Diagnosing severity levels of Autism Spectrum Disorder with Machine Learning

Marcello Cinque
University of Naples Federico II
marcello.cinque@unina.it

Vincenzo Moscato
University of Naples Federico II
vmoscato@unina.it

Marco Postiglione
University of Naples Federico II
marco.postiglione@unina.it

Maria Pia Riccio
Federico II Hospital of Naples
mariapia.riccio@unina.it

Abstract

Autism Spectrum Disorders (ASDs) are neurodevelopmental disorders which inhibits linguistic, cognitive, communication and social skills of affected individuals. Currently, ASD is diagnosed by means of time-consuming and expensive screening tests. Hence, Machine Learning (ML) techniques have been applied to construct predictive models able to diagnose autism at early stages. However, the binary setting (ASD vs not-ASD) and the not-exciting performance reached by such models highlight the need for further de-identified datasets and interdisciplinary work linking computer scientists and Subject Domain Experts (SDEs). In this work, we propose a novel dataset in which labels refer to the severity level of autism as required by the *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* standard reference. Then we analyze the quality of resulting ML models (i.e. Random Forest, XGBoost, Neural Network) based not only on their performance metrics (i.e. precision, recall, F1) but also on the most important features they consider for classification and their similarity with the ones suggested by the SDE.

1 Introduction

According to the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)* (1), Autism Spectrum Disorders (ASDs) are neurodevelopmental conditions which affect communication and behavior and are characterized by impairment of communication and social interaction and by repetitive, restricted and stereotyped interests. The prevalence has increased exponentially in the last few years: most recent estimates suggest a prevalence of 1 in 59 among 8-year-old children from the USA (2). Both genetic and environmental factors can contribute to the pathogenesis, but the etiology is still unknown. To date, ASD is clinically diagnosed by means of behavior-based tests which are costly and time-consuming (3). Hence, the automation of the diagnosis by means of ML approaches would be a great advance in this healthcare field.

Wall et al. (4) are the firsts to apply Machine Learning (ML) approaches for ASD diagnosis. On paper, their results were extremely promising: they obtained 100% accuracy and sensitivity by using only 8 out of the 29 original features. Similarly, Wall et al. (5) show that 7 of the 93 answers resulting from the *Autism Diagnostic Interview-Revised (ADI-R)* were sufficient to classify autism with 99.9% accuracy. However, a subsequent research work from Bone et al. (6) has shown the problems of the above-mentioned works: models were trained on highly unbalanced data and test instances did not contain any negative sample, and test instances were simulated by picking samples from the training set to deal with the lack of data. As a consequence, the absence of clinical domain expertise led to erroneous conclusions, since all the discarded were necessary to obtain a reliable autism score.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Results obtained in recent literature are not exciting (7; 8) and thus suggest the need for additional efforts in this field to use ML models in the actual practice of medicine.

In addition to all of the above, most of the research works focus on a binary classification task (9), i.e. *ASD vs non-ASD* classification, and ignore the numerous facets of ASD which can be decomposed in different classes. However, ASDs are very heterogeneous conditions and symptoms vary widely from subject to subject, hence the term "spectrum". Therefore, according to DSM-5 criteria, clinical severity levels — based on social communication impairments and restricted and repetitive behavioral patterns — can be identified as follows:

- *Level 1: requiring support*. difficulties and possible decreasing interest in social interactions; inflexibility of behavior, difficulties in switching between activities, problems in planning and organizing.
- *Level 2: requiring substantial support*. Impairments in verbal and non-verbal communication skills evident also with supports in place; inflexibility of behavior, difficulty coping with change, evident and interfering restricted/repetitive behaviors, difficulty changing focus or action.
- *Level 3: requiring very substantial support*. Severe communication deficits; inflexibility of behavior and extreme difficulties in restricted/repetitive behaviors, great difficulty changing focus or action.

The binary problem simplification can surely lead to higher levels of sensitivity and specificity, but the resulting ML models are not useful in the actual practice of medicine, where there is a grey area and patients with different levels of disease should receive different treatments.

To sum up, the primary objective of previous studies was to ease physicians in the screening process of patients, thus focusing on performance metrics such as accuracy, specificity and sensitivity, even at the expense of discarding features which could have brought to useful insights (4; 5; 10). The analysis of the current literature in the application of ML techniques to ASD diagnosis reveals the need for a strengthened cooperation between clinical and computational researchers, which could lead to useful insights which can actually improve the clinical processes.

In this work, according to the guidance of a *Subject Domain Expert (SDE)*, we publicly release *Autism Spectrum Disorder - Severity Levels (ASD-SL)*, a multi-class ASD dataset which is compliant with DSM-5 criteria[1]. Furthermore, not only do we analyze the quality of resulting ML models with quantitative performance metrics (precision, recall and F1 scores), but we also use SHAP (11) to compute the importance assigned by each model to each feature and compare the results with the knowledge and experience of the SDE.

2 ASD-SL Dataset

We collected anonymized data from a consecutive sample of children referred to Department of Pediatrics - Unit of Child and Adolescent Neuropsychiatry, University Federico II of Naples, for an evaluation in a clinical suspicion of ASD. About 141 individuals (76,5% males), aged between 18 to 156 months, received a full assessment, including historical information, structured clinical interviews and validated observations. Autism Diagnostic Observation Schedule-2 (ADOS-2) was performed by a licensed clinician both to confirm diagnosis and to evaluate level of symptoms according to comparative score. To determine the development/intellective level, Griffiths Mental Development Scale (GMDS-ER) or Leiter International Performance Test-Revised (Leiter-R) were administered. To establish adaptive competence of all patients, parents were interviewed by Vineland Adaptive Behavior Scales – II edition (VABS-II). Diagnosis of ASD was formulated according to DSM-5. We classified all the specifiers useful to determine the severity level of ASD according to the DSM-5 diagnostic criteria: "With/without accompanying language impairment"; "With/without accompanying intellectual impairment". We integrate data about environmental factors, genetic factors, cognitive/social/language impairments and useful measures which are usually leveraged to diagnose ASD:

- *Intellective Quotient (IQ)/Developmental Quotient(DQ)*: usually measured by means of intelligence or developmental tests (e.g. Wechsler scales, Leiter international performance

[1] `https://picuslab.dieti.unina.it/index.php/asd-sl`

Table 1: Dataset features

Feature	Description	Type	# missing
age_months	numbers indicating the age	int	37
gender	and gender of patients	binary	0
pregnancy_problems	flags indicating environmental factors	binary	39
normally_evolved_perinatal_phenomena		binary	39
birth_anomalies		binary	47
psychiatric_disorders_familiarity		binary	0
QS (developmental age)	useful measures (see Section 2)	int	43
IQ		float	109
DQ		float	71
DQ_IQ		float	39
QA_VABS		int	4
ADOS		int	5
I_intellective_impairment	specifiers of impairments and comorbidities	binary	118
II_language_impairment		binary	46
III_known_medical_condition		binary	56
III_history_environmental_exposure		binary	54
IV_other_mental_behavioral_disorders		binary	55
other_psychiatric_comorbidities		binary	70
nutrition_disorders		binary	17
CGH_array_alterations	genetic factors	binary	45
n_altered_chromosomes		int	45
n_mutations		int	45
n_dup		int	45
n_del		int	45

scale, Griffiths III Mental Development scales), they provide indications on the presence (and the level) of an intellectual/developmental impairments. Determining this value allows to determine the ASD specifier "With/without associated intellectual impairment".

- *Adaptive Quotient (QA)*: it indicates the level of "adaptive behavior", i.e. how effectively a person copes with common life demands and how well they meet the standards of personal independence expected of someone in their particular age group, socio-cultural background and community setting.
- *Autism Diagnostic Observation Schedule (ADOS)*: it measures the responses to standard, social, planned activities (stimuli) that favor communication skills and social interaction. This instrument is the gold standard to confirm ASD diagnosis and it allows to evaluate level of symptoms according to comparative score (ranging from 0 to 10).

Table 1 shows an overview of the complete set of features, with their description, type and the number of missing values.

3 Quality Analysis

In healthcare projects, the standard ML pipeline (data collection - modeling - error analysis) has to be enriched with an interface which allows SDEs to understand the inner mechanisms of ML models. In this perspective, not only do we quantitatively evaluate the performance of three different ML models, but we also compare their SHAP explanations to the ones provided by the SDE.

We trained three different ML models: Random Forest, XGBoost and a neural network. We test their performance with a 10-fold cross-validation and choose the best hyper-parameters for each fold with a 10-fold randomized search. Table 2 reports the results in terms of precision, recall and F1 scores. The high variability of performance highlights the need for further analyses to establish the quality of models.

Table 2: Training results.

Level	Model	Precision	Recall	F1
1	Random Forest	55.00 $\pm$47.17	46.67 $\pm$ 42.03	48.00 $\pm$ 40.86
	XGBoost	65.83 $\pm$ 36.60	65.00 $\pm$ 32.02	60.00 $\pm$ 28.09
	Neural Network	55.83 $\pm$ 32.50	65.83 $\pm$ 36.60	54.67 $\pm$ 28.84
2	Random Forest	63.57 $\pm$ 23.64	79.79 $\pm$ 16.94	68.35 $\pm$ 20.92
	XGBoost	64.31 $\pm$ 23.47	77.17 $\pm$ 13.24	67.13 $\pm$ 18.69
	Neural Network	66.64 $\pm$ 25.35	67.89 $\pm$ 18.13	63.60 $\pm$ 19.97
3	Random Forest	73.17 $\pm$ 12.94	66.44 $\pm$ 14.60	67.93 $\pm$ 07.72
	XGBoost	67.83 $\pm$ 23.29	56.06 $\pm$ 23.07	56.99 $\pm$ 20.67
	Neural Network	55.50 $\pm$ 29.21	53.23 $\pm$ 30.02	51.44 $\pm$ 25.95

Table 3: Global importance rankings

Feature	SDE	Random Forest	XGBoost	Neural Network
QA_VABS	1	6	7	3
QS	2/3/4	7	11	4
IQ	2/3/4	21	22	18
DQ_IQ	2/3/4	20	15	5
ADOS	5	1	2	1

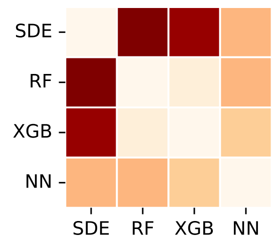

Figure 1: Rank distances

We asked a SDE to rank the most useful features he would consider when diagnosing the level of ASD. He answered by highlighting three levels of importance and five features: (1) first of all, the *QA_VABS* value is of primary importance, since the severity level of autism is established on the basis of the support level needed in the two areas related to the core symptoms of ASD: social communication and restricted/repetitive behaviors; (2) DQ and IQ have a smaller impact since it is possible that a patient with good levels of cognitive development has not established adequate autonomy skills and thus requires a significant support; (3) the *ADOS* scores express the level of autism symptoms: it is useful but it cannot be considered alone since patients with few symptoms could have a significant lack in autonomy.

We used SHAP to compute the global importance assigned by ML models to each feature to measure how their inner workings come towards or away from the SDE knowledge and experience. Ranks of the main features are reported in Table 3: while the ADOS score is considered as an important attribute for classification by the whole set of models, the only one that seems to "agree" with the SDE on the other features is the neural network, which assigns similar ranks with except to the *IQ* feature (but we expected that because *DQ_IQ* has been generated from the aggregation of *DQ* and *IQ* attributes). The matrix of Euclidean distances between ranks in Figure 1 shows that while the Random Forest and XGBoost algorithms provide similar ranks, they completely differ from SDE knowledge.

4 Conclusion & Future Work

In this work, in order to facilitate the research community to diagnose severity levels of ASD according to DSM-5 criteria, we publicly release a dataset of children affected by the disease. Past literature and our experimental results highlight the need for collaborative work between ML and clinical researchers. We observed that quantitative metrics are not always sufficient to choose between different ML solutions (especially in few-shot scenarios, where we are not able to establish significant differences between models), and that interpretability-based comparisons — in conjunction with SDE knowledge and experience — may constitute a valuable support for the decision.

References

[1] V. A. Mittal and E. F. Walker, "Diagnostic and statistical manual of mental disorders," *Psychiatry Research*, vol. 189, pp. 158–159, 2011.

[2] M. J. Maenner, K. A. Shaw, J. Baio, A. Washington, M. E. Patrick, M. Dirienzo, D. L. Christensen, L. D. Wiggins, S. Pettygrove, J. G. Andrews, M. Lopez, A. Hudson, T. Baroud, Y. Schwenk, T. White, C. R. Rosenberg, L. C. Lee, R. A. Harrington, M. Huston, A. S. Hewitt, A. N. Esler, J. A. Hall-Lande, J. N. Poynter, L. Hallas-Muchow, J. N. Constantino, R. T. Fitzgerald, W. M. Zahorodny, J. Shenouda, J. L. Daniels, Z. Warren, A. C. Vehorn, A. Salinas, M. S. Durkin, and P. Dietz, "Prevalence of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2016," *MMWR Surveillance Summaries*, vol. 69, pp. 1 – 12, 2020.

[3] M. Galliver, E. Gowling, W. J. Farr, A. Gain, and I. Male, "Cost of assessing a child for possible autism spectrum disorder? an observational study of current practice in child development centres in the uk," *BMJ Paediatrics Open*, vol. 1, 2017.

[4] D. Wall, J. Kosmicki, T. DeLuca, E. Hårstad, and V. Fusaro, "Use of machine learning to shorten observation-based screening and diagnosis of autism," *Translational Psychiatry*, vol. 2, 2012.

[5] D. Wall, R. Dally, R. J. Luyster, J.-Y. Jung, and T. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," *PLoS ONE*, vol. 7, 2012.

[6] D. Bone, M. S. Goodwin, M. Black, C.-C. Lee, K. Audhkhasi, and S. S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of Autism and Developmental Disorders*, vol. 45, pp. 1121–1136, 2015.

[7] M. Parikh, H. Li, and L. He, "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data," *Frontiers in Computational Neuroscience*, vol. 13, 2019.

[8] K. S. Oma, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A machine learning approach to predict autism spectrum disorder," *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6, 2019.

[9] F. Thabtah, "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward," *Informatics for Health and Social Care*, vol. 44, pp. 278 – 297, 2018.

[10] M. Duda, R. Ma, N. Haber, and D. Wall, "Use of machine learning for behavioral distinction of autism and adhd," *Translational Psychiatry*, vol. 6, 2016.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf