn J. Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation. *PLOS Comput Biol*. 2017;13(3):e1005385. doi:10.1371/journal.pcbi.1005385
- [30] Melnyk S, Fuchs GJ, Schulz E, et al. Metabolic imbalance associated with methylation dysregulation and oxidative damage in children with autism. *J Autism Dev Disord*. 2012;42(3):367–377. doi:10.1007/s10803-011-1260-7
- [31] Delhey LM, Tippett M, Rose S, et al. Comparison of treatment for metabolic disorders associated with autism: reanalysis of three clinical trials. *Front Neurosci*. 2018;12:19. doi:10.3389/fnins.2018.00019
- [32] James SJ, Melnyk S, Fuchs G, et al. Efficacy of methylcobalamin and folinic acid treatment on glutathione redox status in children with autism. *Am J Clin Nutr*. 2009;89(1):425–430. doi:10.3945/ajcn.2008.26615
- [33] Frye RE, Melnyk S, Fuchs G, et al. Effectiveness of methylcobalamin and folinic acid treatment on adaptive behavior in children with autistic disorder is related to glutathione redox status. *Autism Res Treat*. 2013;2013:1. e609705. doi:10.1155/2013/609705
- [34] Frye RE, Slattery J, Delhey L, et al. Folinic acid improves verbal communication in children with autism and language impairment: a randomized double-blind placebo-controlled trial. *Mol Psychiatry*. 2018;23(2):247–256. doi:10.1038/mp.2016.168

- [35] Frye RE, DeLaTorre R, Taylor HB, et al. Metabolic effects of saproterin treatment in autism spectrum disorder: a preliminary study. *Transl Psychiatry*. 2013;3(3):e237. doi:10.1038/tp.2013.14
- [36] Izenman AJ. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 1st ed. New York: Springer; 2008.
- [37] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, CA: Wadsworth; 1983.
- [38] Therneau T, Atkinson B, Ripley B. *Rpart: Recursive Partitioning and Regression Trees*; 2018. <https://cran.r-project.org/web/packages/rpart/rpart.pdf>.
- [39] Jolliffe I. Principal component analysis. In: *Wiley StatsRef: Statistics Reference Online*. Wiley; 2014. <http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06472/abstract>. Accessed January 20, 2016.
- [40] Fisher R. The use of multiple measurements in taxonomic problems. *Ann Eugen*. 1936;7(2):179–188.
- [41] Pohar M, Blas M, Turk S. Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski Zv*. 2004;1(1):143.
- [42] Efron B. The efficiency of logistic regression compared to normal discriminant analysis. *J Am Stat Assoc*. 1975;70(352):892–898. doi:10.1080/01621459.1975.10480319
- [43] Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: *Advances in Neural Information Processing Systems*; 2002:841–848.
- [44] Li G, Lee O, Rabitz H. High efficiency classification of children with autism spectrum disorder. *PLoS One*. 2018;13(2):e0192867. doi:10.1371/journal.pone.0192867

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.