MACHINE LEARNING CLASSIFICATION AND CHARACTERIZATION OF AUTISM SPECTRUM DISORDER IN PRE-SCHOOL AGED CHILDREN

Elliot Huang[4], Lemuel Mojica Vazquez[5], Nicolas Echevarrieta Catalan[1], Laura Vitale[2], Daniel S. Messinger[2, 3], Vanessa Aguiar-Pulido[1]

Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition distinguished by a broad range of symptoms and behaviors that profoundly impact communication, social interaction, and the quality of life of individuals for the entirety of their lives [1]. As a result, the pursuit of comprehensive research in the domain of this developmental disorder becomes paramount in fostering early and effective interventions.

In recent years, advancements in data science and machine learning have offered promising opportunities to aid the diagnosis and treatment of ASD. However, existing algorithms often focus on specific modalities [2], potentially missing crucial aspects of an individual's behavior in their naturalistic settings. To address this limitation, we sought to embrace a more holistic approach that considers the complexities of social interactions in real-world environments.

The primary objective of this research project is to use a broad behavioral approach to predicting and characterizing autism. *Leveraging vast datasets of location tracking and audio data, we aim to explore features and patterns that could serve as potential behavioral biomarkers for identifying ASD individuals.*

Methods: Process

1. Data Collection
Data were collected from 21 Typical Development (TD) and 23 ASD kids who attended a single classroom between 2018 to 2022. The two means of quantification were Ubisense RFID real-time location tracking [2] and LENA Digital Language Processors (Audio Recorders) [3].

2. Data Preprocessing
Data preprocessing was conducted to integrate location and audio data. The noise was reduced by the Kalman Filter, and missing values were handled [5].

3. Feature Engineering
Features were extracted from observations with a time window ranging from 1 to 3 hours. With domain knowledge and data-driven techniques, we carefully selected relevant features that could aid in discriminating between TD and ASD individuals. In total, we obtained features from 74 ASD observations and 139 TD observations, all of which were normalized to ranging from -1 to 1 or 0 to 1.

4. Model Development and Optimization
We explored various machine learning algorithms for classification: K-NN, Decision Tree, Logistic Regression, Ridge Regression, Gradient Boosting, Random Forests, AdaBoost, and SVC. We conducted a hyperparameter tuning process using grid search and leave-one-kid-out cross-validation (variation of LOOCV that uses folds composed of all observations of a single kid).

5. Metrics and Interpretation
The models were evaluated using a range of metrics, including accuracy, F1 score, area under the ROC curve, and confusion matrix. We also closely analyzed models to identify critical features that contributed most significantly to classification outcomes.

Methods: Feature Categories

Demographic
Population factors such as age and language spoken.

Social Contact
Social contact is defined by two criteria: distance between 0.2-2 meters and orientations less than 45 degrees from face-to-face [5] (Fig. 1). Features are extracted from this definition.

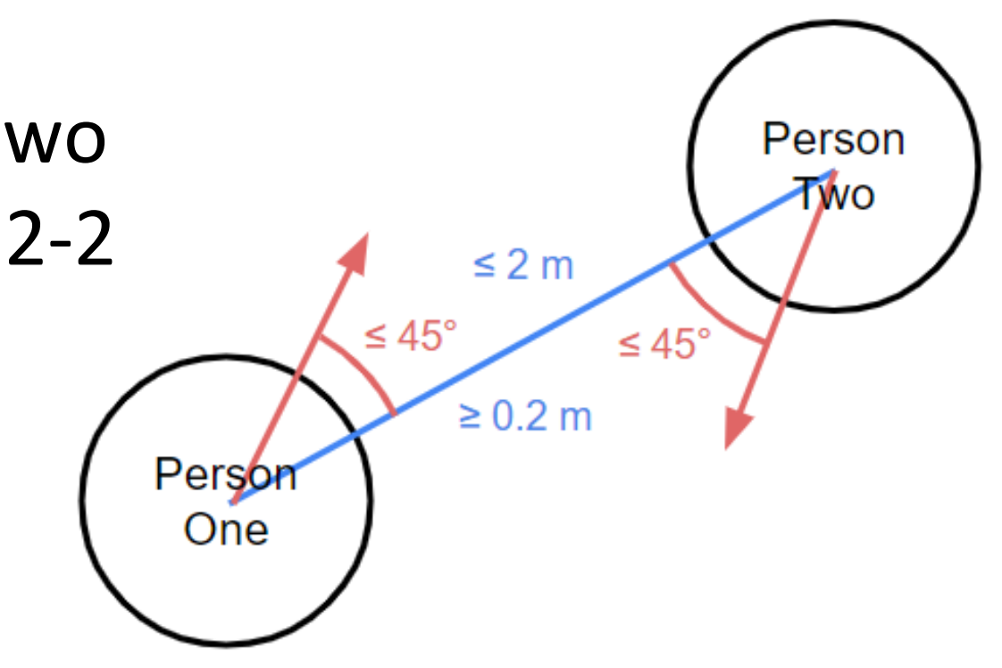

Figure 1. Social Contact Diagram

Social Contact Approach Velocity
When there is contact, the speed of approach is calculated with the temporal derivative of the distance between two individuals using the finite difference method. Features are extracted from this attribute.

Proximity
Features that stem from being in proximity: distance between individuals is less than 1 meter [6].

Speech
Basic metrics of speech such as duration and volume.

Additional Materials

Movement
Measures of the variability in an individual's movements derived from location, orientation and time.

Results

After evaluating all the models, Logistic Regression and Ridge Regression proved to be the most successful. Logistic Regression was able to achieve the highest overall accuracy of 0.822, F1 score of 0.747, and ROC AUC of 0.808 (Fig. 2). However, Ridge Regression performed better when classifying the positive class (ASD), achieving an accuracy of 87.84% (Table 2). We were also able to identify that features related to speech, social approach velocity, and proximity were the primary contributors to the best classifications (Fig. 3).

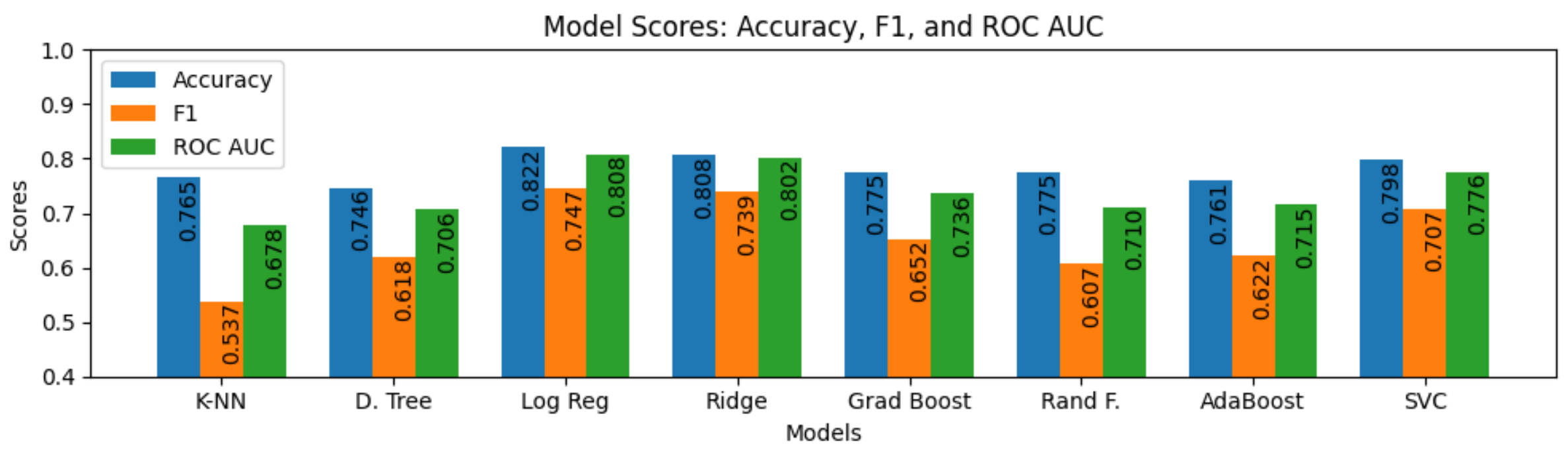

Figure 2. Evaluation Metrics for All Models

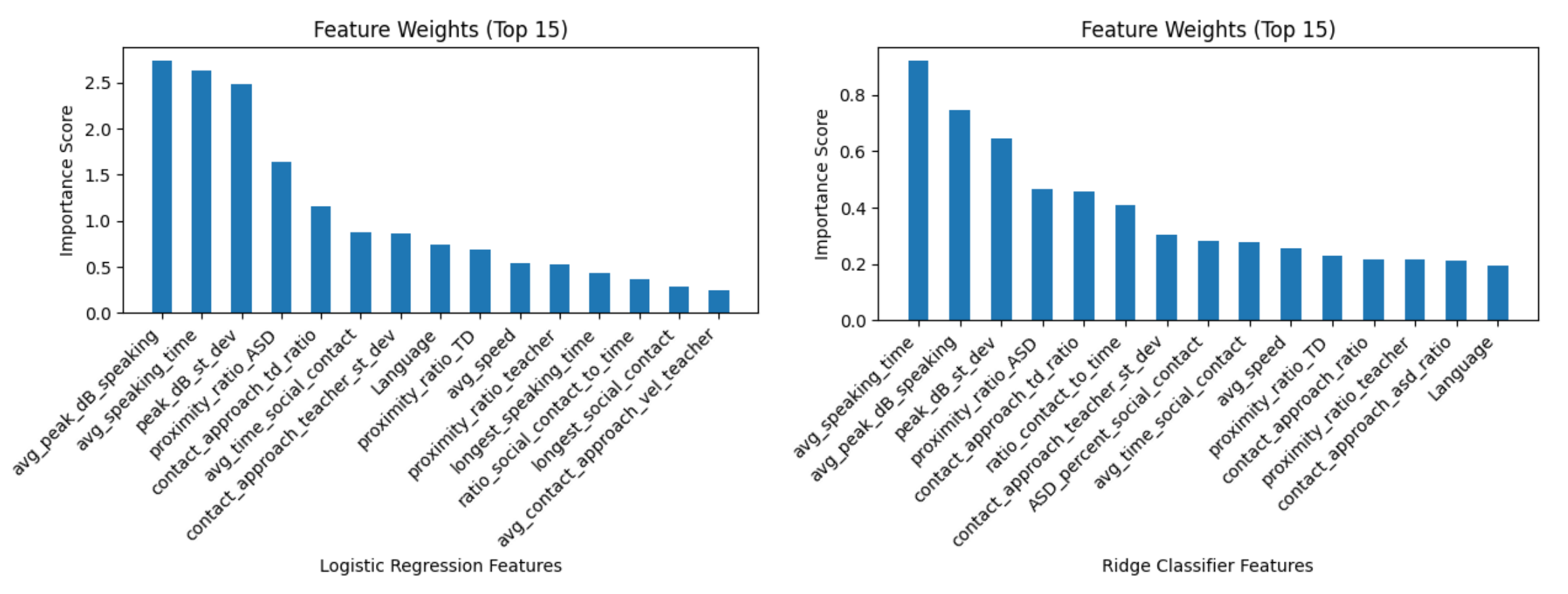

Figure 3. Feature Weights for Logistic Regression (left) and Ridge Classifier (right) with Entire Observations

	Negative (TD)	Positive (ASD)
Negative (TD)	112 (80.58%)	27 (19.42%)
Positive (ASD)	14 (18.92%)	60 (81.08%)

Table 1. Confusion Matrix For Logistic Regression

	Negative (TD)	Positive (ASD)
Negative (TD)	101 (72.66%)	38 (27.34%)
Positive (ASD)	9 (12.16%)	65 (87.84%)

Table 2. Confusion Matrix For Ridge Regression

Conclusion

Our research demonstrates the effectiveness of statistical ML models for autism classification. With only aggregate statistics derived from location and audio data, our ML models were able to accurately predict ASD with 87.84% accuracy.

Furthermore, our investigation of weights (standardized coefficients) enabled us to gain valuable insights. Specifically, we found evidence suggesting the use of decibels as a marker for autism. Additionally, we discovered that a difference in approach velocity to social contact has the prospect to be a new characteristic of ASD. Moreover, the precedence of proximity indicates a possible need to reevaluate the orientation restriction in the definition of social contact.

The implications of our findings extend far beyond classification. Due to the interpretability of statistical ML, we are able to analyze the “why” and “how” a model makes a prediction. In essence, we proved the capacity of ML to not just assist in diagnosis, but also accelerate our understanding of autism.

Future Work

Crucially, it is necessary to expand our dataset to other classrooms to assess the generalizability of the models. There is also potential to improve accuracy through optimizing current features, extracting new attributes, and integrating more data types.

On a broader scale, unsupervised learning can be employed to find underlying patterns and clusters. RNNs and deep learning can also explore more complex behaviors such as circling and following.

References

(1) van Heijst, B. F., & Geurts, H. M. (2015). Quality of life in autism across the lifespan: A meta-analysis. Autism, 19(2), 158–167

(2) Siddiqui S et al. Food for Thought: Machine Learning in Autism Spectrum Disorder Screening of Infants. Cureus. 2021 Oct 12;13(10):e18721. doi: 10.7759/cureus.18721. PMID: 34790476; PMCID: PMC8584605.

(3) LENA SP Digital Language Processor, LENA Foundation.

(4) DIMENSION4 UWB RTLS, Ubisense Group Plc.

(5) Banarjee, C et al. Objective quantification of homophily in children with and without disabilities in naturalistic contexts. Sci Rep 13, 903 (2023).

(6) Messinger, D. S et al. (2019). Continuous measurement of dynamic classroom social interactions. International Journal of Behavioral Development, 43(3), 263–270.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1949972. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

National Science Foundation

1 Department of Computer Science, University of Miami, Coral Gables, FL, USA
2 Department of Psychology, University of Miami, Coral Gables, FL, USA
3 Linda Ray Intervention Center, University of Miami Miller School of Medicine, Miami, FL, USA
4 School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
5 Department of Computer and Electrical Engineering, Ana G. Méndez University Gurabo, PR, USA