Unsupervised Dimensionality Reduction Techniques for the Assessment of ASD Biomarkers

Zachary Jacokes[1], Ian Adoremos[2,3], Arham Rameez Hussain[4], Benjamin T. Newman[4], Kevin A. Pelphrey[5], and John Darrell Van Horn[1,4] for the ACE GENDAAR Consortium

[1]*School of Data Science, University of Virginia*
[2]*College of Computer, Mathematical, and Natural Sciences, University of Maryland*
[3]*The Human Genetics Branch, National Institute of Mental Health*
[4]*Department of Psychology, University of Virginia*
[5]*Department of Neurology, University of Virginia*
Charlottesville, VA 22903, United States of America
Contact Email: jdv7g@virginia.edu

Autism Spectrum Disorder (ASD) encompasses a range of developmental disabilities marked by differences in social functioning, cognition, and behavior. Both genetic and environmental factors are known to contribute to ASD, yet the exact etiological factors remain unclear. Developing integrative models to explore the effects of gene expression on behavioral and cognitive traits attributed to ASD can uncover environmental and genetic interactions. A notable aspect of ASD research is the sex-wise diagnostic disparity: males are diagnosed more frequently than females, which suggests potential sex-specific biological influences. Investigating neuronal microstructure, particularly axonal conduction velocity offers insights into the neural basis of ASD. Developing robust models that evaluate the vast multidimensional datasets generated from genetic and microstructural processing poses significant challenges. Traditional feature selection techniques have limitations; thus, this research aims to integrate principal component analysis (PCA) with supervised machine learning algorithms to navigate the complex data space. By leveraging various neuroimaging techniques and transcriptomics data analysis methods, this methodology builds on traditional implementations of PCA to better contextualize the complex genetic and phenotypic heterogeneity linked to sex differences in ASD and pave the way for tailored interventions.

Keywords: Autism; Neuroimaging; Copy number variation; Gene expression; Conduction velocity

1. Introduction

Autism Spectrum Disorder (ASD) encompasses a broad range of developmental conditions characterized by persistent deficits in social functioning, cognition, and restricted, repetitive behavior[1]. Individuals with ASD often experience challenges in communication, social interactions, and engage in repetitive behaviors or have narrowly focused interests[2]. The prevalence of ASD has been steadily increasing worldwide, affecting between 1 in 36 children and 1 in 45 children according to recent meta-analyses and research by the Centers for Disease Control and Prevention (CDC)[3,4]. This diagnostic increase has brought significant attention to the urgent need for a deeper understanding of the underlying mechanisms of ASD.

Research indicates that ASD is a heterogeneous condition, meaning that it can present very differently from one person to another, complicating efforts to pinpoint its causes[5]. Although it is widely accepted that both genetic and environmental factors contribute to the development of ASD, the exact etiological factors and their interactions remain unclear. Genetic studies have identified numerous genes associated with ASD, suggesting a strong hereditary component[6–8]. However, environmental factors such as prenatal exposure to certain drugs, complications during birth, and advanced parental age have also been identified as potential risk factors for developing ASD[9,10].

Moreover, neuroimaging studies have revealed differences in brain structure and function in individuals with ASD[11–13]. These studies have shown abnormalities in areas of the brain responsible for social behavior, communication, and sensory processing. Despite these advances, there is still much to learn about how these genetic and environmental factors interact to influence brain development and lead to the diverse array of symptoms observed in ASD.

Neuroimaging and genomics exploration is essential for understanding ASD because these approaches provide complementary insights into the biological underpinnings of the condition. Neuroimaging techniques, such as MRI and fMRI, allow researchers to observe structural and functional differences in the brains of individuals with ASD; this imaging data helps to identify patterns and variations in brain development and connectivity that may contribute to ASD symptoms. Concurrently, genomics offers a window into the genetic factors influencing ASD risk, uncovering specific genes and genetic variants associated with the disorder. By integrating genomic information with neuroimaging data, research efforts can better explore how genetic predispositions affect brain structure and function, and vice versa. This combined approach is crucial for elucidating the complex interplay between genetic and neural mechanisms, ultimately enhancing our understanding of ASD and guiding the development of more targeted interventions.

1.1. *Sex-wise disparity in ASD*

A significant aspect of ASD research is the observed sex-wise disparity in its prevalence. Males are diagnosed with ASD more frequently than females, with a ratio of approximately four-to-one[3]. This disparity suggests potential sex-specific biological factors that may influence the development of ASD. Several hypotheses have been proposed to explain this difference, including genetic differences in sex chromosomes, hormonal influences, and differences in brain structure and function between males and females[12,14]. Understanding these sex-specific factors is crucial for developing tailored diagnostic and therapeutic approaches for ASD.

1.2. *Neuronal microstructure analysis in ASD*

Neuroscientific research has increasingly focused on the neuronal microstructure to uncover the subtle differences in brain form and function associated with ASD. Using diffusion MRI, microstructural analysis allows for the examination of small-scale variations in the brain's cellular architecture and can provide insights into the neural underpinnings of ASD. A recently developed microstructural analysis measures axonal conduction velocity, which is derived from parameters such as the g-ratio (the ratio of the inner to the outer diameter of the myelin sheath) and axon diameter[15,16]. Conduction velocity approximates the speed at which action potentials travel along axons, and deviations from the optimal speed can result in impaired neuronal communication.

1.3. *Genetic factors and the pseudo-autosomal region*
Genetic research has identified several candidate genes associated with ASD, many of which are in the pseudo-autosomal regions of the sex chromosomes[17–19]. These regions are of particular interest because they escape the usual X-inactivation process in females, resulting in a unique expression pattern that may contribute to the sex-wise disparity observed in ASD. Exploring these genetic factors, combined with microstructural data, can provide a more comprehensive understanding of the biological basis of ASD.

1.4. *The ACE Network and NDA*
The Autism Centers of Excellence (ACE) program is an initiative funded by the National Institute of Mental Health (NIMH) aimed at advancing the understanding, diagnosis, and treatments of ASD. Established to support large-scale multidisciplinary research projects, the ACE program brings together leading experts from various fields like genetics, neuroimaging, and phenotypic science to foster collaboration. Its structure allows for the integration of novel methodologies and state-of-the-art technologies to ensure that research efforts are at the forefront of scientific discovery. Complementary to the ACE program is the NIMH Data Archive (NDA), a comprehensive database managed by the NIMH that serves to centralize and disseminate the vast array of data collected on mental health research. Together, the ACE program and the NDA create a synergistic environment to nurture and advance the field of ASD research. The ACE program generates rich multimodal datasets that feed into the NDA. By leveraging the comprehensive data available through the NDA, researchers can explore new hypotheses, validate findings, and translate discoveries into clinical applications more effectively.

1.5. *Dimensionality in microstructural analysis*
A significant challenge in the analysis of neuronal microstructure data is the so-called “curse of dimensionality”. Microstructural processing pipelines typically generate data from over 200 distinct brain regions for each individual participant, which when performed on a voxel-wise level results in millions of datapoints for each individual. In our study, which includes 213 participants, this results in a vast multidimensional dataset. Although an N=213 might be considered respectable in human neuroimaging research, the sheer number of predictors poses a challenge for attaining sufficient statistical power, reproducibility, and interpretation. As an addendum to the concept of “big data,” we suggest that researchers consider highly dimensional datasets such as this one as “wide data” that is subject to a different set of equally important challenges.

Traditional approaches to address this issue involve feature selection to reduce the analytic search space. However, such techniques have inherent limitations. Firstly, they rely heavily on domain expertise, which may not always be available or infallible. Secondly, feature selection excludes certain predictors from the analysis before any machine learning algorithms can utilize them, thereby potentially limiting the scope of the analysis. While this approach can be beneficial when domain expertise is available, it can hinder exploratory analyses of new datasets.

1.6. *Multimodal data fusion in health sciences*
The integration of multimodal neuroimaging and genetic data presents a significant opportunity to improve model performance resulting from the synergy of shared and complementary information across modalities. For ASD research, the known genetic and neurological bases provide a strong foundation for exploring the rich multimodal data space afforded by large-scale data repositories like NDA provides for the ACE program. However, emphasizing interpretable methods is of paramount importance if research findings are to be translated into clinical application. It is through this framework this study has sought to provide insights into the multimodal data space generated by combining neuroimaging and genetic features.

1.7. *Novel approach: PCA and machine learning integration*
Our analysis aims to navigate the complex multidimensional space created by combining genetic and microstructural data modalities. To achieve this, we employ a novel implementation of principal component analysis (PCA) to identify unique characteristics of the dataset in an

unsupervised manner. PCA allows us to reduce the dimensionality of the dataset while retaining the within-class variation, thereby addressing the curse of dimensionality without relying on traditional feature selection methods, as well as retaining generalizability to unseen data.

Following the unsupervised feature selection through PCA, we integrate the results into a traditional classification machine learning framework. This approach enables us to leverage the strengths of both unsupervised and supervised learning techniques, providing a more robust analysis of the data. By doing so, we aim to uncover novel insights into the relationship between genetic factors, neuronal microstructure, and ASD.

The integration of advanced neuroimaging techniques and genetic data analysis holds great promise for unraveling the complex etiology of ASD. By addressing the challenges posed by the curse of dimensionality and leveraging advanced analytical methods, we can enhance our understanding of how the neuronal microstructure and genetic factors combine to form the autistic phenotype. This comprehensive approach not only advances our knowledge of ASD but also paves the way for the development of more effective diagnostic and therapeutic strategies tailored to the unique needs of individuals with ASD.

2. Methods

2.1. *Participants*

Participants included 213 (mean age=153.20 [in months], standard dev.=±35.22; age range=96–215; 99 female [46.48%]) volunteers from Wave 1 of an NIH-sponsored Autism Centers for Excellence network. The study sample included 113 autistic individuals (mean age=150.19, standard dev.=±34.56; age range=96–215; 51 female [45.13%]) and 100 non-autistic individuals (mean age=156.60, standard dev.=±35.81; age range=97–215; 48 female [48.00%]). The diagnostic and sex ratios were intended to be balanced. All ACE GENDAAR Wave 1 (9/04/2012-7/31/2022) neuroimaging, phenotypic, and genetic data were collected, processed, and archived on secure local compute servers under the following Internal Review Board (IRB) approvals: USC Approval #HS-13-00668; USC Approval #HS-18-00467; UVA Approval #22078; UVA IRB HSR #21361; GMU #00000169; and UVA #HSR-22-0423. As per the requirements of the US NIMH, de-identified and de-linked copies of all data were regularly submitted to the NDA as part of Collection #2021, where they are freely available for access to approved investigators. Data obtained by subsequent ACE GENDAAR Waves 2 and 3 (ongoing data collection) were not considered in this analysis. Informed consent was obtained from all participants and their legally authorized representatives.

2.2. *Genetic data preparation*

2.2.1. *Analysis of copy number variant densities*

Using Bioconductor R, a karyotype map was created to visualize mutation densities[20]. Statistical differences were assessed between groups to determine mutation loci present in exclusively in ASD females, and vice versa. Loci were systematically compared to the locations of known genes using the UCSC genome browser, along with their exonic sections and prior association with ASD[21]. Copy number variants (CNVs) were identified from a set (N=196) of Manta-annotated variant-call format (VCF) files. The New York Genome Institute preprocessed and designed these files. Manta is a structural variant (SV) calling tool from Chen et al. that utilizes discordant read-pair and split-read evidence to identify various CNVs, including insertions, deletions, translocations, inversions, and tandem duplications[22]. Manta-annotated VCF files for each subject were compared against a Homo sapiens (assembly GRCh38.p14) reference genome, which contains base-pair positions for transcripts, genes, exons, and introns for all 24 chromosomes, including sex-linked chromosomes X and Y.

2.2.2. *Analysis of differential expression and functional enrichment analytics*

Whole blood transcriptome sequencing was performed on 370 individuals. Transcript-level abundances were quantified using Kallisto[23]. Tximport was employed to aggregate these transcript-level abundances into gene-level counts[24]. Differential expression analysis was conducted using the R package DESeq2, facilitating the identification of statistically significant

changes in gene expression across ASD-diagnosed individuals, and were compared across neurotypical cohorts with sex and diagnosis were examined for interaction effects[25].

2.3. *Conduction velocity data preparation*

2.3.1. *Image acquisition*

Diffusion, T1-weighted, and T2-weighted images were acquired from each participant. Diffusion images were acquired with an isotropic voxel size of 2x2x2mm^3, 64 non-colinear gradient directions at b=1000 s/mm^2, and 1 b=0, TR=7300ms, TE=74ms. T1-weighted MPRAGE images with a FOV of 176x256x256 and an isotropic voxel size of 1x1x1mm^3, TE=3.3; T2-weighted images were acquired with a FOV of 128x128x34 with a voxel size of 1.5x1.5x4mm^3, TE=35. All images were preprocessed to correct for common sources of error and bias in accordance with prior published work[11,26]. T1w/T2w ratio was calculated by performing N4-bias correction, rescaling image intensity, then dividing on a voxel-wise basis[27,28]. Diffusion images were analyzed using a single-shell constrained spherical deconvolution (CSD) to obtain 3 tissue CSD (3T-CSD) microstructure compartments (intra- and extra-cellular isotropic signal, and intra-cellular anisotropic signal) and a fixel-based analysis was used to measure axonal fiber density and cross-section on a voxel-wise basis[11,26,29,30]. Despite obtaining multiple microstructure metrics using this methodology, only conduction velocity was examined here.

2.3.2. *Conduction velocity determination*

The aggregate g-ratio was calculated on a voxel-wise basis and was used as Mohammadi & Callaghan suggest; this is displayed in Equation 1[16,31–33]. As a measure of intra-axonal volume, the fiber density cross section was used as the intra-axonal volume fraction (AVF), and as a metric of myelin density, the T1w/T2w ratio was used as the myelin volume fraction (MVF)[34]. Both metrics represent the total sums of each respective compartment across the volume of the voxel and are a volume-based equivalent to the original formulation of g as the ratio of axon diameter (d) to fiber diameter (D).

$$g = \frac{d}{D} = \sqrt{1 - \frac{\mathrm{MVF}}{\mathrm{MVF+AVF}}} \tag{1}$$

Aggregate conduction velocity was calculated based on the calculations of Rushton and Berman et al.; reiterating Rushton's calculation that conduction velocity (θ) is proportional to the length of each fiber segment (l), and that this is roughly proportional to D, which in turn can be defined as the ratio between d and the g-ratio[15,35]. A value proportional to conduction velocity can be calculated using axon diameter and the g-ratio as in equation 2[35]:

$$\theta \propto l \propto Dg\sqrt{-\ln(g)} \propto d\sqrt{-\ln(g)} \tag{2}$$

All imaging metrics, 3T-CSD compartments, T1w/T2w ratio, aggregate g-ratio, and aggregate conduction velocity were averaged across each of 214 ROIs taken from the JHU-ICBM WM atlas (48 ROIs) and the Destrieux Cortical Atlas (164 ROIs)[27,28]. Additionally, two composite ROIs were included, one of all 48 JHU ROIs and one of 150 neocortical regions from the Destrieux Atlas.

2.4. *Initial analysis*

2.4.1. *Data preprocessing*

All conduction velocity and gene expression predictors were included in an initial traditional model for a total of 245 predictors. Participants were removed from the sample if missing either modality. The data was randomly split into training and testing sets, stratified by diagnostic cohort, at a 75-25 ratio. For feature preprocessing, all numeric predictors were normalized; the two

modalities do not occur on the same scale, so in this way we ensured equitable contributions from each in the analysis.

2.4.2. *Principal component analysis*

PCA was performed to reduce the dimensionality of the data. 40 principal components (PCs) were determined to be the maximum number of PCs examined: this number is equal to approximately 25% of the training data points (n=159), and 40 PCs account for approximately 85% of the cumulative explained variance.

2.4.3. *Logistic regression*

Logistic regression modeling for classification was employed to determine how well the PCs separate the two classes. Model complexity was managed by tuning the number of PCs. 10-fold cross-validation was employed to further validate the modeling procedure. The workflow examined a range between one and 40 PCs to identify the number of PCs that maximized the area under the receiver operating characteristic (AUROC) curve, a metric that balances true positive rate against false positive rate. The final model configuration was applied to the entire training data set with the optimal hyperparameters determined by the tuning process. The final model was deployed on the unseen testing dataset, evaluated using both AUROC and accuracy. The results of the training and testing sets for this analysis are displayed in Table 1.

2.5. *Experimental analysis*

2.5.1. *Data preprocessing*

For the second comparative analysis, the existing training data set was split by participant cohort such that all autistic participants comprised one data frame and all non-autistic participants comprised another data frame. All conduction velocity and gene expression predictors were included in each of these data frames (again a total of 245 predictors). All numeric predictors were normalized again for the same reasons outlined above.

2.5.2. *Principal component analyse*s

Separate PCAs were performed on each of the cohort data frames to reduce the dimensionality of the cohort-specific data by exploring the underlying structures. The number of PCs retained were determined independently for each group. First, the number of PCs that account for 70% of the cumulative variance was identified. Then, the number of PCs with a corresponding eigenvalue greater than or equal to one was identified. If these numbers were not equal, the number of retained PCs was decided to be the midpoint between them (rounded down). The results of this process are displayed in scree plots in Figure 1. Consequently, 17 PCs were retained for the autistic cohort and 14 PCs were retained for the non-autistic cohort.

2.5.3. *Feature selection*

Salient features for each group were extracted from the selected PCs systematically using the following procedure. First, the top 25% (75^{th} percentile) of variable loadings (in terms of absolute value) were identified per selected component to focus on those that contributed most to the within-class variance. Then, instances of each of the predictors present in the top 25% were aggregated to identify the unique predictors among and across these PCs, defined as those only appearing once across all selected PCs. This resulted in seven predictors for the autistic group and 29 for the non-autistic group. Finally, four common predictors between the two classes were removed; the remaining 32 predictors were selected for modeling. A full accounting of these predictors is reported in Tables 2 and 3.

2.5.4. *Logistic regression*

Logistic regression modeling for classification was again employed to determine the effectiveness of this dimensionality reduction technique as compared to the traditional method. Predictors for this model included the 36 predictors selected from the procedure above. Model complexity was managed by tuning the number of PCs. 10-fold cross-validation was employed to further validate

the modeling procedure. The workflow examined a range between one and 36 PCs to identify the number of PCs that maximized the AUROC curve. The final model configuration was applied to the entire training data set with the optimal hyperparameters determined by the tuning process. The final model was deployed on the unseen testing dataset, evaluated using both AUROC and accuracy. The results of the training and testing sets for this analysis are displayed in Table 1. All machine learning analyses and plot visualizations were created using the R package TidyModels[36].

3. Results

3.1. *Genetic analysis*

3.1.1. *Sex-wise analysis of CNV densities in autistic participants*

CNVs were detected in 196 Manta-annotated VCF files from the New York Genome Institute. VCF files benchmarked against the reference genome were assessed for sex-wise differences in the pseudo-autosomal region using pairwise *t*-tests; the results were statistically significant (*T*-statistic = -7.21; $p < 0.001$).

3.1.2. *Differential expression analysis*

Differential expression analysis in DESeq2 showed that 3,707 genes exhibited significant differences when sex and diagnosis are considered as interacting factors. Differentially expressed genes showed statistical significance ($p < 0.01$ after false discovery correction) within or near the pseudo-autosomal boundary and the heterochromatic regions of the Y chromosome. Among these, the homologously encoded zinc finger transcription factors ZFX and ZFY emerged as highly significant genes. After adjustment, ZFX and ZFY showed exceptionally low *p*-values.

3.2. *Traditional analysis*

The results of the traditional modeling procedure were as follows. The 10-fold cross validation procedure for tuning the number of principal components showed that the best training AUROC was 0.693 at 12 PCs. The associated training accuracy was 61.267%. For the unseen testing dataset, the AUROC was 0.618, and the accuracy was 57.407%. These values are reported in Table 1; ROC curves are displayed in Figure 2.

3.3. *Experimental analysis*

3.3.1. *Scree plot description*

Scree plots were generated for each of the autistic and non-autistic cohort PCAs. Thresholds were determined based on the intersection of cumulative percent variance explained (greater than 70% were considered) as well as the principal components with eigenvalues greater than or equal to one; the average PC of these two metrics was used as the final threshold. These thresholds are shown in Figure 1; the former is indicated in smaller dashed red lines, the latter is indicated by longer dashed red lines, and the average of these two is also displayed as a solid red line. These

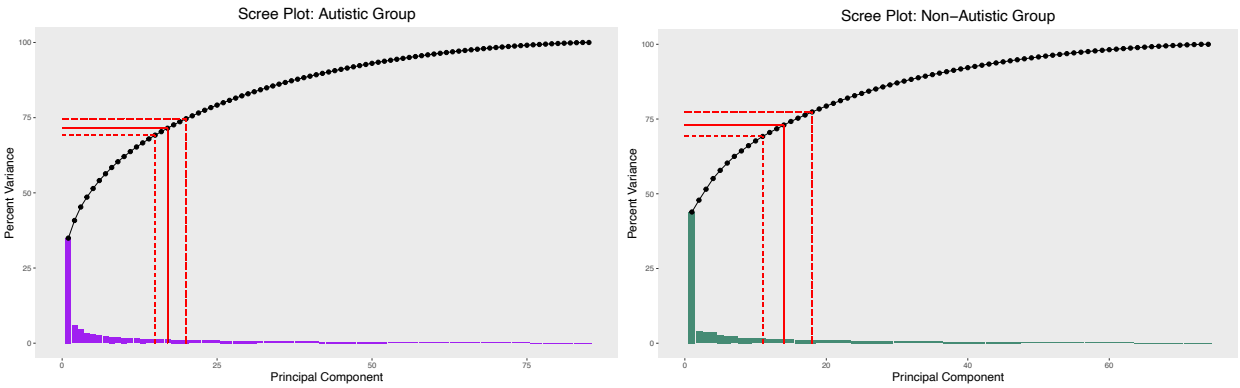

Fig. 1. Scree plots of the autistic and non-autistic cohort PCAs. Short-dashed lines indicate the number of PCs that account for 70% of the cumulative variance, long-dashed lines indicate the number of PCs with eigenvalues greater than or equal to one; the solid red lines indicate the average of these two values, rounded down.

values were as follows: greater than 70% cumulative variance was explained by 15 PCs in the autistic group and 11 PCs in the non-autistic group, eigenvalues greater than or equal to one included 20 PCs in the autistic group and 18 PCs in the non-autistic group, and the final threshold for the autistic group was 17 PCs and 14 PCs for the non-autistic group.

3.3.2. *Model evaluation*

Table 1 contains the logistic regression performance results from the two approaches. The training AUROC and accuracy values were comparable across both approaches, while the testing AUROC of 0.668 was greatly improved in the experimental approach, indicating more robust generalizability. Overall, the accuracy metrics were poor for both models, but an accuracy value of 59.259% for the experimental approach showed improvement over the traditional approach. Visualizations of the AUROC curves are available in Figure 2.

Table 1. Area under ROC curve and accuracy for the traditional model and experimental model.

	Training		Testing	
	AUROC	Accuracy	AUROC	Accuracy
Traditional	0.6931	61.2672%	0.6181	57.4074%
Experimental	0.6935	60.2892%	0.6676	59.2593%

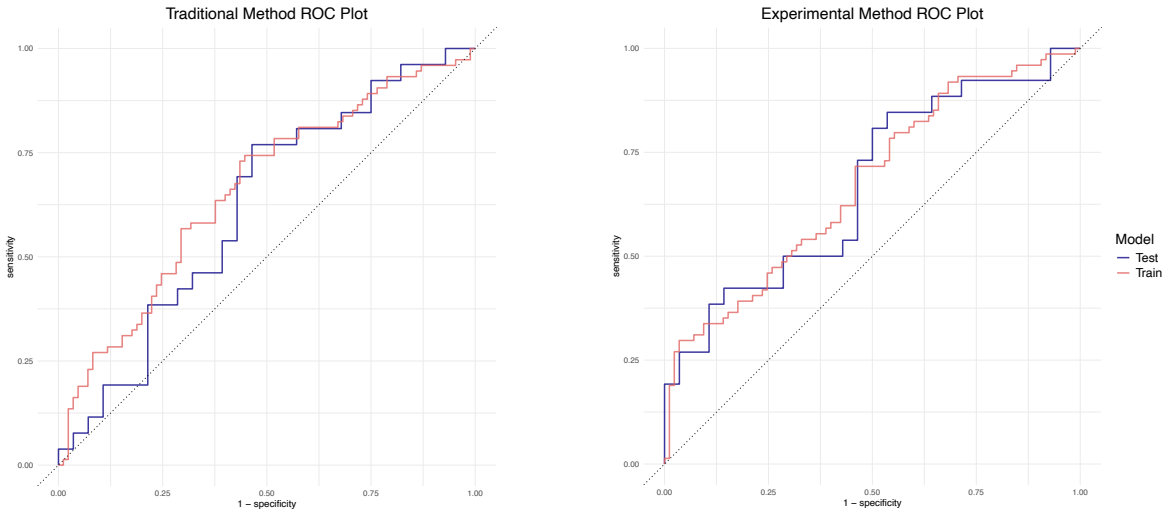

Fig. 2. ROC curves for the traditional logistic regression results (left) and experimental results (right).

3.3.3. *Feature selection*

Tables 2 and 3 display the features selected by the experimental procedure, ordered by PC number and then loading value. After removing the predictors that appeared in PCA procedure for both the autistic and non-autistic group, the experimental analysis contained 36 predictors. These features were mostly loaded onto the first principal component for each group (25/40; 62.5%). The value reported in the final column of these tables represents the loading value of a given predictor on the PC where higher absolute values represent a stronger relationship between predictor and PC. Directionality is also relevant here: positive values indicate a positive relationship between predictor and PC, whereas negative values indicate the opposite. These values only apply within the context of a given PC and should not be compared across PCs. The relevant cortical, subcortical, and white matter regions can be found highlighted in Figure 3.

Table 2. Top predictors from the autistic cohort PCA, sorted by component number, then loading value within each component.

Predictor	Region type	Hemisphere	Component	Value
Frontal superior gyrus	Gray matter	Right	PC1	0.0990
Lateral fissure (posterior part)	Gray matter	Right	PC1	0.0890
Lateral superior temporal gyrus	Gray matter	Right	PC2	0.1511
Frontal inferior sulcus	Gray matter	Right	PC9	-0.1263
Dorsal posterior cingulate gyrus	Gray matter	Left	PC9	-0.1068

Table 3. Top predictors from the non-autistic cohort PCA, sorted by component number, then loading value with each component.

Predictor	Region Type	Hemisphere	Component	Value
Superior corona radiata	White matter	Right	PC1	0.0898
Body of corpus callosum	White matter	-	PC1	0.0893
Posterior corona radiata	White matter	Right	PC1	0.0890
Anterior corona radiata	White matter	Left	PC1	0.0888
Posterior limb of internal capsule	White matter	Left	PC1	0.0886
Posterior thalamic radiation	White matter	Right	PC1	0.0881
Superior circular sulcus of the insula	Gray matter	Left	PC1	0.0877
Posterior corona radiata	White matter	Left	PC1	0.0875
Mid./posterior cingulate gyrus/sulcus	Gray matter	Right	PC1	0.0864
External capsule	White matter	Right	PC1	0.0859
Genu of corpus callosum	White matter	-	PC1	0.0850
Posterior thalamic radiation	White matter	Left	PC1	0.0847
Caudate	Subcortical	Left	PC1	0.0845
Sub-parietal sulcus	Gray matter	Left	PC1	0.0833
Precuneus gyrus	Gray matter	Left	PC1	0.0814
Superior temporal sulcus	Gray matter	Left	PC2	-0.0744
Superior temporal gyrus (transverse)	Gray matter	Left	PC5	-0.0755
Anterior circular sulcus of the insula	Gray matter	Right	PC5	0.0676
Inferior frontal sulcus	Gray matter	Left	PC7	0.1026
H-shaped orbital sulcus	Gray matter	Left	PC7	0.0806
Superior occipital gyrus	Gray matter	Right	PC8	0.0814
Hippocampus	Subcortical	Left	PC9	-0.1093
Superior temporal gyrus (transverse)	Gray matter	Right	PC10	0.1104
Tapetum	White matter	Left	PC11	-0.1104
Inferior parietal gyrus (supramarginal)	Gray matter	Left	PC12	-0.1337
Paracentral lobule gyrus and sulcus	Gray matter	Right	PC13	0.1560
Transverse temporal sulcus	Gray matter	Left	PC14	-0.1019

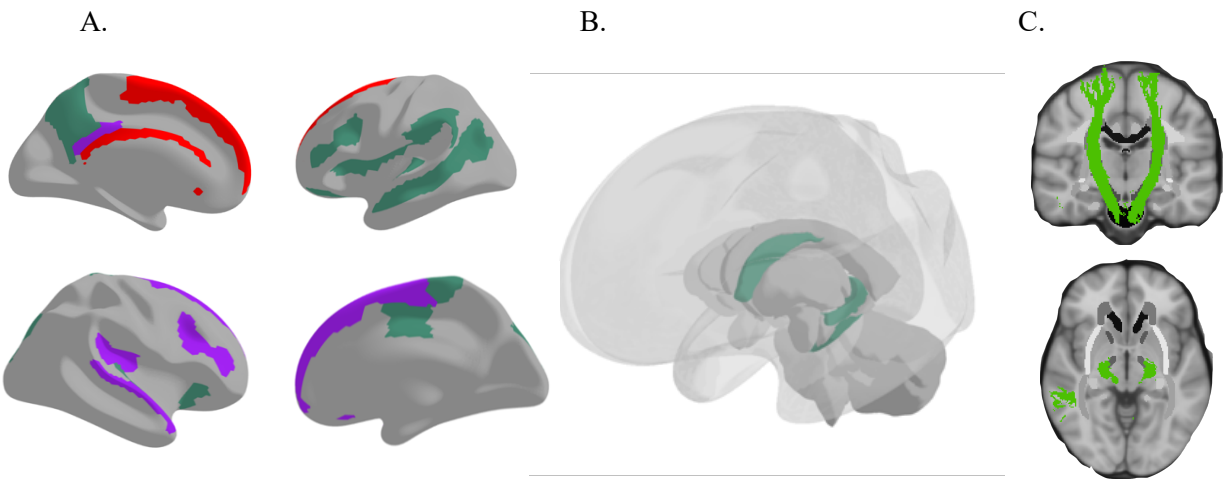

Fig. 3. (A) Cortical regions extracted from the PCA procedure. Top left: medial view of the left hemisphere; top right: lateral view of the left hemisphere; bottom left: lateral view of the right hemisphere; bottom right: medial view of right hemisphere. (B) Subcortical regions extracted from the PCA procedure. (C) White matter tracts extracted from the PCA procedure. Purple regions were found to be characteristic of the autistic group; green regions were found to be characteristic of the non-autistic group; red regions represent the overlapping regions between both the autistic and non-autistic group.

4. Discussion

The results of the experimental dimensionality reduction procedure are promising. In the context of classification for autistic vs. non-autistic individuals using neuroimaging and genetic features, the AUROC performance in this study is acceptable, especially using traditional machine learning frameworks (and not deep learning, which brings its own set of challenges)[37–40]. PCA is effective in this context because it better addresses the issue of overfitting, as evidenced by the improved testing AUROC metric. By capturing within-class variability, the modeling effort performs better on unseen testing data and generalizes more readily to other datasets. Despite failing to achieve performance that could provide actionable clinical insights and true inference of the underlying mechanisms, the feature selection methodology succeeded for multiple reasons.

First, the marked improvement in testing AUROC performance (over the traditional approach) demonstrates that the extracted features capture many of the relevant aspects that differentiate the classes. AUROC is better suited to many classification tasks, including this one, since it provides a balance between true positive rate and false positive rate, whereas accuracy is a simpler metric that measures the ratio of correct predictions to total predictions. AUROC is also the preferred metric for datasets with imbalanced classes; while the classes in this study are not exceptionally imbalanced, AUROC is equipped to handle even slight imbalances and, as such, is the preferred metric here. AUROC improvements in the experimental analysis demonstrate this methodology's internal validity and robustness to variations in unseen testing data.

Additionally, many of the extracted features represent notable regions of cortical, subcortical, and white matter connectivity in ASD research. ASD is characterized by abnormalities in brain structure, function, and connectivity, and many of the established areas of study are present in the extracted features[12,41,42]. The ability of the proposed procedure to pinpoint differences in relevant brain regions validates the methodology and necessitates further exploration both within and without the context of ASD research.

This analysis does not provide much evidence for the role of the pseudo-autosomal region on autism development, as none of the examined genetic predictors outperformed the microstructural predictors in terms of principal component loading. The low N of the sample is

likely a contributing factor to this phenomenon, though it is also possible that the pseudo-autosomal region is not nearly as contributory to the etiology of ASD as microstructural metrics. Indeed, when the two modalities were examined separately, the genetic data performed poorly as predictors for classification within the same framework.

4.1. *Feature selection*

4.1.1. *Cortical features*

Of the many cortical gray matter regions extracted by this methodology, two have been implicated in ASD research previously: the superior occipital gyrus and the frontal superior gyrus[43,44]. The frontal superior gyrus in particular is known to play a role in executive functioning, a domain previously identified as having deficits for autistic individuals relative to non-autistic individuals[45,46]. Other extracted cortical regions not directly implicated in ASD research do pertain to neurological processes relevant to areas previously identified as lacking in ASD individuals, including social cognition (anterior circular sulcus of the insula, inferior parietal gyrus), language processing (inferior frontal sulcus, superior temporal sulcus) and executive function (inferior frontal sulcus)[47–50]. It should be noted that certain cortical regions previously identified as differentially active in autistic and non-autistic individuals were not highlighted by this method, including the dorsal medial frontal cortex, anterior cingulate cortex, and orbitofrontal cortex[51–53].

4.1.2. *Subcortical features*

Subcortical features extracted using this method included the hippocampus and caudate nucleus. The hippocampus is known to be heavily involved in memory-related functions, and specific to ASD, both encoding and retrieval processes of episodic memory have been implicated as altered in ASD[54]. The caudate nucleus has been shown to have decreased connectivity in autistic individuals and is implicated in restricted and repetitive behavior development and increased autistic symptom severity as well[55–57].

4.1.3. *White matter features*

Many of the white matter features extracted in this study are also characteristic of the differences observed between autistic and non-autistic individuals. Corpus callosum tracts are most relevant here (body and genu of corpus callosum, superior/anterior/posterior corona radiata), but the tapetum has also been found to be under-connected in ASD relative to non-autistic individuals[58,59].

4.2. *Alternative approaches*

4.2.1. *PCA procedure on different data frames*

This experimental technique was deployed on this dataset in other ways to assess its effectiveness in different contexts. PCA was performed on each modality without first separating classes to attempt to capture modality-specific variability. Many of the extracted microstructure predictors remained the same as the focus of this study; however, this method also incorporated several genetic predictors as well. The resulting logistic regression yielded poor classification performance, likely due to an inability to extract the most salient features for each class.

Further, separate PCAs were performed on the four groups defined by the two different modalities and the two classes (autistic genetic, autistic microstructure, non-autistic genetic, non-autistic microstructure). Again, the microstructure metrics were comparable to those extracted in the main analysis of the study, and again this method allowed for more genetic predictors to contribute to the machine learning framework. This methodology performed even worse than before, however. The results of both attempts further cements the conclusion that the pseudo-autosomal region does not contribute to differences between autistic and non-autistic participants in this study and it is possible the genetic basis of ASD may lie elsewhere on the genome.

4.2.2. *Other machine learning models*

Two other types of machine learning models were employed for the classification part of this analysis: random forest (RF) and quadratic discriminant analysis (QDA). These models are appropriate for data that is not expected to display a linear decision boundary and as such are more

flexible. Logistic regression does expect the data to be linearly separable, and while that may appear to be a significant limitation of the modeling efforts of this study, RF and QDA performed far worse than logistic regression in both the traditional and experimental dimensionality reduction frameworks. One explanation for this could be that the data is not complete enough to allow for flexible models to generalize well. Microstructure and genetics are only two pieces of a larger puzzle that can include many other modalities like functional imaging, EEG, and behavioral data. Relatedly, while the extracted features comprise the major group differences in this dataset, they only capture part of the global within-group variability and therefore further limit the generalizability of the results; a phenomenon exacerbated by flexible machine learning methods.

4.3. *Future directions*

In the pursuit of assessing putative neurogenetic markers of ASD through the integration of neuroimaging, genomic, and phenotypic data, built upon the approach described here, several critical future directions emerge. One primary consideration is the utilization of data imputation to increase the sample size. While genetic data imputation may not be valid due to the potential introduction of biases and inaccuracies, it can be more appropriately applied to other metrics such as conduction velocity, pending further exploration and validation of the technique in this context.

In terms of machine learning applications, while classification remains a viable approach, regression-based predictive modeling presents an avenue with the potential for more nuanced and informative results. Incorporating behavioral phenotyping outcome surveys, including measures of language, executive function, and social interaction, could provide rich data for these models, enhancing their predictive power and relevance.

An interesting observation from the experimental model is the failure of the gene expression features to contribute significantly following the selection procedure. When modalities were analyzed independently absent the experimental procedure, the resulting classification performance was suboptimal compared to traditional methods. This issue was further compounded when PCA was applied separately to four classes based on diagnostic groups and modalities (e.g., gene expression-autistic, gene expression-non-autistic, etc.). This suggests that the variance captured through the main PCA feature selection approach is sufficient for robust case classification, outperforming more granular feature selection strategies. Some recent studies have attempted to balance modality-specific contributions; these procedures tend to utilize regularization and differential weighting to achieve modality balance and could provide a more nuanced representation of the influence of each modality[60,61].

The feature selection approach could be applicable in individual nuances in autistic individuals; the initial provenance of salient features provides a starting point from which individual similarities and differences can be assessed. Additionally, sex-specific disparities in ASD are another critical area that warrants further examination and could be addressed by an exacting feature selection approach. Conducting separate PCAs for different sexes within the autistic group may reveal unique and actionable insights, potentially improving the performance of downstream machine learning models.

Moreover, several advanced analytical methods offer promising future directions, in particular deep learning**.** Employing deep learning techniques for data fusion to integrate multimodal data could capture complex relationships between neuroimaging, genomic, and phenotypic data. This is an emerging area with promising results but no unified optimal strategy as of yet[62,63].

In summary, future research in the integration of neuroimaging, genomic, and phenotypic data in ASD will need to explore advanced data imputation techniques, leverage regression-based predictive modeling, and consider sex-specific analyses. Employing deep learning, sophisticated weighting and thresholding strategies, and advanced dimensionality reduction methods could significantly enhance the understanding and predictive power of these complex datasets.

4.4. *Conclusions*

The results of the experimental dimensionality reduction procedure for classifying autistic versus non-autistic individuals using neuroimaging and genetic features are promising. The AUROC performance achieved in this study is acceptable, especially within traditional machine learning

frameworks. PCA effectively addresses overfitting, as indicated by the improved testing AUROC metric. By capturing within-class variability, the model performs better on unseen testing data and generalizes more readily to other datasets.

Firstly, the marked improvement in testing AUROC performance over traditional approaches indicates that the extracted features capture many relevant aspects differentiating the classes. AUROC is a balanced metric that accounts for both true positive and false positive rates, making it particularly suitable for datasets with even slight class imbalances. The improvements in AUROC demonstrate the methodology's internal validity and robustness to variations in unseen testing data.

Many of the extracted features represent notable regions of cortical, subcortical, and white matter connectivity, which are well-documented in ASD research. Interestingly, the analysis did not provide substantial evidence for the role of the pseudo-autosomal region in autism development. None of the examined genetic predictors outperformed the microstructural predictors in terms of principal component loading. This may be due to the low sample size, but it also raises the possibility that the pseudo-autosomal region is not as contributory to the etiology of ASD as microstructural metrics. When examined separately, genetic data performed poorly as predictors for classification within the same framework, further supporting this conclusion.

Cortical features extracted from the analysis highlight critical regions involved in ASD, such as areas related to social cognition, language processing, and executive function. These regions are consistent with the existing literature on ASD, reinforcing their importance in understanding the disorder's neurobiological underpinnings. Likewise, subcortical features identified include regions involved in emotion regulation, reward processing, and motor functions. Abnormalities in these areas are frequently reported in ASD studies, underscoring their relevance to the disorder's phenotype and supporting the validity of the feature selection process. Finally, white matter features point to connectivity issues between different brain regions, which are a hallmark of ASD. Disruptions in white matter integrity can affect communication between cortical and subcortical regions, contributing to the diverse symptomatology of ASD.

Applying PCA to each modality without separating classes aimed to capture modality-specific variability. While some microstructure predictors remained consistent, this approach also included several genetic predictors. However, the resulting logistic regression yielded poorer than expected classification performance, likely due to an inability to extract the most salient features for each class. Separate PCAs for the four groups (autistic genetic, autistic microstructure, non-autistic genetic, non-autistic microstructure) also performed poorly, reaffirming that the pseudo-autosomal region may not significantly contribute to ASD classification.

Exploring more flexible machine learning methods, such as quadratic discriminant analysis and tree-based models, did not improve performance over logistic regression. This suggests that the proposed feature selection method is most effective with less flexible machine learning models, highlighting the need for careful selection of analytical techniques based on the data and research goals. The identification of critical cortical, subcortical, and white matter features aligns with existing ASD research, reinforcing their relevance in understanding the disorder's neurobiological underpinnings. While the role of genetic predictors remains less clear, these findings highlight the need for meticulous selection of analytical techniques tailored to the specific characteristics of the data. Such comprehensive and data-driven strategies are vital for understanding the nuances of ASD and advancing for the field toward more effective and personalized diagnostics and interventions.

References

1. Werling DM, Geschwind DH. Sex differences in autism spectrum disorders: *Curr Opin Neurol*. 2013;26(2):146-153. doi:10.1097/WCO.0b013e32835ee548

2. Elsabbagh M, Johnson MH. Infancy and autism: progress, prospects, and challenges. *Prog Brain Res*. 2007;164:355-383. doi:10.1016/S0079-6123(07)64020-5

3. Zeidan J, Fombonne E, Scorah J, et al. Global prevalence of autism: A systematic review update. *Autism Res*. 2022;15(5):778-790. doi:10.1002/aur.2696

4. Maenner MJ, Warren Z, Williams AR, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR Surveill Summ*. 2023;72(2):1-14. doi:10.15585/mmwr.ss7202a1

5. Masi A, DeMayo MM, Glozier N, Guastella AJ. An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neurosci Bull*. 2017;33(2):183-193. doi:10.1007/s12264-017-0100-y

6. Ramaswami G, Geschwind DH. Genetics of autism spectrum disorder. In: *Handbook of Clinical Neurology*. Vol 147. Elsevier; 2018:321-329. doi:10.1016/B978-0-444-63233-3.00021-X

7. Choi L, An JY. Genetic architecture of autism spectrum disorder: Lessons from large-scale genomic studies. *Neurosci Biobehav Rev*. 2021;128:244-257. doi:10.1016/j.neubiorev.2021.06.028

8. Woodbury-Smith M, Scherer SW. Progress in the genetics of autism spectrum disorder. *Dev Med Child Neurol*. 2018;60(5):445-451. doi:10.1111/dmcn.13717

9. Bagasra O, Heggen C, Hossain MI. *Autism and Environmental Factors*. 1st ed. Wiley; 2018. doi:10.1002/9781119042280

10. Landrigan PJ. What causes autism? Exploring the environmental contribution. *Curr Opin Pediatr*. 2010;22(2):219-225. doi:10.1097/MOP.0b013e328336eb9a

11. Newman BT, Jacokes Z, Venkadesh S, et al. Conduction velocity, G-ratio, and extracellular water as microstructural characteristics of autism spectrum disorder. Bray S, ed. *PLOS ONE*. 2024;19(4):e0301964. doi:10.1371/journal.pone.0301964

12. Ilioska I, Oldehinkel M, Llera A, et al. Connectome-wide Mega-analysis Reveals Robust Patterns of Atypical Functional Connectivity in Autism. *Biol Psychiatry*. 2023;94(1):29-39. doi:10.1016/j.biopsych.2022.12.018

13. Yoon N, Huh Y, Lee H, et al. Alterations in Social Brain Network Topology at Rest in Children With Autism Spectrum Disorder. *Psychiatry Investig*. 2022;19(12):1055-1068. doi:10.30773/pi.2022.0174

14. Gata-Garcia A, Porat A, Brimberg L, Volpe BT, Huerta PT, Diamond B. Contributions of Sex Chromosomes and Gonadal Hormones to the Male Bias in a Maternal Antibody-Induced Model of Autism Spectrum Disorder. *Front Neurol*. 2021;12:721108. doi:10.3389/fneur.2021.721108

15. Rushton W a. H. A theory of the effects of fibre size in medullated nerve. *J Physiol*. 1951;115(1):101-122. doi:10.1113/jphysiol.1951.sp004655

16. Mohammadi S, Callaghan MF. Towards in vivo g-ratio mapping using MRI: Unifying myelin and diffusion imaging. *J Neurosci Methods*. 2021;348:108990. doi:10.1016/j.jneumeth.2020.108990

17. Maxeiner S, Benseler F, Krasteva-Christ G, Brose N, Südhof TC. Evolution of the Autism-Associated Neuroligin-4 Gene Reveals Broad Erosion of Pseudoautosomal Regions in Rodents. Nowick K, ed. *Mol Biol Evol*. 2020;37(5):1243-1258. doi:10.1093/molbev/msaa014

18. Mclellan A, Wynne F, Ball M, Moore T. *Sexual Antagonism and Autism Susceptibility in the Xq/Yq Pseudoautosomal Region (PAR2).*; 2007.

19. Wang S, Wang B, Drury V, et al. Rare X-linked variants carry predominantly male risk in autism, Tourette syndrome, and ADHD. *Nat Commun*. 2023;14(1):8077. doi:10.1038/s41467-023-43776-0

20. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80

21. Kent WJ, Sugnet CW, Furey TS, et al. The Human Genome Browser at UCSC. *Genome Res*. 2002;12(6):996-1006. doi:10.1101/gr.229102

22. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222. doi:10.1093/bioinformatics/btv710

23. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

24. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*. 2016;4:1521. doi:10.12688/f1000research.7563.2

25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

26. Newman BT, Dhollander T, Reynier KA, Panzer MB, Druzgal TJ. Test-retest reliability and long-term stability of three-tissue constrained spherical deconvolution methods for analyzing diffusion MRI data. *Magn Reson Med*. 2020;84(4):2161-2173. doi:10.1002/mrm.28242

27. Mori S, Crain BJ. *MRI Atlas of Human White Matter*. Elsevier; 2006.

28. Destrieux C, Fischl B, Dale A, Halgren E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*. 2010;53(1):1-15. doi:10.1016/j.neuroimage.2010.06.010

29. Raffelt DA, Tournier JD, Smith RE, et al. Investigating white matter fibre density and morphology using fixel-based analysis. *NeuroImage*. 2017;144(Pt A):58-73. doi:10.1016/j.neuroimage.2016.09.029

30. Newman BT, Patrie JT, Druzgal TJ. An intracellular isotropic diffusion signal is positively associated with pubertal development in white matter. *Dev Cogn Neurosci*. 2023;63:101301. doi:10.1016/j.dcn.2023.101301

31. Campbell JSW, Leppert IR, Narayanan S, et al. Promise and pitfalls of g-ratio estimation with MRI. *NeuroImage*. 2018;182:80-96. doi:10.1016/j.neuroimage.2017.08.038

32. Stikov N, Perry LM, Mezer A, et al. Bound pool fractions complement diffusion measures to describe white matter micro and macrostructure. *NeuroImage*. 2011;54(2):1112-1121. doi:10.1016/j.neuroimage.2010.08.068

33. Stikov N, Campbell JSW, Stroh T, et al. In vivo histology of the myelin g-ratio with magnetic resonance imaging. *NeuroImage*. 2015;118:397-405. doi:10.1016/j.neuroimage.2015.05.023

34. Raffelt D, Tournier JD, Rose S, et al. Apparent Fibre Density: a novel measure for the analysis of diffusion-weighted magnetic resonance images. *NeuroImage*. 2012;59(4):3976-3994. doi:10.1016/j.neuroimage.2011.10.045

35. Berman S, Filo S, Mezer AA. Modeling conduction delays in the corpus callosum using MRI-measured g-ratio. *NeuroImage*. 2019;195:128-139. doi:10.1016/j.neuroimage.2019.03.025

36. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. Published online 2020. https://www.tidymodels.org

37. Reiter MA, Jahedi A, Fredo ARJ, Fishman I, Bailey B, Müller RA. Performance of machine learning classification models of autism using resting-state fMRI is contingent on sample heterogeneity. *Neural Comput Appl*. 2021;33(8):3299-3310. doi:10.1007/s00521-020-05193-y

38. Mellema CJ, Nguyen KP, Treacher A, Montillo A. Reproducible neuroimaging features for diagnosis of autism spectrum disorder with machine learning. *Sci Rep*. 2022;12(1):3057. doi:10.1038/s41598-022-06459-2

39. Santana CP, De Carvalho EA, Rodrigues ID, Bastos GS, De Souza AD, De Brito LL. rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. *Sci Rep*. 2022;12(1):6030. doi:10.1038/s41598-022-09821-6

40. Yassin W, Nakatani H, Zhu Y, et al. Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. *Transl Psychiatry*. 2020;10(1):278. doi:10.1038/s41398-020-00965-5

41. Walsh MJM, Wallace GL, Gallegos SM, Braden BB. Brain-based sex differences in autism spectrum disorder across the lifespan: A systematic review of structural MRI, fMRI, and DTI findings. *NeuroImage Clin*. 2021;31:102719. doi:10.1016/j.nicl.2021.102719

42. Khundrakpam BS, Lewis JD, Kostopoulos P, Carbonell F, Evans AC. Cortical Thickness Abnormalities in Autism Spectrum Disorders Through Late Childhood, Adolescence, and Adulthood: A Large-Scale MRI Study. *Cereb Cortex N Y N 1991*. 2017;27(3):1721-1731. doi:10.1093/cercor/bhx038

43. Arunachalam Chandran V, Pliatsikas C, Neufeld J, et al. Brain structural correlates of autistic traits across the diagnostic divide: A grey matter and white matter microstructure study. *NeuroImage Clin*. 2021;32:102897. doi:10.1016/j.nicl.2021.102897

44. Zhao X, Zhu S, Cao Y, et al. Abnormalities of Gray Matter Volume and Its Correlation with Clinical Symptoms in Adolescents with High-Functioning Autism Spectrum Disorder. *Neuropsychiatr Dis Treat*. 2022;Volume 18:717-730. doi:10.2147/NDT.S349247

45. Ball G, Stokes PR, Rhodes RA, et al. Executive Functions and Prefrontal Cortex: A Matter of Persistence? *Front Syst Neurosci*. 2011;5. doi:10.3389/fnsys.2011.00003

46. Jacokes Z, Jack A, Sullivan CAW, et al. Linear discriminant analysis of phenotypic data for classifying autism spectrum disorder by diagnosis and sex. *Front Neurosci*. 2022;16:1040085. doi:10.3389/fnins.2022.1040085

47. Ruland SH, Palomero-Gallagher N, Hoffstaedter F, Eickhoff SB, Mohlberg H, Amunts K. The inferior frontal sulcus: Cortical segregation, molecular architecture and function. *Cortex*. 2022;153:235-256. doi:10.1016/j.cortex.2022.03.019

48. Wymbs NF, Nebel MB, Ewen JB, Mostofsky SH. Altered Inferior Parietal Functional Connectivity is Correlated with Praxis and Social Skill Performance in Children with Autism Spectrum Disorder. *Cereb Cortex*. 2021;31(5):2639-2652. doi:10.1093/cercor/bhaa380

49. Kortz M, Lillehei K. Insular Cortex. In: *StatPearls [Internet]*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK570606/

50. Beauchamp MS. The social mysteries of the superior temporal sulcus. *Trends Cogn Sci*. 2015;19(9):489-490. doi:10.1016/j.tics.2015.07.002

51. Zoltowski AR, Lyu I, Failla M, et al. Cortical Morphology in Autism: Findings from a Cortical Shape-Adaptive Approach to Local Gyrification Indexing. *Cereb Cortex*. 2021;31(11):5188-5205. doi:10.1093/cercor/bhab151

52. Moradi E, Khundrakpam B, Lewis JD, Evans AC, Tohka J. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *NeuroImage*. 2017;144:128-141. doi:10.1016/j.neuroimage.2016.09.049

53. Di Martino A, Ross K, Uddin LQ, Sklar AB, Castellanos FX, Milham MP. Functional Brain Correlates of Social and Nonsocial Processes in Autism Spectrum Disorders: An Activation Likelihood Estimation Meta-Analysis. *Biol Psychiatry*. 2009;65(1):63-74. doi:10.1016/j.biopsych.2008.09.022

54. Banker SM, Gu X, Schiller D, Foss-Feig JH. Hippocampal contributions to social and cognitive deficits in autism spectrum disorder. *Trends Neurosci*. 2021;44(10):793-807. doi:10.1016/j.tins.2021.08.005

55. Turner KC, Frost L, Linsenbardt D, McIlroy JR, Müller RA. Atypically diffuse functional connectivity between caudate nuclei and cerebral cortex in autism. *Behav Brain Funct*. 2006;2(1):34. doi:10.1186/1744-9081-2-34

56. Qiu T, Chang C, Li Y, et al. Two years changes in the development of caudate nucleus are involved in restricted repetitive behaviors in 2–5-year-old children with autism spectrum disorder. *Dev Cogn Neurosci*. 2016;19:137-143. doi:10.1016/j.dcn.2016.02.010

57. O'Dwyer L, Tanner C, Van Dongen EV, et al. Decreased Left Caudate Volume Is Associated with Increased Severity of Autistic-Like Symptoms in a Cohort of ADHD Patients and Their Unaffected Siblings. Hu VW, ed. *PLOS ONE*. 2016;11(11):e0165620. doi:10.1371/journal.pone.0165620

58. Payabvash S, Palacios EM, Owen JP, et al. White Matter Connectome Edge Density in Children with Autism Spectrum Disorders: Potential Imaging Biomarkers Using Machine-Learning Models. *Brain Connect*. 2019;9(2):209-220. doi:10.1089/brain.2018.0658

59. Gibbard CR, Ren J, Seunarine KK, Clayden JD, Skuse DH, Clark CA. White matter microstructure correlates with autism trait severity in a combined clinical–control sample of high-functioning adults. *NeuroImage Clin*. 2013;3:106-114. doi:10.1016/j.nicl.2013.07.007

60. Sheng J, Xin Y, Zhang Q, Wang L, Yang Z, Yin J. Predictive classification of Alzheimer's disease using brain imaging and genetic data. *Sci Rep*. 2022;12(1):2405. doi:10.1038/s41598-022-06444-9

61. Bi X an, Hu X, Wu H, Wang Y. Multimodal Data Analysis of Alzheimer's Disease Based on Clustering Evolutionary Random Forest. *IEEE J Biomed Health Inform*. 2020;24(10):2973-2983. doi:10.1109/JBHI.2020.2973324

62. Calhoun VD, Sui J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2016;1(3):230-244. doi:10.1016/j.bpsc.2015.12.005

63. Kalamkar S, A. GM. Multimodal image fusion: A systematic review. *Decis Anal J*. 2023;9:100327. doi:10.1016/j.dajour.2023.100327