A Comparative Evaluation of Supervised Machine Learning Models for Causal Gene Recovery in Autism Spectrum Disorder

Charles Jia, MSc.[1, 2]

[1]Department of Epidemiology and Biostatistics, Western University, Ontario, Canada

[2]ICES Western, Ontario, Canada

Abstract

Introduction

Autism Spectrum Disorder is a complex neurodevelopmental condition with a strong polygenetic architecture. Copy number variations across multiple genetic loci have been consistently implicated as a primary mechanism of disease. While machine learning methods have been increasingly applied to genomic data for ASD, systematic benchmarking is necessary to assess their capacity to determine true causal genetic features.

Methods

This case-control study used a synthetic CNV dataset comprising 1,250 cases and 1,250 controls across 250 simulated genomic loci. The dataset was partitioned into training and test sets using stratified random sampling. Four supervised machine learning models were developed and evaluated: penalized logistic regression (PLR), random forest (RF), extreme gradient boosting (XGB), and support vector machines (SVM). Model performance was assessed using area under the receiver operating characteristic curve, sensitivity, specificity and balanced accuracy. Pairwise differences between models were assessed using the DeLong test. Genes determined through feature importance were compared.

Results

XGB achieved the highest ROC-AUC, followed by RF, SVM and PLR. DeLong tests confirmed that XGB significantly outperformed all three competing models. With respect to causal gene analysis, 18 genes appeared consistently among the top 25 ranked features across at least three of the four models. PLR, RF and XGB models demonstrated strong agreement, while SVM was distinctly inconsistent.

Discussion

XGB demonstrated superior predictive performance across all pairwise comparisons. PLR, RF and XGB models converged substantially in their causal gene ranking, supporting the preferential use of these architectures in CNV-based genomic discovery. The divergence of SVM models in feature ranking suggest kernel-based classifiers may be less reliable. These findings highlight the importance of evaluating both predictive accuracy and feature recovery when selecting ML models for genomic applications.

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition primarily affecting brain connectivity, immune function, and gastrointestinal systems, and is characterized by restricted interests, repetitive behaviors, and impaired social communication[1]. The disorder is broadly understood as multifactorial, with both genetic and non-genetic components contributing to disease risk. Although the precise etiology of ASD remains incompletely defined, a strong hereditary component has been consistently demonstrated through familial studies[2]. In particular, alterations within neuronal and axonal developmental pathways appear to play a central role in ASD pathogenesis, driven in part by single-gene mutations and copy number variations (CNVs) across multiple genomic loci.

CNVs, defined as duplications or deletions of genomic segments resulting in deviations from the expected diploid copy number, have emerged as the primary genetic mechanism underlying ASD susceptibility. Genome-wide profiling using next-generation sequencing (NGS) has revealed that rare and recurrent CNVs are significantly enriched in individuals with ASD relative to neurotypical controls, implicating a distributed set of genomic loci rather than a single causal variant[3]. This polygenic architecture, wherein numerous genes of individually modest effect collectively contribute to disease risk, presents a fundamental challenge for conventional statistical approaches and underscores the need for advanced computational methods capable of capturing complex, high-dimensional genetic signals.

Machine learning (ML) has increasingly been integrated with NGS workflows to address this challenge, offering powerful tools for identifying disease-relevant genetic features from large, high-dimensional datasets[4]. However, the inherent black-box nature of many ML algorithms limits their biological interpretability, constraining their utility in genomic discovery contexts where mechanistic understanding is as important as predictive accuracy[5].

In the present study, we address this gap by developing a synthetic CNV benchmarking framework in which a known set of 30 causal genes is embedded within a dataset of 2,500 patients (1,250 cases, 1,250 controls) across 250 genomic features. Four supervised ML models are systematically compared: penalized logistic regression (PLR), random forest (RF), extreme gradient boosting (XGB), and support vector machines (SVM). Each model is evaluated on predictive performance on held-out test data. By dissociating predictive accuracy from biological

interpretability, this study provides a comparative framework for selecting ML models in CNV-based ASD biomarker discovery, with direct implications for the design of future genomic studies in complex neurodevelopmental disorders.

2. MATERIALS AND METHODS

2.1 Study design

This study adopted a case-control methodology, which is consistent with established conventions in genomic associations studies of ASD[6].

2.2 Data sources

Claude.ai was used to simulate a synthetic computationally generated CNV dataset, given difficulties in sourcing from an existing genomic database. The synthetic data generation procedure was designed to reflect the statistical properties of a real NGS-derived CNV dataset, including continuous copy-number ratios reported to two decimal points, a normal diploid baseline of 2.00 copies, and a population-level variation in CNV magnitude consistent with already published ASD genomic studies. Genetic features were labelled GENE_001 to GENE_250, representing a simplified panel of genomic loci.

2.3 Patient population

The synthetic dataset comprised 2,500 patients in total, allocated equally between 1,250 cases and 1,250 controls. This balanced case-control ratio was chosen to ensure equal representation of both classes during model training and to avoid class imbalance artefacts in performance metrics.

2.4 Exposure

The primary exposure of interest was CNV across 250 stimulated genomic loci.

2.5 End points

The primary endpoint of interest is the causal relationship between each of the 250 stimulated genomic loci to the diagnosis of ASD. This endpoint is highly dependent on the secondary endpoint of interest: the comparative predictive performance of the four machine learning models.

2.6 Statistical analysis

Four supervised machine learning models were developed and evaluated. First, **PLR** was implemented with a regularization parameter λ selected through 10-fold cross-validation

optimizing ROC-AUC. Second, a **RF** classifier was trained with hyperparameter optimizations performed through a full Cartesian grid search over candidate values of mtry, minimum node size, sample fraction, and bootstrap replacement strategy. Out-of-bag misclassification rate was the optimization criterion of interest. Third, a **XGB** model was implemented using a binary logistic objective function, with hyperparameter tuning conducted through random search over 20 candidate combinations of learning rate, tree depth, regularization penalties (λ, α), and subsampling fractions, optimized using 5-fold cross-validation log-loss. Finally, a **SVM** model with a radial basis function (RBF) kernel was trained with cost (C) and kernel width (σ) selected through 5-fold cross-validation grid search optimizing ROC-AUC.

The patient population was partitioned into a training set (70%) and a test set (30%). Following the stratified train-test partitioning, CNV features were preprocessed by centering and scaling using parameters derived from the training set only, applied subsequently to the test set to prevent data leakage. A standardized mean difference (SMD) greater than 0.10 or *p*-value less than 0.05 were determined to be statistically significant. All statistical analyses were conducted using RStudio (Posit Software, PBC, Boston, MA).

3. RESULTS

3.1 Patient population

The patient population comprised of 2,500 patients (1,250 cases, 1,250 controls). The dataset was partitioned into training (n = 2,000) and test (n = 500) sets using stratified random sampling. There were no significant differences between any of the demographic characteristics (Table 1). Both patient groups were similarly aged. Patients diagnosed with ASD were an average age of 54.14 years [standard deviation (SD) 11.79] compared to the average of 53.75 years (SD 12.22) in the control group. Patient ethnicity was predominantly European (41.8%), followed by African (20.3%), East Asian (19.4%), South Asian (9.7%) and American (8.7%).

3.2 Comparative predictive performance

All four supervised machine learning models demonstrated above-89 chance classification performance on the held-out test set (Table 2). The XGB model achieved the highest ROC-AUC of 1.000, followed by RF (AUC = 0.9984), SVM (AUC = 0.9750), and PLR (AUC = 0.9591). Interestingly, the XGB model does not perform with the highest accuracy. Instead, the RF model appears to have higher sensitivity, specificity and balanced accuracy compared to the XGB model.

Pairwise DeLong tests revealed that the XGB model significantly outperformed the RF ($p = 0.0438$), the SVM ($p < 0.001$) and the PLR models ($p < 0.001$). The RF model significantly outperformed the SVM ($p < 0.001$) and the PLR models ($p < 0.001$). The SVM model did not outperform the PLR model ($p = 0.0537$).

3.3 Causal relationship of 250 simulated genetic loci

Across all four models, dozens of genes were consistently ranked among the top in terms of feature importance. GENE_058, GENE_236, GENE_232, GENE_010, GENE_016, GENE_003, GENE_009, GENE_216, GENE_197, GENE_246, GENE_100, GENE_037, GENE_149, GENE_162, GENE_129, GENE_086, GENE_173, and GENE_159 appeared three times in the top 25 features determined through permutation importance or average gain. Interestingly, the PLR, RF and XGB models tended to agree with regards to which features appeared in the top 25. SVM tended to completely disagree with the other three models about the top 25 features.

4. DISCUSSION

This study employed a synthetic CNV benchmarking framework to evaluate the comparative predictive ability of four supervised ML models to identify causal predictors amongst 250 genetic loci. Our findings found the XGB model significantly outperformed other models in regards to the ROC-AUC and pairwise DeLong tests. However, this model was lacking compared to the RF model in terms of test sensitivity, specificity and balanced accuracy.

Many different genes were identified several times across the four different models. However, there was a distinct disagreement with the top 25 features in the SVM model compared to the PLR, RF and XGB models.

4.1 Predictive performance

The XGB model achieved the highest ROC-AUC of 1.000, followed by RF (AUC = 0.9984), SVM (AUC = 0.9750), and PLR (AUC = 0.9591). The superior performance of the XGB model is consistent with its established strengths in high-dimensional tabular data, where its gradient boosting framework and built-in regularization allow it to iteratively correct residual errors while controlling for overfitting[7]. Similarly, the strong performance of the RF model reflects the well-documented robustness of ensemble tree techniques in genomic feature spaces[8].

The DeLong tests confirmed that the XGB model significantly outperformed all three competing models, while the RF model significantly outperformed both SVM and PLR. Interestingly, there was not a significant difference in performance between the SVM and PLR models. This suggests comparable discriminative capacity irrespective of their fundamentally different decision boundaries. The modest performance of the PLR model is consistent with known limitations in non-additive CNV settings, even with regularization imposing sparsity on the feature space[9]. It was hypothesized that the SVM model would perform better than was seen. This lower-than-predicted performance may be due to the sensitivity of the RBF kernel to the choice of cost and kernel width hyperparameters, where the optimal margin boundary is difficult to define when the number of features approaches the number of observations as it did in the test set[10].

4.2 Causal relationships between genetic loci and ASD

The central objective of this study was to identify genomic loci that may be causally associated with ASD. Across all four models, a consistent set of genes emerged repeatedly among the top 25 ranked features. These genes are GENE_058, GENE_236, GENE_232, GENE_010, GENE_016, GENE_003, GENE_009, GENE_216, GENE_197, GENE_246, GENE_100, GENE_037, GENE_149, GENE_162, GENE_129, GENE_086, GENE_173, and GENE_159. Each appeared in the top 25 features at least three times out of the four models. The convergence of these results strengthens confidence that these genes represent genuinely recoverable causal signals rather than model-specific artefacts.

A particularly interesting finding was the divergence of the SVM model from the other three models. While the PLR, RF and XGB models showed a substantial agreement in their feature rankings, the SVM identified a distinct set of features. This divergence is unlikely to reflect a superior biological insight on the part of the SVM model, given its lower comparative ROC-AUC relative to the other supervised methods. Instead, it is more plausible that this difference is attributable to its fundamentally different decision-making mechanism: kernel-based classifiers. This finding has practical implications for causal genomic studies, where tree-based models may outperform kernel methods.

4.3 Limitations

Several limitations should be acknowledged in this study. First, while the use of synthetic data enables this interesting experiment in causal genomic study, it may also simplify the genomic architecture of ASD. Real CNV data pertaining to ASD exhibits complex patterns of linkage of equilibrium, population stratification and gene-gene interactions that were not modelled in this framework. Second, the simulated CNV effect size was prompted to be 0.20 units. While this is a subtle effect size, it is also uniform across all 30 causal genes. In reality, causal loci would likely exhibit heterogeneous effect sizes. Third, the ancestry covariates in the present dataset were simulated without association to disease status or CNV distributions. This, unfortunately, does not reflect the population stratification that is a well-recognized confound in real genomic association studies. Future work should evaluate model performance at larger sample sizes and with more complex simulated genetic architectures to better approximate real-world genomic discovery conditions.

5. CONCLUSION

This study presents a systematic benchmarking comparison of four supervised ML models in their capacity to identify causal genetic features from synthetic CNV data. The XGB model significantly outperformed other models in terms of predictive performance. The XGB, RF and PLR models showed strong agreement in ranking the top 25 features, with 18 genes consistently appearing amongst the top 25 features across these three models. The SVM model diverged substantially from the three other supervised learners. These findings support the preferential use of gradient boosting and ensemble tree methods in CNV-based ASD genomic discovery studies. As ML methods continue to mature and real-world ASD genomic datasets grow in scale and resolution, this framework provides a reproducible and extensible foundation for future comparative ML studies.

6. REFERENCES

1. Chow ML, Pramparo T, Winn ME, et al. Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLOS Genetics*. 2012;8(3):e1002592. doi:10.1371/journal.pgen.1002592

2. Eissa N, Al-Houqani M, Sadeq A, Ojha SK, Sasse A, Sadek B. Current Enlightenment About Etiology and Pharmacological Treatment of Autism Spectrum Disorder. *Front Neurosci*. 2018;12. doi:10.3389/fnins.2018.00304

3. Luo R, Sanders SJ, Tian Y, et al. Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders. *The American Journal of Human Genetics*. 2012;91(1):38-55. doi:10.1016/j.ajhg.2012.05.011

4. Hassan M, Awan FM, Naz A, et al. Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. *International Journal of Molecular Sciences*. 2022;23(9):4645. doi:10.3390/ijms23094645

5. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. *arXiv*. Preprint posted online 2017. doi:10.48550/ARXIV.1705.07874

6. Purdue MP, Dutta D, Machiela MJ, et al. Multi-ancestry genome-wide association study of kidney cancer identifies 63 susceptibility regions. *Nat Genet*. 2024;56(5):809-818. doi:10.1038/s41588-024-01725-7

7. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Published online 2016. doi:10.48550/ARXIV.1603.02754

8. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

9. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Royal Statistical Society Journal Series B: Methodological*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

10. Schölkopf B, Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press; 2001. doi:10.7551/mitpress/4175.001.0001

Funding

No funding was used to support this mock study.

Author Contributions

Charles Jia (CJ). Conceptualization: CJ. Data curation: CJ. Data analysis: CJ. Writing – original draft: CJ. Writing – review and editing: CJ.

Data Availability

A .csv formatted version of the data used within this study can be found on the associated github repository linked here.

Disclosures

All authors report no disclosures

Table 1. Baseline characteristics amongst patients diagnosed with or without autism spectrum disorder (n = 2,500)

		Total (n = 2,500 patients)	Controls (n = 1,250)	Cases (n = 1,250)	SMD
Age (mean, SD)		53.95 (12.01)	53.75 (12.22)	54.15 (11.79)	0.034
Sex (n, %)	Female	1222 (48.9)	610 (48.8)	612 (49.0)	0.003
Ethnicity (n, %)	African	508 (20.3)	239 (19.1)	269 (21.5)	0.066
	American	218 (8.7)	110 (8.8)	108 (8.6)	
	East Asian	485 (19.4)	248 (19.8)	237 (19.0)	
	European	1046 (41.8)	525 (42.0)	521 (41.7)	
	South Asian	243 (9.7)	128 (10.2)	115 (9.2)	

Abbreviations: SD = standard deviation; SMD = standardized mean differences.

Table 2. Comparison of the area under curve, sensitivity, specificity and balanced accuracy of four supervised machine learning models: extreme gradient boosting, random forest, support vector machines, and penalized logistic regression.

Model	Area Under Curve	Sensitivity	Specificity	Balanced Accuracy
Extreme Gradient Boosting	1.000	0.9120	0.9093	0.9107
Random Forest	0.9984	0.9867	0.9813	0.9840
Support Vector Machines	0.9750	0.0912	0.9093	0.9107
Penalized Logistic Regression	0.9591	0.8400	0.9547	0.8973