Behavioral Classification and Characterization of Autism Spectrum Disorder in Naturalistic Settings using Classical Machine Learning

Elliot Huang
School of Computer Science
Georgia Institute of Technology
Atlanta, GA 30332
elliot.huang@gatech.edu

Lemuel Mojica Vazquez
Department of Computer Science
and Electrical Engineering
Ana Mendez University
Gurabo, Puerto Rico
lmojica48@email.uagm.edu

Nicolas Echevarrieta-Catalan
Department of Computer Science
University of Miami
Coral Gables, FL 33146
nxe272@miami.edu

Laura Vitale
Department of Psychology
University of Miami
Coral Gables, FL 33146
lcv31@miami.edu

Daniel S. Messinger
Department of Psychology
University of Miami
Coral Gables, FL 33146
dmessinger@miami.edu

Vanessa Aguiar-Pulido
Department of Computer Science
University of Miami
Coral Gables, FL 33146
vanessa@cs.miami.edu

Abstract

Autism Spectrum Disorder (ASD) is a group of complex neurodevelopmental disorders that affects about 1% of the world's population, impacting the quality of life of not only the diagnosed individuals but also their communities. Early detection and intervention are paramount to limit its effect on a child's development, however overlap with other disorders and medical comorbidities make these tasks challenging. The present study explores the use of a novel multimodal, interpretable approach to characterize ASD children's behavior in a naturalistic environment. Spatial (real-time location tracking), audio and demographic data from children in a classroom setting are integrated and analyzed to identify traits potentially connected to ASD. Our findings point to the use of this type of approach as a potential tool for screening individuals in a naturalistic setting, allowing for further evaluation and, ultimately, earlier diagnosis by a clinician. Results show good classification performance and suggest vocalization, speech, proximity and certain movement-related features to be impacted in ASD.

1 Introduction

Autism Spectrum Disorder (ASD) is a developmental disability that is characterized by changes in communication, social interaction and behavior. Challenges emerging from this neurodevelopmental disorder influence the totality of an individual's experience of life, with a markedly lower quality of life at every stage [1]. Currently, it is estimated that over 75 million people worldwide are living with this disorder, with 1 in every 36 children at the age of 8 years affected by it in the US [2]. Traditionally,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

ASD has been diagnosed with evaluations conducted by trained specialists, based on guidelines such as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), and the Modified Checklist for Autism in Toddlers (M-CHAT) [3]. However, this approach is not without issues and despite numerous attempts to diagnose autism utilizing chemical and biological means, no reliable biomarkers exist that could enable earlier diagnosis [4] as there is much to elucidate regarding ASD's etiology and pathogenesis.

Advancements in technology and machine learning (ML) offer a promising avenue to better characterize this heterogeneous disorder. Noteworthy examples include the utilization of contemporary clinical assessment data in a federated learning approach, which featured statistical learning classifiers [5], and the application of deep learning to neuroimaging [6]. Moreover, the fusion of diverse data types has shown the potential to further bolster predictive capabilities. A study by Kollias et al. demonstrated that a multimodal data integration approach, incorporating eye-tracking, kinematics, and electroencephalography (EEG), outperformed purely eye-tracking-based training in terms of accuracy [7]. Nonetheless, it is important to acknowledge that the translation of ML research findings into clinical outcomes remains a sizeable challenge, marked by factors such as limited sample sizes and inconsistent findings [8].

In this work, we sought to adopt a holistic approach, capturing some of the complexities inherent to interaction in real-world environments. We propose a multimodal, interpretable approach to analyze spatial, audio and demographic data collected in a classroom environment which included children with and without ASD. The goal of the present work is two-fold: (1) gain a better understanding of ASD, and (2) identify features that could be further explored as potential (behavioral) biomarkers of ASD.

2 Materials and Methods

The approach proposed here integrates three different types of data, which are carefully curated to be used as input to a machine learning-based algorithm for data analysis (Figure 1). The different tasks performed as part of this workflow are detailed subsequently.

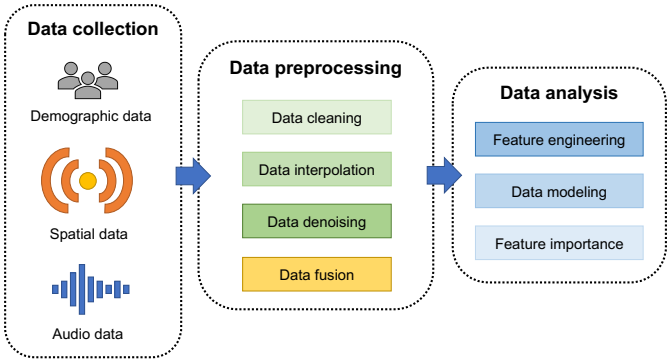

Figure 1: Overall workflow utilized in the present work.

2.1 Data collection

A total of 213 observations were used as input to the predictive models, further divided into 74 pertaining to individuals diagnosed with ASD and 139 from individuals with typical development (TD). Observations were obtained from 44 children (21 TD and 23 ASD) from six different classrooms who participated in a particular learning program between the years 2018 and 2022. Demographic information was manually collected. Spatial and audio data were collected once a month for each class, each instance spanning 1-3 hours and with a total average of 5 instances per class. Data collection procedures adhered to ethical standards, with participation requiring written informed consent from the parents or legal guardians of the children. Furthermore, our research protocol fully complied with the guidelines set by the Helsinki Committee and received approval from the University of Miami Social & Behavioral Sciences Institutional Review Board (IRB) under protocol number #20160509.

During data collection, children wore vests containing a LENA-SP audio recorder, which allowed acquisition of audio data, and RFID tags on their left and right sides, allowing determination of position and orientation. Vest wearing compliance rate exceeded 97%, with only a few isolated instances of non-compliance [9]. Ubisense DIMENSION4 Radio Frequency Identification Real-Time Location System with Research Upgrade was employed to track the real-time location of children within classroom spaces. The classrooms are equipped with four radio cell sensors attached to the corners of the classrooms that detect the location of the tags 2-4 times per second with an accuracy of 15-30 cm [10]. Location measurement was based on a xy coordinate system, where the origin (0,0) was set as the sensor in the corner closest to the primary entrance. The LENA recorder captures the surrounding sound during observations, in addition to a child's vocalizations.

2.2 Data preprocessing

The spatial and audio data collected during classroom observations were preprocessed for downstream analysis, including data cleaning, interpolation, denoising and fusion. Onset, offset and intensity of each child's vocalization were extracted using the LENA Pro pattern recognition software. Motion data was interpolated to the precision of one-tenth of a second, orientation of subjects was determined, and spatial separation between the left and right tags was computed from the real-time location tracking data. To reduce measurement anomalies and data disturbances, Kalman filters were employed. Finally, audio, spatial and demographic data were integrated for each individual subject.

2.3 Feature engineering

A total of 39 features across seven categories were extracted for each observation. Spatial and audio features underwent a normalization process with robust min-max scaling techniques, ensuring that their values fell within the ranges of -1 to 1 or 0 to 1 where appropriate. Features are described below.

- **Demographic information.** A variable encoding whether a child was monolingual or bilingual was included.
- **Movement.** Stereotyped patterns of movement have been linked to autism. Three general statistics of movement are thus computed: mean speed, speed variability (standard deviation), and rotational speed variability. Additionally, to measure variability in larger intervals, we calculated the proportion of location changes (leaving set 1-meter zones) to time.
- **Proximity.** The term "proximity" is used to denote individuals less than 1 meter apart [11]. Each individual could be in proximity of ASD children, TD children, and teachers. We quantified the ratio of instances in which each specific group was found in proximity of the subject relative to the aggregate instances of proximity involving any group.
- **Social contact.** Two individuals are considered to be in "social contact" if the separation between them is between 0.2 to 2 meters and their orientations are within 45 degrees of face-to-face contact [9]. A buffer of one second was provided to accommodate brief interruptions. The proportion of time, average duration, and the longest period of social contact were calculated. Additionally, the ratio of cumulative duration spent in social contact with teachers, ASD and TD children in relation to the overall aggregate duration of social contact was also computed.
- **Approach velocity to social contact.** This term denotes the speed of an individual in motion (a maximum of 2.5 seconds) prior to engaging in social contact, relative to the initial position of the counterpart participant. Nine features were calculated within this category depending on who is involved in the contact: 3 capturing the average approach velocity towards each type of individual (i.e., teacher, ASD child, TD child), 3 quantifying the variability in approach velocity exhibited towards each group, and 3 gauging the proportion of approaches directed at each type of individual relative to the total times of contact with that group. Additionally, 3 global features were obtained, representing the overall average approach velocity, the proportion of approach to social contact events, and the mean of the three average approach velocities.
- **Vocalizations.** The mean, variability, and maximum of peak intensities were calculated.
- **Speech.** We introduce the concept of "speaking events", indicated by semi-continuous sequences of vocalizations distinguished by gaps of less than one second between successive

vocal elements. We also further subdivided these events, creating a more specific category within the domain of speaking events called "response". This grouping identifies instances where a child's speaking event is preceded by another individual's speaking event within a maximum time frame of six seconds, with both individuals engaging in social contact for the duration of the time gap. The mean and maximum duration of speaking events were computed, as well as the cumulative duration of speech vs. the overall duration of social contact. Additionally, we explored two response-related features: one that analyzes responses in relation to the total speaking duration and another that calculates the average time it takes for a child to respond.

2.4 Data modeling

A comprehensive set of interpretable ML models were built using as input the features described in 2.3, including: K-Nearest Neighbors (KNN), logistic regression (LR), ridge regression (RR), decision trees, random forests, extra-trees, gradient boosting, AdaBoost, and support vector machines (SVM). All models were implemented employing the scikit-learn library [12]. Model performance was optimized through a grid-search hyperparameter tuning process. Additionally, R was utilized to train a generalized linear model (GLM) on all features and extract p-values to determine which of these were statistically significant at α=0.05.

It is worth noting that LR and RR classify an observation by estimating the probability of it belonging to the positive class. Thus, an observation is classified as positive if the probability is greater than or equal to a default threshold, and as negative otherwise. Using a default value for the threshold may become problematic in situations in which the dataset used as input is imbalanced. As such, in the current scenario, the regression models could suffer from a bias towards predicting the class with the greatest number of observations. To minimize the amount of false negatives, we introduced an additional hyperparameter termed the "threshold." This hyperparameter serves the crucial function of adjusting the minimum required probability for the model to classify an observation as the positive class, thus contributing to optimal predictive performance.

Traditional cross-validation could not be applied to our unique dataset. For example, in the standard k-fold CV, data are divided into k portions and a randomly selected subset of the data is designated as the validation set, while the model is trained on the remaining portions. Then, the model's performance is evaluated utilizing this validation set to ascertain accuracy. This process is repeated k times and performance metrics are averaged to prevent selecting a model that performs optimally for one validation set and which may not generalize well. However, even if employing the stratified version of this validation technique, there is a high possibility for an individual child's data observations to be present in both the training and validation sets, thus introducing a bias into the model's evaluation. The metrics would then reflect a model's proficiency in predicting specific children's attributes rather than its capability to discern autism-related features. Hence, in this work, we propose the use of what we denominated leave-one-kid-out cross-validation (LOKOCV). Building upon the concept of leave one out cross-validation CV (LOOCV), a subtype of k-fold CV in which k=number of total input observations, we introduced a unique adaptation that changes what the term "one" represents. In our method, we replaced the exclusion of a single data point with the exclusion of an entire individual. Essentially, all data points associated with a specific child are omitted from the training set and designated as the validation set. This process is iteratively performed for each child within the dataset, ensuring that an individual child's data samples are not part of both the training and validation sets. In our particular experimental setup, the LOKOCV involved 44 folds, with each fold including data pertaining to 43 kids for the training set while the observations of 1 child are reserved for validation purposes.

3 Results

3.1 Classification performance

Models were evaluated using four different metrics: accuracy, F1 score, area under the receiver operating characteristic curve (AUROC) and true positive rate (TPR). Given the imbalance in our dataset (139 TD vs. 74 ASD observations), our primary focus was on optimizing the F1 score, which measures both precision and recall, and AUROC, which measures the ability of a model to

Table 1: Model performance

Model	Accuracy	TPR	F1 score	AUROC
Logistic Regression	0.79	0.82	**0.73**	**0.79**
Support Vector Machine	**0.80**	0.70	0.71	0.78
Ridge Regression	0.76	**0.84**	0.71	0.78
Decision Tree	0.78	0.68	0.68	0.76
Gradient Boosting	0.78	0.62	0.66	0.74
AdaBoost	0.76	0.55	0.61	0.71
Random Forest	0.78	0.51	0.61	0.71
Extra-Trees	0.77	0.50	0.60	0.70
K-Nearest Neighbors	0.73	0.42	0.50	0.65

discriminate between the two classes. TPR was important to assess the performance of the proposed approach in correctly identifying ASD individuals.

Table 1 shows the best metrics for each of the optimized models built in this work, with three models performing remarkably better than the rest: LR, RR, and SVM. Notably, all three of these models employed a linear decision boundary, suggesting a mainly linear relationship between input features and the dependent variable. LR emerged as the top-performing model, achieving an F1 score of 0.726 and a AUROC of 0.793. Closely behind, SVM and RR obtained F1 scores of 0.712 and 0.705 and AUROC values of 0.779 and 0.775 respectively. The optimal configuration for LR involved a liblinear solver with an L1 penalty and a threshold of 0.34, SVM performed best with a gamma-scaled linear kernel, and RR achieved the best results employing an SVD solver and a threshold of -0.27. In terms of TPR, RR performed the best, reaching a value of 0.838. LR and SVM achieved a TPR of 0.824 and 0.703 respectively. Higher TPR values for LR and RR could be attributed to the optimization of the "threshold" hyperparameter (see 2.4).

3.2 Feature importance

Both LR and RR models operate by modifying the positive class probability through the application of specific weights to input features. As such, the magnitudes of these weights offer valuable insights into the relative importance of individual features in making predictions, which we will analyze next.

Table 2 includes the 10 highest weighted features of the LR model. In particular, vocalization features emerge as the most prominently weighted attributes, suggesting a substantial disparity in vocalization intensity between ASD and TD individuals. Additionally, speech-related features, particularly those associated with duration, are within the next most highly ranked features. Proximity features also command considerable weight, especially in relation to ASD children and teachers. Similarly, features related to the duration of social contact capture attention. Lastly, a single movement feature, specifically speed variability, emerges as the 10th highest ranked feature. A similar narrative unfolds in the case of RR. The top six features remain consistent with those found for LR, reinforcing the importance of vocalizations, proximity, and speech-related features. However, a noteworthy addition is the appearance of approach velocity to children, displacing the lone movement feature from the top 10. Additionally, the ratio of social contact with ASD compared to total social contact replaces the proportion of time allocated to social contact as a highly ranked attribute in the model.

Next, we sought to calculate feature-specific statistical significance information. For this purpose, we built a GLM (which in practice was set to be a LR model) employing all features as input in R. At a significance level (α) of 0.05, we observed that seven features had statistically significant p-values (Table 3). These features fall into five different categories, namely, vocalization, proximity, speech, movement and approach velocity. Only two categories (vocalization and speech) contain more than one statistically significant attribute. It is worth noting that, while the number of social contact events the subject approached and the average speed of movement were not features that ranked within the top 10 for LR and RR, in the GLM model they exhibited statistical significance. Overall, the results of this analysis reinforce the notion of vocalization, speech, and proximity being impacted in ASD.

Table 2: Top 10 Logistic Regression weights

Feature	Category	Weight
Average peak decibels	Vocalizations	2.99
Standard deviation of peak decibels	Vocalizations	2.84
Proximity ratio to ASD group	Proximity	1.77
Average speaking duration	Speech	1.59
Longest speaking duration	Speech	0.92
Proximity ratio to teacher	Proximity	0.82
Duration ratio of speaking contact	Speech	0.41
Ratio of social contact to time	Social Contact	0.39
Average duration of social contact	Social Contact	0.39
Standard deviation of movement speed	Movement	0.36

Table 3: Statistically significant features

Feature	Category	P-value
Average peak decibels	Vocalization	0.00000577
Standard deviation of peak decibels	Vocalization	0.000211
Proximity ratio to ASD group	Proximity	0.000456
Average speaking duration	Speech	0.007525
Average movement speed	Movement	0.027029
Approach ratio	Approach Velocity	0.031794
Longest speaking duration	Speech	0.033559

4 Conclusions and Future Work

Our research has showcased the promising capabilities of statistical ML models to deepen our understanding of ASD. This work proposed a novel multimodal approach that represents a first step toward the development of tools that could potentially be used for screening individuals in naturalistic settings (e.g., a classroom environment). By detecting patterns that could be linked to ASD, an individual could be referred for further evaluation to ensure prompt diagnosis and intervention. It is worth noting that no protected health information is collected as part of this study and that all data collected are de-identified prior to data preprocessing and analysis, thus limiting any concerns that may arise regarding privacy and security. Although referring individuals for additional testing may be felt as potentially stigmatizing to some, given the impact of early intervention on child development, we believe that contributing to early diagnosis outweighs the potential risks.

Our models achieve performance values of about 80% and flag interesting features for further exploration, including speech-related attributes, proximity patterns, and social contact duration, contributing toward pinpointing specific behaviors in autism and, consequently, to the characterization of this disorder. Nevertheless, there is still substantial work ahead. Our approach can be further refined through the optimization of existing features, the extraction of additional attributes, and the integration of other data types. Moreover, the features identified to be relevant by the models will require further research to solidify these into well-defined behavioral biomarkers. For this purpose, statistical analyses will be performed. Furthermore, additional data will be collected to validate the proposed model and its findings. Finally, our models have paved the way for more complex machine learning techniques to deepen our comprehension of ASD, for example, combining latent space representations and classical machine learning or employing deep learning to detect more complex patterns.

Acknowledgments and Disclosure of Funding

The work presented in this paper is supported in part by NSF Grant No. DS-2150830 and CNS-1949972. The completion of this research was made possible thanks to hardware support and facilities provided by the Department of Computer Science, the Department of Psychology and the Frost Institute for Data Science & Computing at the University of Miami.

References

[1] van Heijst, B.F. & Geurts, H.M. (2014) Quality of life in autism across the lifespan: A meta-analysis, *Autism* **19**(2):158–167.

[2] Maenner, M.J., Warren, Z., Williams, A.R., Amoakohene, E., Bakian, A.V., Bilder, D.A., Durkin, M.S., Fitzgerald, R.T., Furnier, S.M., Hughes, M.M., Ladd-Acosta, C.M., McArthur, D., Pas, E.T., Salinas, A., Vehorn, A., Williams, S., Esler, A., Grzybowski, A., Hall-Lande, J., Nguyen, R.H.N., Pierce, K., Zahorodny, W., Hudson, A., Hallas, L., Mancilla, K.C., Patrick, M., Shenouda, J., Sidwell, K., DiRienzo, M., Gutierrez, J., Spivey, M.H., Lopez, M., Pettygrove, S., Schwenk, Y.D., Washington, A. & Shaw, K.A. (2023) Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020, *Morbidity and Mortality Weekly Report – Surveillance Summaries* **72**(2):1-14.

[3] Lordan, R., Storni, C. & De Benedictis, C.A. (2021) Autism spectrum disorders: Diagnosis and treatment, *Autism Spectrum Disorders*, Exon Publications, pp. 17–32.

[4] Shen, L., Liu, X., Zhang, H., Lin, J., Feng, C. & Iqbal, J. Biomarkers in autism spectrum disorders: Current progress, Clinica Chimica Acta **502**:41-54.

[5] Farooq, M.S., Tehseen, R., Sabir, M. & Atal, Z. (2023) Detection of autism spectrum disorder (ASD) in children and adults using machine learning, Scientific Reports **13**(1):9605.

[6] Eslami, T., Almuqhim, F., Raiker, J.S., Saeed, F. (2021) Machine Learning Methods for Diagnosing Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder Using Functional and Structural MRI: A Survey, Frontiers in Neuroinformatics **14**:575999

[7] Kollias, K.-F., Syriopoulou-Delli, C.K., Sarigiannidis, P. & Fragulis, G.F. (2021) The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review, *Electronics*, **10**(23):2982.

[8] Bahathiq, R.A., Banjar, H., Bamaga, A.K. & Jarraya, S.K. (2022) Machine learning for autism spectrum disorder diagnosis using structural magnetic resonance imaging: Promising but challenging, *Frontiers in Neuroinformatics* **16**:949926.

[9] Banarjee, C., Tao, Y., Fasano, R.M., Song, C., Vitale, L., Wang, J., Shyu, M.L., Perry, L.K. & Messinger, D.S. (2023) Objective quantification of homophily in children with and without disabilities in naturalistic contexts, *Scientific Reports* **13**(1):903.

[10] Irvin, D.W., Crutchfield, S.A., Greenwood, C.R., Kearns, W. D. & Buzhardt, J. (2017) An automated approach to measuring child movement and location in the early childhood classroom, *Behavior Research Methods* **50**(3):890–901.

[11] Messinger, D.S., Prince, E.B., Zheng, M., Martin, K., Mitsven, S., Huang, S., Stölzel, T., Johnson, N., Rudoplh, U., Perry, L., Laursen, B. & Song, C. (2019) Continuous measurement of dynamic classroom social interactions, *International Journal of Behavioral Development* **43**(3):263–270.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* 12:2825-2830.