Predictive Modeling of Autism Spectrum Disorder: Socioeconomic, Prenatal, and Environmental Influences

Seyoung Anna Park[*]
The Webb School
Claremont, CA

Abstract

Autism spectrum disorder (ASD) arises from socioeconomic, prenatal, perinatal, and environmental influences. Using data from the 2022–2023 National Survey of Children's Health (N=82,068, ages 2 to 17), we built predictive models with logistic regression, random forest, support vector machines, gradient boosting, and neural networks. Gradient boosting and neural networks achieved the best performance across accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC. Key predictors spanned socioeconomic, prenatal, and environmental domains, highlighting the value of multifactorial modeling for ASD risk prediction.

Keywords: Autism spectrum disorder, machine learning, socioeconomic status, prenatal and perinatal risk factors, environmental exposures, predictive modeling

1 Introduction

1.1 Background

Autism spectrum disorder (ASD) is a neurodevelopmental condition marked by difficulties in social interaction and communication, along with limited interests and repetitive actions [1]. The reported prevalence of ASD has continuously increased over recent decades. For instance, in the United States, the prevalence of ASD was approximately 0.7% (1 in 150 children) in 2000, rising to about 2.8% (1 in 36 children) by 2020 [2]. The etiology of ASD is complex and cannot be explained by a single factor; rather, it is known to involve interactions between genetic and various non-genetic factors [3, 4].

1.2 Literature Review

Recent studies suggest a multifactorial interplay involving parental socioeconomic status (SES), prenatal and perinatal conditions, and environmental exposures [5, 3, 6, 7, 4]. For

[*] seyounganna.park@gmail.com

example, large-scale population-based studies have reported varying relationships between SES and ASD risk: some studies, such as a Taiwanese cohort study, indicated a positive association between higher parental SES and ASD risk, potentially due to differential access to diagnostic services or underlying biological and immunological factors associated with higher SES [8]. Conversely, several U.S.-based population studies found increased ASD prevalence among children from lower-income households and with lower maternal educational attainment, suggesting SES may influence ASD risk through environmental stressors or differential access to healthcare and early interventions [9].

Prenatal and perinatal risk factors have also been robustly documented. Meta-analyses have consistently reported associations between ASD and maternal infections, pre-pregnancy obesity, diabetes, preterm birth, and birth complications, indicating these conditions can disrupt critical neurodevelopmental processes [6, 10]. Furthermore, systematic reviews of environmental epidemiology have estimated that approximately 40–50% of ASD liability is attributable to non-genetic factors, including prenatal and early-life exposure to air pollution, pesticides, and heavy metals, further highlighting the importance of environmental factors in ASD etiology [11].

Previous studies have independently explored socioeconomic, prenatal, perinatal, and environmental factors associated with ASD, but few have comprehensively integrated these domains into a predictive framework. To fill this critical research gap, there is a clear need for an integrated approach that simultaneously considers the interactions among socioeconomic status (SES), prenatal and perinatal conditions, and environmental exposures. Therefore, the primary aim of our study is to develop and validate a comprehensive machine learning-based prognostic model that integrates SES, prenatal and perinatal factors, and environmental exposures, to enhance the prediction and understanding of ASD risk.

1.3 Data Description

This study utilizes child health data from the National Survey of Children's Health (NSCH), a comprehensive annual cross-sectional survey conducted by the Child and Adolescent Health Measurement Initiative (CAHMI) [12]. The NSCH collects nationally representative data covering a wide range of topics, including demographic characteristics, socioeconomic factors, prenatal and perinatal conditions, environmental exposures, family health history, and child health outcomes. For this research, we specifically combined data from the two most recent survey years (2022 and 2023) to enhance statistical power, encompassing responses from a total of 104,995 children aged 0–17 across the United States.

1.3.1 Sample and Survey Design

The NSCH utilizes a stratified, address-based sampling design targeting households across all 50 states and Washington D.C. One child per household is randomly selected, and the survey is completed by a parent or guardian familiar with the child's health history. After excluding children under the age of 2 years, as ASD diagnoses are uncommon below this age, and addressing missing data issues (21.84% of cases), our final analytical sample included 82,068 children aged 2–17 years. Each child's data record includes sampling weights that ensure generalization to the broader U.S. child population.

1.3.2 Variables Collected

The NSCH dataset covers critical variables across multiple domains:

- Socioeconomic Factors: Household income relative to the federal poverty level, parental education, family structure, and health insurance status.
- Prenatal Factors: Maternal smoking, alcohol, and drug use during pregnancy, maternal health conditions (diabetes, hypertension), and prenatal care utilization.
- Perinatal Factors: Birth weight categories, gestational age (preterm status), birth complications, breastfeeding history, and multiple birth status.
- Environmental Factors: Household smoking, vaping, and broader neighborhood conditions (safety, housing quality).
- Familial Factors: Family medical history of ASD or developmental delays, parental mental health, and parental age at child's birth.
- Child Health and ASD Outcome: Doctor-diagnosed ASD status, along with other health conditions (e.g., ADHD, developmental delays).

1.3.3 Data Quality and Preprocessing

To ensure robust analyses, detailed data preprocessing, including handling missing values through listwise deletion for substantial missing data and multivariate imputation by chained equations (MICE) for isolated cases was performed. Categorical variables were encoded numerically using dummy encoding, and continuous variables were standardized (z-scored). Due to class imbalance in ASD outcomes (3.29% prevalence), stratified sampling methods were employed. The cleaned dataset was sorted into a training set (70%, n=57,448) and a testing set (30%, n=24,620), maintaining a consistent prevalence of ASD across subsets (3.29%). Detailed data processing rsteps (initial sample size to final dataset partitioning) are visually summarized in Figure 1.

1.4 Paper Organization

This paper is structured into four primary sections: Introduction, Methods, Results, and Discussion. The Methods section details the predictive modeling frameworks employed, including logistic regression, random forest, support vector machines (SVM), gradient-boosted machines (GBM), and neural network models, alongside feature selection techniques. The Results section presents the predictive performance of each model and identifies significant predictive factors for ASD. Finally, the Discussion Section interprets these key factors in the context of the existing literature, discusses implications, limitations, and suggests directions for future research.

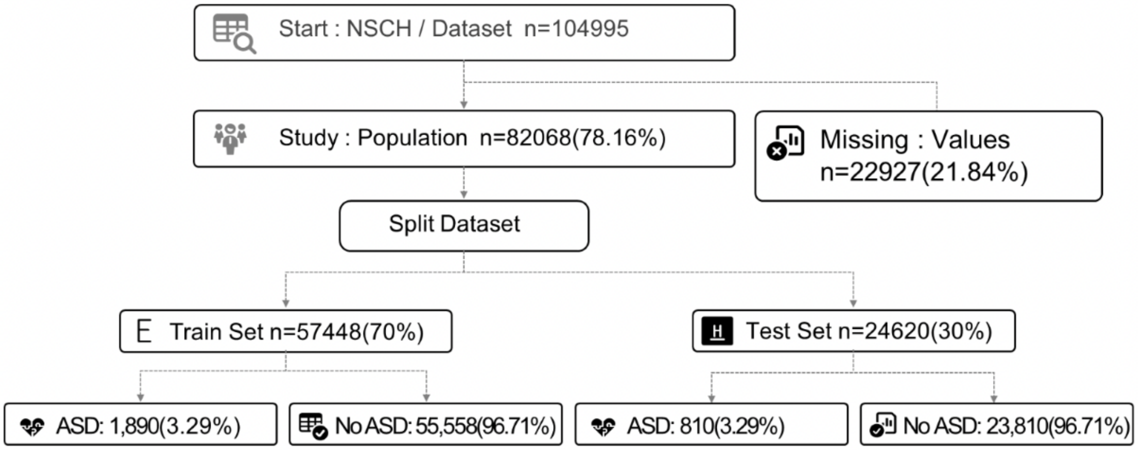

Figure 1: Data Processing Flowchart for NSCH Study

2 Methods

We used the preprocessed analytic dataset from the NSCH 2022–2023 combined files (see Sections 1.3–1.4 for data source, variable definitions, survey design, and initial preprocessing). From the original sample ($\approx$ 100,000 children), observations with extensive missingness in key predictors ($\approx$ 22%) were excluded, yielding a final analytic cohort of 82,068 children. We applied a stratified 70:30 train–test split that preserved ASD prevalence ($\approx$ 3.3%) in both sets; the overall data-processing workflow is summarized in Figure 1.

2.1 Exploratory Analysis and Feature Selection

We conducted descriptive summaries and visualizations; between-group differences (ASD vs. non-ASD) were screened using chi-square tests for categorical variables and t-tests for continuous variables. Given the large sample size, we reported *standardized mean differences* (SMDs) alongside p-values to assess effect sizes. Potential multicollinearity was examined via correlation matrices. Guided by these diagnostics, we performed feature reduction within the modeling pipeline using LASSO and recursive feature elimination (RFE).

2.2 Modeling and Training Procedure

To predict Autism Spectrum Disorder (ASD), this study developed and compared five different machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, and a Neural Network. To ensure a scientifically rigorous and fair comparison, all models were trained and evaluated using a strict **k-fold cross-validation** procedure. This method prevents **information leakage** by ensuring that the model is tested only on data it has never seen before. Each model's hyperparameters were tuned, and the best-performing version was selected based on the Area Under the ROC Curve (AUC) metric [13].

1. Logistic Regression

Logistic Regression is a foundational statistical method used to model the probability of a binary outcome. For this study, it calculates the probability of an ASD diagnosis based on a linear combination of the input variables. The model is defined by the equation:

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp\big(-(\beta_0 + \sum_j \beta_j X_j)\big)}$$

This model served as our **interpretable baseline**. Its primary advantage is transparency; the coefficients (β_j) directly show the strength and direction (positive or negative) of each variable's influence on the outcome. This provided a clear benchmark for comparison with more complex models [14].

2. Random Forest

A Random Forest is a powerful ensemble learning method that operates by constructing a multitude of decision trees. A final prediction is made by taking a majority vote from all the individual trees in the "forest."

$$\hat{y}(x) = \arg\max_c \sum_{b=1}^{B} \mathbf{1}\{T_b(x) = c\}$$

We included this model for its ability to capture **complex, non-linear relationships and interaction effects** between variables that a linear model like logistic regression might miss. It is a robust method that is less prone to overfitting than a single decision tree [15, 13].

3. Support Vector Machine (SVM)

The SVM is a classification algorithm that works by finding the optimal hyperplane that best separates the two data groups (ASD and non-ASD). For complex, non-linear data, it uses the "kernel trick" to map the data to a higher dimension where a linear separation is possible. We used the Radial Basis Function (RBF) kernel:

$$y_i\,(w^\top x_i + b) \geq 1 \quad \text{and} \quad K(x_i, x_j) = \exp\big(-\gamma \|x_i - x_j\|^2\big)$$

The SVM was chosen to provide a **fundamentally different classification approach**. Unlike models based on probability or decision rules, the SVM is based on finding a maximum-margin boundary, offering a unique perspective for our comparative analysis [16, 17].

4. Gradient Boosting

Gradient Boosting is another advanced ensemble technique that builds decision trees **sequentially**. Each new tree is trained to correct the errors made by the previous ones, effectively learning from its mistakes.

$$F_m(x) = F_{m-1}(x) + \eta\, h_m(x)$$

This model was selected for its **high predictive accuracy**. By focusing on residual errors in a step-by-step manner, it can capture very subtle patterns and create a highly effective predictive model, which we hypothesized would perform strongly in this study [18, 13].

5. Neural Network
Inspired by the structure of the human brain, our Neural Network (a multilayer perceptron) consists of interconnected layers of nodes. Each layer transforms the data it receives, allowing the model to learn increasingly abstract and complex patterns.

$$a^{(l)} = f\big(W^{(l)}a^{(l-1)} + b^{(l)}\big)$$

The Neural Network was implemented for its ability to **automatically learn and model high-order interactions** within the data without them being explicitly defined. This makes it a powerful tool for discovering hidden relationships that other models might not identify [19, 20].

2.3 Evaluation Metrics

The primary performance metric was ROC–AUC, interpreted as the probability that a randomly selected ASD case receives a higher predicted risk than a randomly selected non-case (1.0 perfect, 0.5 random) [13]. Secondary metrics—accuracy, sensitivity (recall), specificity, precision, and F1—were computed in a consistent fashion across cross-validation and the held-out test set.

2.4 Variable-importance Integration

To facilitate cross-model interpretation, we normalized model-specific importance summaries (e.g., permutation importance for tree ensembles, absolute standardized coefficients for logistic regression, RFE/SVM rankings) to comparable ranks and aggregated them into a combined heatmap. This procedure emphasizes features consistently prioritized across distinct inductive biases, while deferring all empirical rankings to the Results.

2.5 Class Imbalance Considerations

Given the low ASD prevalence, we used class weighting when available, favored threshold-agnostic discrimination (AUC), and reported thresholded metrics to reflect sensitivity–specificity trade-offs appropriate for potential screening scenarios.

3 Results

Table 1 presents baseline characteristics of the study population, highlighting significant differences between children diagnosed with Autism Spectrum Disorder (ASD) and those without ASD.

Table 1: Baseline Characteristics of Study Population (ASD vs. Non-ASD)

Variable	**Levels**	**No ASD (N=101,009)**	**ASD (N=3,446)**	**p-value**	**SMD**
Age (years)		8.33 ± 5.30	9.98 ± 4.66	<0.001	0.330
Sex (%)	Female	49,555 (49.1)	799 (23.2)	<0.001	0.559
	Male	51,454 (50.9)	2,647 (76.8)		
Birth Weight (grams)		3307.04 ± 544.08	3270.76 ± 599.67	0.001	0.063
Prematurity (%)	No	89,251 (89.3)	2,847 (83.6)	<0.001	0.167
	Yes	10,724 (10.7)	560 (16.4)		
Ever Breast-fed (%)	No	6,015 (15.3)	236 (28.6)	<0.001	0.325
	Yes	33,197 (84.7)	588 (71.4)		
Parental Age at Birth (years)		30.38 ± 5.69	29.90 ± 6.07	<0.001	0.083
Maternal Education (%)	Less than high school	2,663 (2.6)	91 (2.6)	<0.001	0.200
	High school	13,035 (12.9)	554 (16.1)		
	Some college/Assoc.	21,081 (20.9)	942 (27.3)		
	College or higher	64,230 (63.6)	1,859 (53.9)		
Smoke (%)	No	87,079 (88.3)	2,815 (83.1)	<0.001	0.150
	Yes	11,503 (11.7)	573 (16.9)		
Access to Healthcare (%)	No	4,325 (4.3)	81 (2.4)	<0.001	0.108
	Yes	96,209 (95.7)	3,349 (97.6)		
Allergies History (%)	No	74,364 (73.8)	2,112 (61.4)	<0.001	0.267
	Yes	26,437 (26.2)	1,329 (38.6)		
Family Poverty Ratio (%)	0–99%	10,259 (10.2)	497 (14.4)	<0.001	0.236
	100–199%	16,650 (16.5)	768 (22.3)		
	200–399%	36,685 (36.3)	1,196 (34.7)		
	≥400%	37,415 (37.0)	985 (28.6)		
Family History of Mental Disorders (%)	No	75,107 (89.4)	2,114 (74.2)	<0.001	0.402
	Yes	8,909 (10.6)	735 (25.8)		
Prenatal/Perinatal Complications (%)	No	93,478 (92.5)	2,978 (86.4)	<0.001	0.201
	Yes	7,531 (7.5)	468 (13.6)		

SMD = Standardized Mean Difference.
Values are presented as mean ± SD or N (%).

Children with ASD were significantly older on average, with a mean age of 9.98 years (SD ± 4.66), compared to 8.33 years (SD ± 5.30) among non-ASD children ($p < 0.001$). The ASD group predominantly consisted of males, accounting for 76.8% compared to 50.9% in the non-ASD group ($p < 0.001$, Standardized Mean Difference [SMD] = 0.559). Furthermore, the prevalence of breastfeeding was lower among children with ASD (71.4% vs. 84.7%, $p < 0.001$, SMD = 0.325). ASD cases also demonstrated higher rates of prematurity, allergy history, exposure to household smoking, prenatal and perinatal complications, and lower socioeconomic status, indicated by a lower family poverty ratio (all p-values < 0.001, SMD ranging from 0.150 to 0.267).

Figure 2 illustrates a correlation heatmap that reveals important relationships among selected variables. Notably, maternal education exhibited a strong positive correlation with family income, suggesting that higher maternal educational attainment is associated with better economic conditions. Additionally, preterm birth was negatively correlated with birth weight, aligning with established clinical knowledge.

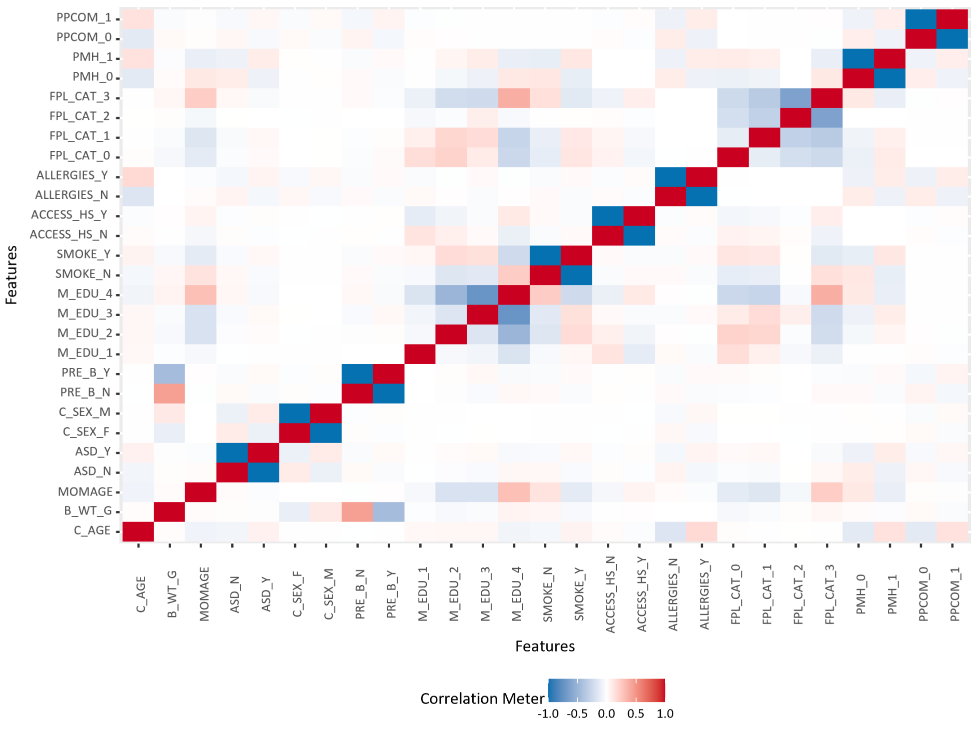

Figure 2: Correlation Heatmap of Variables Associated with Autism Spectrum Disorder (ASD).

Figure 3 illustrates the variable importance rankings generated by the five machine learning models used in this study: Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting Machines (GBM), and Neural Networks. Each model produced slightly different importance scores. Gradient Boosting concentrated on a smaller set of variables with high weight, whereas Logistic Regression and SVM distributed importance more evenly across multiple predictors. To avoid focusing on model-specific charts, the results were further summarized in a consolidated heatmap.

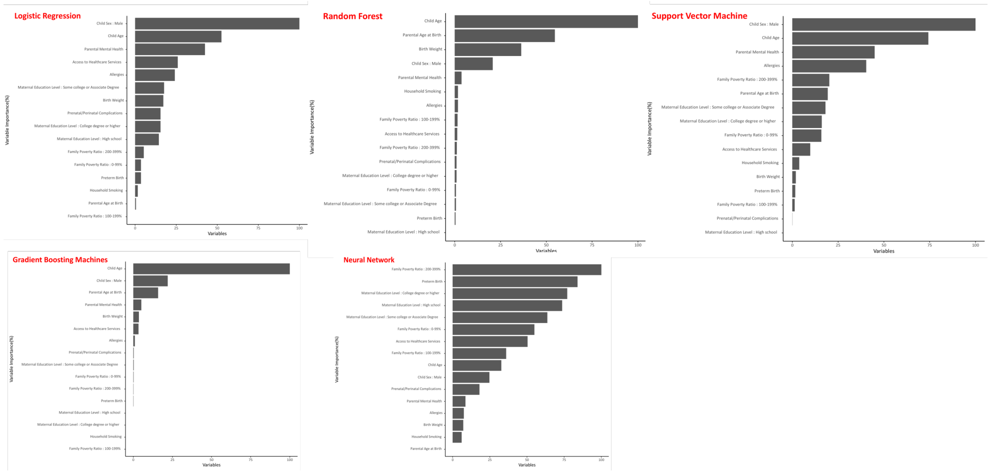

Figure 3: Variable Importance Across Machine Learning Models Predicting Autism Spectrum Disorder (ASD).

Figure 4 synthesizes per-model variable-importance rankings (encoded on a 1–5 scale, with darker red indicating higher rank) across the five learners, enabling direct comparison. Two demographic/biologic features—child sex (male) and child age—recur at or near the top across diverse modeling families (logistic regression, SVM, gradient boosting, and random forest), indicating a stable signal that is robust to differences in functional form. A second tier comprises family and perinatal context: parental age at birth and parental mental health are repeatedly prioritized across multiple models, while prematurity and birth weight contribute more moderately. Together, the heatmap points to a compact set of core predictors that persist despite algorithmic differences, consistent with the study's multifactorial framing.

Model-specific emphases help contextualize this consensus. Tree ensembles (gradient boosting, random forest) concentrate importance on age-related variables, reflecting their capacity to capture thresholds and interactions; linear-margin learners (logistic regression, SVM) foreground sex and age while distributing weight across additional covariates; and the multilayer perceptron elevates socioeconomic context by assigning high rank to maternal education and, to a lesser extent, the family poverty ratio, consistent with nonlinear combinations among SES and clinical history. Because the heatmap encodes *relative* ranks rather than raw scales, these patterns should not be interpreted causally; nevertheless, convergence across orthogonal modeling paradigms strengthens confidence that biologic (sex, age) and family context (parental age, parental mental health), with complementary socioeconomic signals, form the central feature set for ASD risk stratification in this survey-based setting.

Taken together, these results indicate that child sex and child age are consistently strong predictors across all models. In addition, parental age at birth and parental mental health also received high scores, reinforcing their importance. The consistency across diverse models provides robust evidence that both biological characteristics (such as age and sex) and family-related factors (such as parental age and mental health) are critical in predicting ASD risk. Comparing results across multiple models thus allows a more reliable

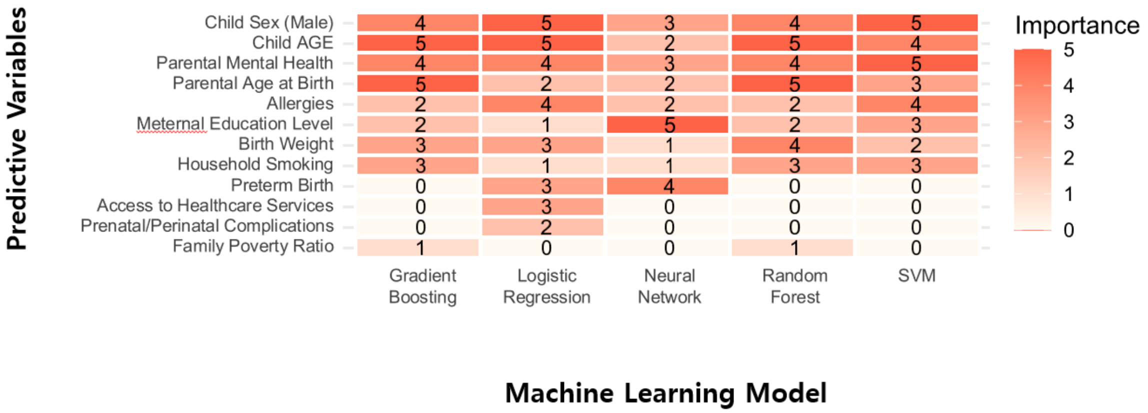

Figure 4: Heatmap of Variable Importance Scores Across Five Machine Learning Models Predicting Autism Spectrum Disorder (ASD).

identification of the factors that truly matter.

Model performance was evaluated using Receiver Operating Characteristic (ROC) curve analysis, shown in Figure 5. All five models performed better than random classification, with Gradient Boosting Machines achieving the highest performance (AUC = 0.986). Logistic Regression and Neural Networks followed closely, while Random Forest exhibited very high sensitivity (0.980) but extremely low specificity (0.048), leading to frequent misclassification of non-ASD cases. Logistic Regression and SVM produced more balanced results, though not as strong overall as Gradient Boosting.

Table 2 reports six performance metrics—accuracy, sensitivity, specificity, precision, F1-score, and ROC–AUC—for all five models. As prespecified, ROC–AUC serves as the primary, threshold-agnostic discrimination metric, while the other metrics describe threshold-level trade-offs for each learner. Across models, precision values are uniformly high (approximately 0.97–0.98), whereas sensitivity and specificity vary more widely, reflecting different operating points.

Gradient boosting shows the strongest discrimination (AUC = 0.986) and a high F1-score (0.919), combining precision 0.974 with sensitivity 0.870. Its accuracy is 0.851, and its specificity (0.310) indicates that, at the evaluated threshold, many non-ASD cases are flagged as positive relative to its recall for ASD cases. These values summarize a model that strongly ranks ASD cases above non-cases while adopting a recall-oriented operating point.

For *random forest*, sensitivity is very high (0.980) but specificity is very low (0.048), yielding nominal accuracy of 0.949 and the *highest* F1-score in Table 2 (0.974). Because F1 aggregates precision and recall only, the combination of high recall and high precision produces a large F1 despite the low specificity. This profile documents a strongly recall-favoring operating point.

The remaining models exhibit more balanced sensitivity–specificity pairs. *Logistic regression* reports AUC = 0.705, sensitivity = 0.641, specificity = 0.683, precision = 0.983,

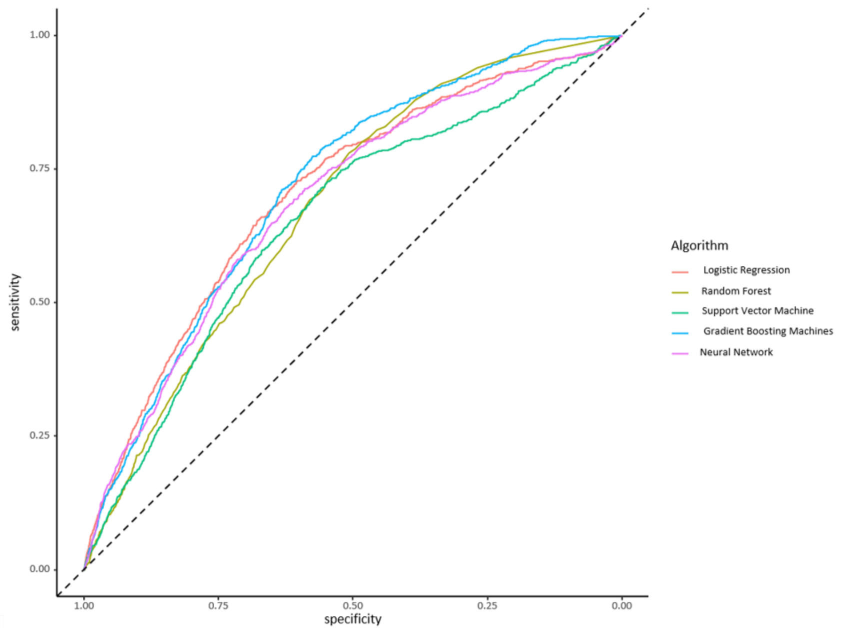

Figure 5: ROC Curves Comparing Machine Learning Models for Predicting Autism Spectrum Disorder (ASD).

F1 = 0.776, and accuracy = 0.643. *SVM* shows AUC = 0.654 with sensitivity = 0.699 and specificity = 0.549 (precision = 0.979, F1 = 0.816, accuracy = 0.695). *MLP* reports AUC = 0.690 with sensitivity = 0.652 and specificity = 0.647 (precision = 0.982, F1 = 0.784, accuracy = 0.652). These summaries indicate that, relative to the tree ensembles, the linear-margin and neural-network models operate closer to a balanced sensitivity–specificity setting at the evaluated threshold.

Overall, Table 2 complements Figure 5 by quantifying threshold-level trade-offs while preserving the AUC-based ranking. In a low-prevalence setting, AUC and F1 are informative summary measures, and the reported sensitivity and specificity clarify the direction of each model's operating point.

Table 2: Performance Metrics of Machine Learning Models for Predicting Autism Spectrum Disorder (ASD).

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
Logistic Regression	0.643	0.641	0.683	0.983	0.776	0.705
Random Forest	0.949	0.980	0.048	0.968	0.974	0.680
SVM	0.695	0.699	0.549	0.979	0.816	0.654
Gradient Boosting	0.851	0.870	0.310	0.974	0.919	0.986
Neural Network	0.652	0.652	0.647	0.982	0.784	0.690

In summary, the results from Figures 3 to 5 and Table 2 demonstrate that the most critical predictors of ASD include child sex (male), child age, parental mental health, parental age at birth, and maternal education level. Among the tested models, Gradient Boosting consistently delivered the highest predictive accuracy, highlighting the potential of advanced machine learning techniques for understanding the interplay of biological and socio-environmental factors in ASD.

4 Discussion

This study built and compared machine learning models for ASD risk using a nationally representative survey dataset that combines demographic, socioeconomic, prenatal/perinatal, environmental, and familial factors. Under a uniform, leakage-safe training and evaluation pipeline, two predictors—child sex (male) and child age—consistently appeared as the strongest signals across models, while family-context variables such as parental mental health and parental age at birth were repeatedly prioritized. Socioeconomic indicators (e.g., maternal education, family poverty ratio) contributed additional information depending on the modeling family, indicating that multiple domains provide complementary predictive value.

Differences in model behavior help explain the patterns in Figure 4. Tree ensembles tended to assign their highest ranks to age-related variables; logistic regression and SVM placed sex and age at the top while distributing weight over several additional covariates; and the multilayer perceptron elevated socioeconomic indicators. Importantly, the heatmap reports *relative* ranks within each model rather than effect sizes, so it is intended for comparing how predictors are prioritized across learners rather than for causal interpretation.

From a practical standpoint, the models operate on survey-like features and can therefore support low-cost, scalable *pre-screening.* Table 2 and Figure 5 show that sensitivity and specificity trade off at the evaluated thresholds, while AUC summarizes threshold-agnostic discrimination. In prospective use, settings that value early capture could favor higher sensitivity, whereas settings that prioritize minimizing false positives could favor higher specificity; selecting an operating point should be based on a bundle of metrics (AUC, F1, sensitivity, specificity) rather than a single value.

This work has several strengths. First, it leverages a large, nationally representative sample, improving generalizability. Second, all five learners were compared within the *same* cross-validated pipeline, with preprocessing and feature selection refit inside folds to minimize leakage and ensure a fair comparison. Third, variable-importance summaries were normalized and aggregated to highlight signals that recur across distinct inductive biases.

Limitations should also be acknowledged. The survey is cross-sectional and based on parent report, which introduces potential misclassification and limits causal inference. ASD prevalence is low ($\sim$3.3%), so class imbalance affects operating characteristics and can yield sensitivity–specificity asymmetries. Validation was internal only, without external validation or calibration. Finally, subgroup performance (e.g., by sex, race/ethnicity, socioeconomic strata) and fairness metrics were not assessed here.

Future work should include external validation on independent NSCH waves or clinical cohorts; probability calibration (e.g., calibration curves, Brier scores); and decision-curve analysis to quantify net benefit in realistic screening workflows. Reporting subgroup performance and fairness metrics would clarify equity considerations. Model explainability (e.g., SHAP) can aid transparency. Pilot implementation in schools or community clinics could help define operating thresholds and referral pathways for a two-step workflow (pre-screen → standardized screening).

In conclusion, readily collected features—including sex, age, parental mental health, parental age at birth, maternal education, and perinatal indicators—carry substantial predictive signal for ASD at the population level. The proposed models are best positioned to *complement* rather than replace existing screening tools, by prioritizing children for standardized screening and specialist evaluation. Progress toward deployment will require external validation, calibration, fairness auditing, and protocol development for threshold selection and clinical integration.

Supplemental Materials

The source code and additional materials used in this study are available at the following GitHub repository:

`https://github.com/Annapark070705/ASD_Prognostic.git`

Acknowledgments

This study was initially presented at the Southern California Applied Statistics Association (SCASA) Poster Competition, where I received first place and was subsequently invited to the SCASA 44th Applied Statistics Workshop. The connections and support gained through this experience were instrumental in the successful completion of this manuscript.

I further acknowledge the SCASA organizing committee for creating this valuable opportunity. In particular, I thank Dr. Rebecca Le for her generous guidance during the poster submission stage, and Dr. Olga Korosteleva for her sustained scientific feedback and advice throughout the manuscript preparation process. Their invaluable contributions and unwavering support were critical to the completion of this work.

I am sincerely grateful to my family for their encouragement and dedication throughout this project, as well as to Autism Tree in San Diego for their ongoing support and inspiration.

References

[1] Al-Dewik, Nada et al. “Overview and Introduction to Autism Spectrum Disorder (ASD)”. In: *Autism Spectrum Disorder*. Cham: Springer, 2020, pp. 3–42. DOI: 10.1007/978-3-030-30402-7_1.

[2] Maenner, Matthew J. 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”. In: *MMWR. Surveillance Summaries* 72.2 (2023), pp. 1–14. DOI: 10.15585/mmwr.ss7202a1.

[3] Bai, D. et al. "Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort". In: *JAMA Psychiatry* 76.10 (2019), pp. 1035–1043. DOI: 10.1001/jamapsychiatry.2019.1411.

[4] Rossignol, Daniel A. et al. "Environmental Toxicants and Autism Spectrum Disorders: A Systematic Review". In: *Translational Psychiatry* 4.2 (2014), e360. DOI: 10.1038/tp.2014.4.

[5] Ashwini et al. "The Attitude Regarding Social Interaction and Communication Problems among Children with Autism Spectrum Disorder". In: *European Journal of Medical and Health Research* 2.6 (2024), pp. 182–185. DOI: 10.59324/ejmhr.2024.2(6).24.

[6] Gardener, H. et al. "Perinatal and Neonatal Risk Factors for Autism: A Comprehensive Meta-analysis". In: *Pediatrics* 128.2 (2011), pp. 344–355. DOI: 10.1542/peds.2010-1036.

[7] Maenner, Matthew J. 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". In: *MMWR Surveillance Summaries* 72.2 (2023), pp. 1–14. DOI: 10.15585/mmwr.ss7202a1.

[8] Yu, Tsung et al. "Parental Socioeconomic Status and Autism Spectrum Disorder in Offspring: A Population-Based Cohort Study in Taiwan". In: *American Journal of Epidemiology* 190.5 (2021), pp. 807–816. DOI: 10.1093/aje/kwaa241.

[9] Durkin, Maureen S. et al. "Autism Spectrum Disorder among US Children (2002–2010): Socioeconomic, Racial, and Ethnic Disparities". In: *American Journal of Public Health* 107.11 (2017), pp. 1818–1826. DOI: 10.2105/AJPH.2017.304032.

[10] Wang, C. et al. "Prenatal, Perinatal, and Postnatal Factors Associated with Autism: A Meta-analysis". In: *Medicine (Baltimore)* 96.18 (2017), e6696. DOI: 10.1097/MD.0000000000006696.

[11] Modabbernia, A. et al. "Environmental Risk Factors for Autism: An Evidence-based Review of Systematic Reviews and Meta-analyses". In: *Molecular Autism* 8.1 (2017), p. 13. DOI: 10.1186/s13229-017-0121-4.

[12] Child and Adolescent Health Measurement Initiative (CAHMI). *National Survey of Children's Health (NSCH)*. https://www.childhealthdata.org/learn-about-the-nsch/NSCH. Data Resource Center for Child and Adolescent Health. 2023.

[13] Hastie, Trevor et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York: Springer, 2009. DOI: 10.1007/978-0-387-84858-7.

[14] Hosmer, David W. et al. *Applied Logistic Regression*. 3rd. Wiley, 2013. DOI: 10.1002/9781118548387.

[15] Breiman, Leo. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[16] Cortes, Corinna et al. "Support-Vector Networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018.

[17] Hastie, Trevor et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer, 2009.

[18] Friedman, Jerome H. “Stochastic Gradient Boosting”. In: *Computational Statistics & Data Analysis* 38.4 (2002), pp. 367–378. DOI: 10.1016/S0167-9473(01)00065-2.

[19] LeCun, Yann et al. “Deep Learning”. In: *Nature* 521 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[20] Glorot, Xavier et al. “Deep sparse rectifier neural networks”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 15 (2011), pp. 315–323.