Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Across the Spectrum: A Study of Autism in National Survey Data Using Machine Learning

David He

Rockville, U.S.A

david_he@brown.edu

Abstract. Autism, a neurological disorder and developmental impairment, affects roughly 1 in 36 children in the US. However, relatively few machine learning algorithms, the majority being Logistic Regression models, have been used to predict autism using national data. In this study, other than the Logistic Regression algorithm, four other Machine Learning (ML) algorithms, namely the Random Forest, KNN, Extreme Gradient Boosting, and Support Vector Classifier algorithms, are applied to the NSCH data collected in 2021-2022 (namely, the National -Survey-'(*Data*)' of-Children's-Health), with a 7:3 training to testing split. Additionally, three resampling methods—'random over sampling, 'random under sampling', and 'SMOTE'—are leveraged to address class imbalance issues. Furthermore, the Shapley Additive ex-Planation (SHAP) values of specific features are visualized and evaluated for their importance. The scalability and potential of these ML algorithms in predicting ASD is explored. Among the children with autism, 77.5% were male, 45.1% had ADHD, 39.0% had allergies, 28.1% had a genetic condition, 29.8% had experienced parents' divorce, and 22.0% had experienced household hardship. From the SHAP selection, sex, ADHD, genetic conditions, low birth weight, allergies, certain ACEs, and other family factors are identified as important features. The precisions across five ML models are consistent around 95% while the recalls vary from 77% to 92% and F1 Scores range from 84% to 93%. The accuracies are between 0.77 and 0.92. This study demonstrates potential ML models applications in further understanding autism features of children, towards providing early assessment and tailoring data-driven evidence-based interventions.

Keywords: autism; machine learning algorithms; social demographic; health condition; NSCH data.

1. Introduction

Autism, also known as ASD (i.e., Autism -Spectrum -Disorder), is a syndrome of varied neurodevelopmental disorder that may impact one's communication abilities, behavioral routines, and social performance [1,2]. Worldwide, roughly 1 in 100 children have suffer from ASD [2]. It is estimated that in the United States, about 27.6 in 1000 children at the age of eight have been diagnosed with ASD [3]. The prevalence varies significantly depending on geographic, region, socio-economic status, sex, race, and community site and awareness [2,4].

Contemporary research is providing new insights into various facets of ASD. Researchers have studied the association of socioeconomic status, behavior, metabolic, and genetic markers with ASD to better understand the underlying causes of the disorder [5-8]. In the last ten years, the integration of prediction models with screening has made promising impacts in early assessment and treatment. Machine learning (ML) algorithms have improved in effectiveness, enhancing their ability to differentiate individuals with ASD from those without. Some ML models have been developed using RNA expression datasets to identify autism dysregulated genes [6], while others have measured the relationships of gut microbiome to ASD [7], and others have examined sensory processing scales to detect behavioral ASD [8].

Studies focusing on school-aged children with autism have shed light on targeted therapeutic interventions, supportive environments, and peer mentoring and connection as means to enhance and facilitate growth and development among children with autism [9]. Recent literature has suggested that sociodemographic factors, health comorbidity, parent education, and family-household experience could be particularly impactful to children with autism and/or ADHD [10-12]. Autistic children face health needs much more often than their peers, as well as geographic challenges regarding access to

much needed services [4,13]. Hence, accurate information for children with autism, specifically about their severity indication, health needs, family condition, and other social factors, is critical in tailoring both prevention and intervention measures [1,2]. Furthermore, shortcomings of current services and resources need to be identified to ensure and advance programs delivery or initiatives to support these children.

One method of solving these challenges is the utilization of predictive models to provide accurate factor estimates about not only social demographics but health characteristics of autistic children compared to their peers. Past publications on predicting autism have often had limitations such as focusing on singular models or lacking generalizability. Efforts are still needed to develop reproducible ML models that can accurately and effectively identify discerning traits and conditions in children with autism.

In this study, relevant factors, including social demographic characteristics, comorbidity conditions, and family and household experiences, will be investigated and utilized to predict the presence of autism using the recent NSCH (2021-2022) dataset. In addition to Logistic Regression, four other ML model algorithms will be utilized and combined with resampling techniques to leverage imbalanced data.

2. Methods

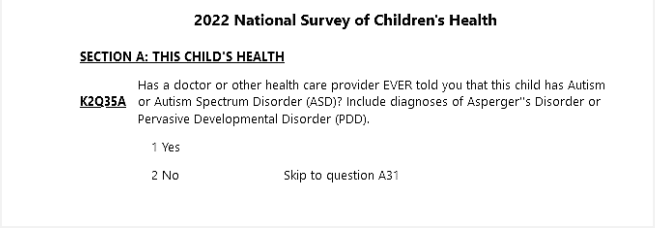
2.1. Study Data Source

The NSCH is an address-based survey conducted among households containing children using web-based and mail collection systems [14]. The survey provides nationally representative responses on a range of topics, including physical wellness, chronic conditions, emotional and mental health status, disabilities, healthcare access, insurance coverage, and family dynamics. Further detailed information on family routines, parental health, and social determinants of health like socio-economic and environmental factors has also been incorporated in recent years. The survey data has been released annually on Child Health Day since 2016 [14].

Data from the NSCH is used to determine the overall well-being of children and various aspects regarding their growth environments nationwide. The information is used by researchers, healthcare professionals, and policymakers to identify children's health trends and disparities, and to target intervention areas [14,15].

2.2. Study Population

The NSCH public use data was analyzed in this study. Data from the 'Screener/Demographic' and 'Topical' sub-datasets were merged into one dataset using their Household Identification System (HHIDS) to match corresponding responses.



2022 National Survey of Children's Health

SECTION A: THIS CHILD'S HEALTH

K2Q35A Has a doctor or other health care provider EVER told you that this child has Autism or Autism Spectrum Disorder (ASD)? Include diagnoses of Asperger"s Disorder or Pervasive Developmental Disorder (PDD).

1 Yes

2 No Skip to question A31

This study categorized the autistic children in the NSCH 2021-2022, who answered 'Yes' to the question below [14].

When removing all responses where at least one relevant feature was missing or left blank, 520 children with autism were excluded and 16,090 children without autism were excluded, leaving a total of 87,845 valid observations. Of those 87,845 responses, 2,926 (roughly 3.3% of the responses) answered 'Yes' to the above question and 84,919 (roughly 96.7%), answered 'No'. More detailed information can be found on the NSCH's webpage and a related Interactive Data Query [14,15].

2.3. Features Selected from the NSCH Data

The data features are extracted from the NSCH 2021 and 2022 dataset based on the public data accompanying codebooks, and contents map [14,15], and prior autism study publications [5,10-12]. Factors selected from the related measures are as shown the following.

i. Demographic Factors -- Age, gender, race.
ii. Health History -- Birth weight history, and chronic comorbid conditions such as ADHD, Allergy, and genetic disorders.
iii. Healthcare Access -- Health insurance coverage, healthcare visits, access to special care, health needs met/not met.
iv. Family and Household Factors -- Adversities children experienced (ACEs) such as hardship to cover family basics, parents/guardians divorced, deceased, in jail, child victim of violence, living with someone of mental illnesses, alcohol/drug problems.
v. Environmental Factors -- Home environment of house tenure status, parental age, highest parent education, and occupation.

2.4. Python Libraries/Packages Imported in the Study

The functions of several Python libraries used in the study are shown in TABLE I.

Table 1. List of Python libraries imported for the study#

Python Library/ Package	**Utilization for Library Imported**	
	Functionality	***Description***
Pandas	EDA: data processing and manipulation	Providing data structures and functions necessary for efficient large dataset cleaning and analysis
Numpy	EDA: data manipulation and analysis	Providing support for multi-dimensional arrays and matrices along with mathematical functions
Matplotlib	EDA: data plots and visualized analysis	Creating extensive customization of plots and integrating well with other scientific libraries for data visualization.
SHAP (SHapley Additive exPlanations)	EDA and ML: data analysis and features visualization	Generating the ML output explanation with values help to understand the contribution of each feature in visualization images
Scikit-learn	ML: data training and testing on algorithms, and performace evaluation	Providing various ML algorithms for regression, classification, and dimensionality reduction. Also producing data feature extraction, model selection, and performance metrics evaluation.
IMBlearn (Imbalanced-learn)	ML: resampling technology for unbalanced data	Modifying algorithms with resampling methods that adapt to be more sensitive to classify minority class instances. Three resampling methods, a) randomly Over-Sampling, b) randomly Under-Sampling, and c) SMOTE (Synthetic Minority Over-sampling Technique), were incorporated in the ML models.

2.5. Machine Learning Algorithms Applied in the Study

The ML algorithms of regression or classifiers include as the followings, available from the online source of "*Scikit-Learn User Guide*" [16].

- Logistic Regression – A supervised learning algorithm using logistic or sigmoid functions to map predicted values to probabilities. The algorithm evaluates the maximum likelihood estimation iteratively to find the best-fitting parameters until the predicted probabilities align closely with the actual class labels.
- Support Vector Machines (SVM) – A supervised learning algorithm defining the support vectors or data points, and thru the maximum boundary further finding the hyperplane. For Non-Linear data, kernel tricks such as Polynomial, Gauss, or Sigmoid are employed.
- Random Forest (RF) – A versatile ensemble learning algorithm by processing multiple decision trees and fitting them to reduce the variance and get a more stable and accurate classification.
- K-Nearest Neighbors (KNN) – A proximity instance-based learning algorithm classifying a data point based on how its neighbors were clustered or its similarity to previously labeled samples.
- Gradient-Boosted Decision Tree (with XGBoost) – A robust learning implementation on gradient-boosting decision trees and combining the predictions in an iterative sequential manner.

3. Results

3.1. Exploratory Data Analysis (EDA)

The EDA of this study includes a descriptive analysis and an exploratory check of the features in the data. The social-demographic, family and household, and health related characteristics from the topical public use data are compared for the groups of children with autism and children without (as shown in Table II). The autistic children were roughly 10 years old on average, around 1.7 years older than the mean age of non-autistic children. Meanwhile, the proportion of males in autistic children was 77.5%, significantly higher than the proportion of 50.9% male for non-autistic children.

Table 2. Descriptive Analysis for Base Characteristics of Autistic Children Compared to Others, NSCH 2021-2022

Survey Screener or Topicals	**Study Subjects Characteristics**		
	Selected Questions/Features Distribution %, or Mean (SE)	***Autistic Children***	***Non-Autistic Children***
Screener Demographics	Age, Mean (SE)	10.0 (0.1)	8.3 (0.1)
	Sex (Male)	77.5	50.9
	Race (White)	75.8	78.0
	Mother's Age, Mean (SE) *	29.9(0.1)	30.5(0.2)
This Child's Health Condition	ADHD (Yes)	45.1	8.6
	Allergy (Yes)	39.0	26.7
	Asthma (Yes)	15.7	9.6
	Diabetes (Yes)	7.4	4.6
	Genetic (Yes)	28.1	3.9
	Headache (Yes)	6.7	3.0
	Heart (Yes)	5.9	2.6
This Child as an Infant	Birth weight is low (Yes)	12.5	8.8
Health Insurance Coverage,	Currently covered (Yes)	98.0	96.6
	Coverage gap past 12 Months (Yes)	3.4	4.7

Survey Screener or Topicals	Study Subjects Characteristics		
	Selected Questions/Features Distribution %, or Mean (SE)	***Autistic Children***	***Non-Autistic Children***
HealthCare Services	Needed health care not received (Yes)	13.4	3.1
About Your Family and Household: The Child Experienced	Household hardship to cover basics, either on food, or on housing (Yes)	22.0	9.9
	Parent divorced, or guardian divorced (Yes)	29.8	19.0
	Lived in a family or with someone mentally ill (Yes)	21.0	8.7
	Ever been treated unfairly that's related to entity's health conditions (Yes)	27.1	2.3
General Household Information	Number of people living at address (Mean) *	4.0	4.0
	Highest level of education, adult (high school or lower)	17.4	14.0
	language speaking at home (Non-English)	5.9	7.5
	Tenure conditions, building held or occupied (Rent)	24.1	16.9

Note: All percentages are unweighted. The P values <0.01 (t-test or chi-square test), except *.

In the group of children with autism, 45.1% had been diagnosed with ADHD, 39.0% had allergies, 15.7% had asthma, and 28.1% had genetic disorders. Meanwhile, in the non-autistic group, 8.6% of children had ADHD, 26.7% had allergies, 9.6% had asthma, and 3.9% had genetic disorders.

Additionally, a significantly higher proportion of children in the autism group had Adverse Childhood Experiences (ACEs), with 22.0% having experienced hardships regarding food/housing, 29.8% having experienced parents/guardians divorcing, 21.0% having lived with someone mentally ill and 27.1% having experienced been treated unfairly due to their health conditions. Of the group of non-autistic children, only 9.9% had experienced hardships, 19.0% had experienced parents/guardians divorcing, 8.7% had lived with someone mentally ill, and 2.3% had been treated unfairly due to their health conditions.

The heatmap of the features correlation matrix (as shown in Figure 1) presents the relationship of selected variables. Of particular note is the relatively strong correlation between the ACEs, indicating that when one ACE occurs, others are likely to follow.

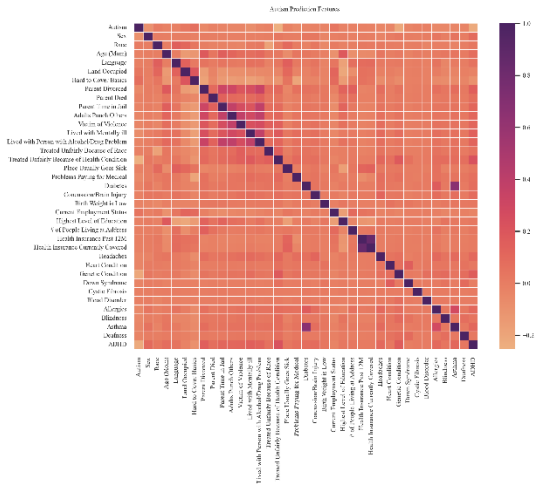


Figure 1. Features Correlation Heatmap for Autism Related Factors, NSCH 2021-2022

To generate the SHAP feature scores and the SHAP feature influences in the prediction, a sample subset selection of 4800 subjects from the RF model was deployed and plotted (as shown in Figure 2 & Figure 3). ADHD, sex, genetic disorders and low birth weight, and several ACEs are shown to have higher feature scores, indicating a higher contribution to the overall identification.

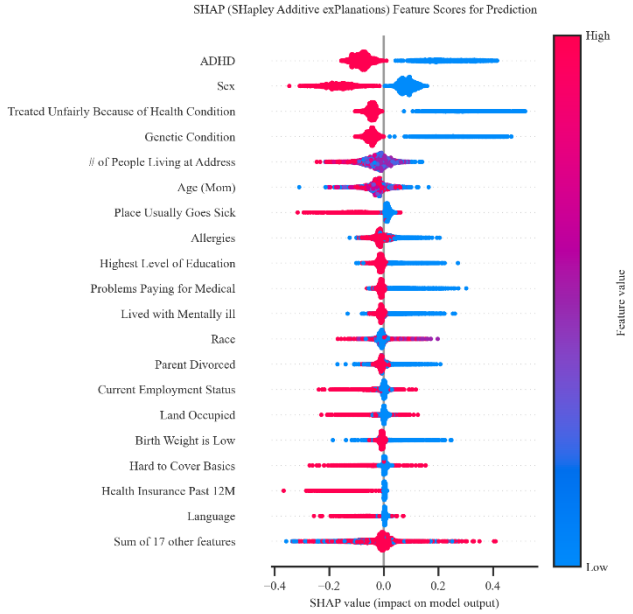


Figure 2. Features Selection by SHAP Scores for the Random Forest Model to Predict Autistic Children, NSCH 2021-2022

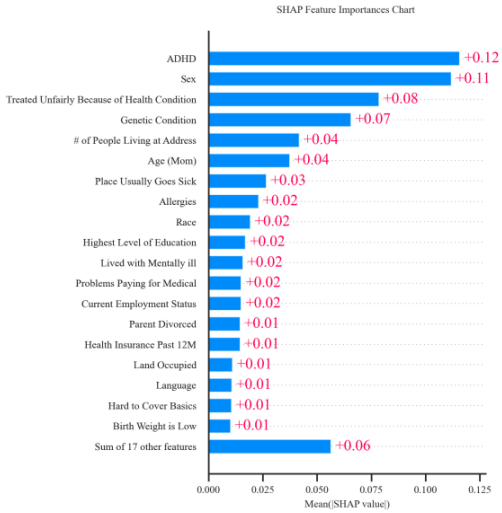


Figure 3. SHAP Importance Bar Chart for Features in the Random Forest Model to Predict Autistic Children, NSCH 2021-2022

3.2. Analysis on Machine Learning Models

After the exploratory analysis, the data was randomly split, with a 70%: 30% ratio, into a training dataset and a testing dataset. Based on the confusion matrices of the ML models when applied to the testing dataset, several model metrics were calculated with sci-kit learn [16]. The performances of the five ML algorithms along with the sampling techniques were evaluated (as shown in Table III) according to the metrics of -- i) Precision, ii) Recall, iii) F1-score, and iv) Accuracy.

Table 3. Evaluation Comparison for Metrics of Weighted Precision, Recall, F1- Score, & Accuracy in five ML Algorithms to Predict Autistic Children, NSCH 2021-2022

Machine Learning Algorithms	**Model Evaluation**				
	Imbalanced Technique Applied	***Precision***	***Re call***	***F1 Score***	***Accuracy***
Logistic Regression	OverSampling	0.96	0.85	0.89	0.85
	UnderSampling	0.96	0.85	0.89	0.85
	SMOTE	0.96	0.86	0.90	0.86
SVM	OverSampling	0.96	0.87	0.90	0.87
	UnderSampling	0.96	0.87	0.91	0.87
	SMOTE	0.96	0.87	0.91	0.87
Random Forest	OverSampling	0.94	0.92	0.93	0.92
	UnderSampling	0.96	0.77	0.84	0.77
	SMOTE	0.95	0.90	0.92	0.90
KNN	OverSampling	0.94	0.91	0.93	0.91
	UnderSampling	0.95	0.88	0.90	0.88
	SMOTE	0.95	0.90	0.92	0.90
XGBoost	OverSampling	0.95	0.88	0.91	0.88
	UnderSampling	0.96	0.80	0.86	0.80
	SMOTE	0.95	0.87	0.90	0.87

As shown above, the precision is relatively consistent across all five models, with a range of 0.94-0.96 and an overall mean of 0.954. The recall varies more and is also smaller than the precision across the five models, with a mean of 0.867, and is especially high for the KNN and RF Algorithms. The F1-

Score is more consistent than the recall but less consistent than the precision, with a mean of 0.901. Finally, the accuracy values range from 0.85-0.92 for over-sampling method, 0.86-0.90 for SMOTE method, and 0.77-0.88 for under-sapling method.

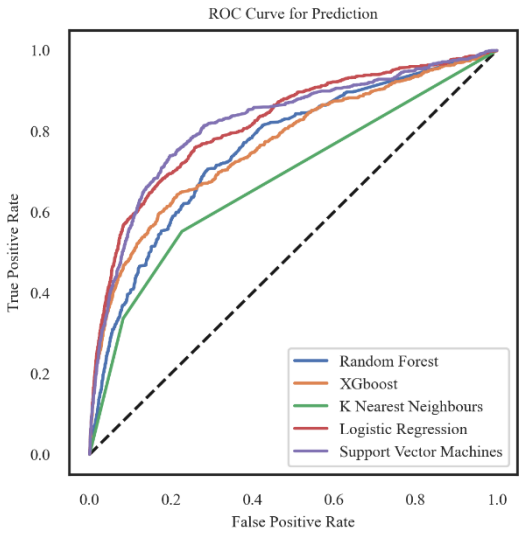


Figure 4. ROCs of Five ML Models to Predict Autistic Children, NSCH 2021-2022

The Receiver Operating Characteristic (ROC) curves for five models using the SMOTE technique were also visualized (as shown in Figure 4). The SVM, Logistic Regression, XGBoost, and RF models displayed similar optimal effects, with related scores of 0.820, 0.818, 0.764, and 0.758 for the Area Under Curve (AUC); the ROC curve for the KNN algorithm was consistently the lowest.

4. Discussion

4.1. Strength and Contribution of the Study

Statistically significant differences were found in several features such as age, sex, chronic comorbidities (including ADHD, allergy, asthma), birth weight, health insurance gaps, unmet healthcare needs, household hardships and certain ACEs. The importance of these features is corroborated by these factors being included and identified as significant from prior research [10-12]. However, these studies primarily focus on specific factors, and typically limit on using a logistic regression model.

In this study, in addition to a logistic regression model, four other ML algorithms were deployed along with a relatively recent dataset, reinforcing prior findings. These ML algorithms show promising potential, especially in combination with oversampling and SMOTE techniques to address imbalanced data, which are necessary due to the small proportion (less than 5%) of children who have ASD.

In the future, data-evidence-based interventions could be feasibly introduced, and personalized treatment could be realized should ML algorithms be further developed to improve models and enhance the interpretability of comprehensive features' contributions to autism in children. Crucial factors associated with autistic children could be better addressed with collaboration from family members, clinicians or therapists, healthcare services, schools, communities, and society as a whole.

4.2. Limitations of the Study and Future Direction

The survey response rate for weighted overall completion was 40.3% and 39.1% for 2021 and 2022, respectively [14]. Because this data is from a cross-sectional survey, no factors of causality could be determined; however, strong correlations between specific factors and the prevalence of autism were observed. Biases that may have emerged due to the low rate (<5%) of autism in the dataset have been addressed using oversampling techniques in the ML models; overfitting might have been consequently

introduced. More stable and replicable algorithms still need to be established and evaluated, along with more in-depth data cleaning and features processing to improve the model's performance.

Other information not addressed in this study, such as factors explored from other studies, regarding family medical history, clinical information, MRI measurements, and gut microbiome, could be explored also in future ML models with these specified features being available for the national study. Additionally, building ML models on strata of groups, collecting longitudinal cohort information, and inspecting the features' connections upon the health, genetic, metabolic, environmental, and social characteristics could be some other valuable areas for future ML autism studies.

5. Conclusion

In this study, five ML algorithms—Logistic Regression, SVM, RF, KNN, and XGBoost—were modified using resampling techniques and applied to the survey data of the 2021-2022 NSCH. This study analyzed 87,845 individual observations that had no missing responses in certain relevant features. By applying feature selection and leveraging resampling techniques, higher precisions of 95% and accuracies in the range of 77%-92% were achieved across the five ML algorithms. Oversampling and SMOTE techniques improved the models' performances; these techniques can be utilized to perform ML analysis on similarly imbalanced data in the future. The findings regarding variables such as social demographics, health and comorbidity conditions, ACEs, and unmet health needs could also be used to explore and expand ML algorithms in an evidence-based autism research field. Additionally, this study illustrates the potential applicability of ML models in conjunction with SMOTE techniques to predict autism, which may promote broader implementations of ML, as well as continuous improvement of ML models in related medical and health research fields.

References

[1] T Hirota, and B King, "Autism spectrum disorder: a review," Journal of American Medical Association, vol 329, no 2, Jan 2023, pp 157-168.

[2] J Zeidan, E Fombonne, 'J Scorah, 'A Ibrahim, and MS Durkin, "Global prevalence of autism: A systematic review update," Autism Res, vol 15, no 5, May 2022, pp 778-790.

[3] MJ Maenner, Z Warren, AR Williams, E Amoakohene, and AV Vakian, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network 11 sites, united states, 2020," MMWR Surveillance Summaries, vol 24, no 72, March 2023, pp 1–14.

[4] J Bradshaw, JM Eberth, A Zgodic, A Federico, and K Flory, "County-level prevalence estimates of autism spectrum disorder in children in the United States," J Autism Developmental Disorders, vol 54, no 7, July 2024, pp 2710-2718

[5] TJ Wong, and T Yu, "Association between socioeconomic status and prevalence of hypersensitivity diseases and autism: a nationwide study of children", Maternal and Child Health Journal, vol 27, no 12, Dec 2023, pp 2194-2202.

[6] I Voinsky, OY Fridland, A Aran, RE Frye, and D Gurwitz, "Machine learning-based blood RNA signature for diagnosis of autism spectrum disorder," International Journal of Molecular Sciences, vol 24, no 3, Jan 2023, pp 2082.

[7] Q Su, OWH Wang, W Lu, Y Wan, L Zhang, "Multikingdom and functional gut microbiota markers for autism spectrum disorder," Nat Microbiol, July 2024, online, ahead of print.

[8] H Alateyat, S Cruz, E Cernadas, M Tubio-Fungueirino, and A Sampaio, "A machine learning approach in autism spectrum disorders: from sensory processing to behavior problems," Frontiers in Molecular Neuroscience, vol 15, May 2022, 889641.

[9] L Hasson, S Keville, J Gallagher, D Onagbesan, and AK Ludlow, "Inclusivity in education for autism spectrum disorders:eexperiences of support from the perspective of .parent carers, 'school teaching staff and young people on the autism spectrum," International Journal of Developmental Disabilities, vol 70, May. 2022, pp /201-212.

[10] M Salehi, A Ahmad, A Lotfi, and S Gunturu, "Characteristics and co-morbidities of autism spectrum disorder as risk factors for severity: a national survey in the United States," version 1, Preprint, Retrieved [07/16/24], available at Research Square, Feb 2024.

[11] CM Kerns, CJ Newschaffer, S Berkowitz, and BK Lee, "Examining the association of autism and adverse childhood experiences in the national survey of children's health: the important role of income and co-occurring mental health conditions," Journal of Autism and Developmental Disorders, vol 47, July 2017, pp 2275-81.

[12] A Federico, /A Zgodic, /K Flory, /RM Hantman, and /JM Eberth, “Predictors of autism spectrum disorder and ADHD: Results from the National Survey of Children's Health,” Disability and Health Journal, vol 17, no 1, Jan 2024, pp 101512

[13] RM Hantman, A Zgodic, K Flory, AC McLain, J Bradshaw, JM Eberth, “Geographic disparities in availability of general and specialized pediatricians in the us and prevalence of childhood neurodevelopmental disorders,” J Pediatr, vol 12, Jul 2024, pp 114188.

[14] National Survey of Children’s Health, Health Resources and Services Administration, Maternal and Child Health Bureau, Retrieved [08/02/24],’[mchb.hrsa.gov/data/national--surveys]

[15] National /Survey of /Children’s /Health /Interactive Data Query, Data Resource Center for Child and Adolescent Health, Retrieved [08/02/24],/[www.childhealthdata.org]./

[16] Scikit-Learn User Guide, Retrieved [08/02/24],/[www.scikit-learn.org/stable/user_guide.html]