Exploring Machine Learning for Early Autism Spectrum Disorder Prediction Using Umbilical Cord Blood Gene Expression

Laura G. Speggiorin[1,2], **Thayne W. Kowalski**[2,3,4], **Mariana Recamonde-Mendoza**[1,2]

[1]Institute of Informatics – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[2]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA)
Porto Alegre – RS – Brazil

[3]Graduate Program in Genetics and Molecular Biology, Genetics Department,
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[4]Medical Genetics Service, Hospital de Clínicas de Porto Alegre (HCPA)
Porto Alegre – RS – Brazil

lgspeggiorin@inf.ufrgs.br, tkowalski@hcpa.edu.br, mrmendoza@inf.ufrgs.br

Abstract. *This study presents a proof-of-concept machine learning (ML) model for early Autism Spectrum Disorder (ASD) prediction using transcriptomic data from umbilical cord blood. We analyzed 224 samples (53 ASD, 80 with non-typical development [Non-TD], 91 with typical development [TD]) from high-risk cohorts, proposing a two-step classification pipeline based on eight distinct algorithms and ensemble approaches. The first model (TD vs. Non-TD/ASD) achieved an F1.5-score of 0.89 and recall of 1.0; the second model (ASD vs. Non-TD) yielded 75% accuracy and an F1.5-score of 0.54. Results suggest subtle, yet detectable, transcriptomic signals in perinatal blood that may support early ASD risk stratification, warranting further investigation in larger cohorts.*

1. Introduction

An increase in the diagnosis of Autism Spectrum Disorder (ASD) has been observed in recent years, with studies mentioning it could be as prevalent as 1 in 36 people [Zablotsky et al. 2017]. ASD is a group of neurodevelopmental disorders characterized by impairment in social communication and interaction as well as repetitive patterns of behaviors. It often co-occurs with other conditions, like attention deficit hyperactivity disorder (ADHD), as well as anxiety, depression, and epilepsy [Lord et al. 2020], leading to additional physical or mental challenges, higher treatment costs, and increased demands on families [Sharma et al. 2018].

Some of these challenges can be mitigated through early interventions, which also underscores the importance of early diagnosis. However, the heterogeneous origin of ASD is not completely understood. It is known that the occurrence of the disorder is influenced by both genetic and environmental factors, making it a highly heritable disorder. Transcriptomic data, which reflects both influences, has shown promise for investigating such complexity [Tylee et al. 2017]. Previous studies have reported expression alterations in immune and neuronal pathways [Hodges et al. 2020], supporting the relevance

of blood and brain tissue in ASD research. These findings have fostered increasing interest in genomic approaches to investigate the causes and consequences of ASD, with the goal of deepening our understanding of its molecular basis and exploring complementary strategies to enable earlier diagnosis.

The most common form of ASD is multifactorial, involving the contribution of hundreds of gene variants. As a result, diagnosis is still based on standardized neurodevelopmental assessments conducted by professionals. Some machine learning (ML) studies have proposed the use of supervised learning to learn complex underlying patterns in ASD through classification algorithms, essentially offering a preliminary diagnosis based on some specific types of data [Omar et al. 2019, Liu et al. 2016, Zhang et al. 2018]. The combination of ML algorithms with gene expression data has shown promising results, although most studies to date rely on blood samples collected during toddlerhood.

In contrast, this work explores the use of genome-wide gene expression profiles from umbilical cord blood collected at birth, capturing genetic and early environmental factors during a highly sensitive developmental window. We aim to investigate the viability of using ML models to identify higher ASD risk in infants based on this data. We propose a two-step ML model to support early ASD prediction: first, classifying samples as typical development (TD) or not; then, among non-typical cases (Non-TD), identifying ASD. Ensemble learning and dimensionality reduction were used to address class imbalance and high dimensionality. Despite moderate performance, the results suggest that cord blood transcriptomics can contribute to early ASD risk stratification.

2. Materials and Methods

This section presents the methodological details of our work. We begin by describing the dataset and the preprocessing steps used to address the challenge of high dimensionality. We then outline the model development pipeline, covering data partitioning, model training, and performance evaluation. We present our two-step classification strategy, which integrates multiple well-established algorithms through an ensemble approach.

2.1. Dataset description and dimensionality reduction

The collected data consisted of genome-wide transcript levels measured in umbilical cord blood samples from two early ASD studies: the Early Autism Risk Longitudinal Investigation (EARLI) [Newschaffer et al. 2012] and the Markers of Autism Risk in Babies – Learning Early Signs (MARBLES) study [Hertz-Picciotto et al. 2018]. Both studies recruited families with an older child previously diagnosed with ASD. In each case, umbilical cord blood samples were collected at birth from younger siblings, who were later assessed at 36 months and classified, based on algorithm-derived diagnoses, as having ASD, Typical Development (TD), or Non-TD (neither ASD nor TD).

The dataset was initially pre-processed and aggregated by Mordaunt et al. [Mordaunt et al. 2019], and is publicly available under accession number GSE123302. Preprocessing steps included assessing signal distribution within each study, quality control, normalization, probe annotation and filtering, followed by surrogate variable analysis to correct for substantial effects on gene expression.

From the original 271 preprocessed samples, we used 224 for which consent for data sharing was granted. The retained dataset included gene expression levels, probe

identifiers (used as feature names), anonymized patient identifiers, and diagnostic labels (used as the target variable). The initial dataset contained 36,459 features (gene expression probes) and 224 samples, distributed as 53 ASD, 80 Non-TD, and 91 TD. Although a class imbalance was observed, it was not considered severe enough to require data augmentation or resampling techniques. Ethical review was not required for this study, since it uses anonymized data with informed consent obtained during the original collection.

Due to the high dimensionality of transcriptomic data and the relatively small sample size, which can hinder generalization in ML models, dimensionality reduction was a necessary step. We first filtered the features based on metadata from the sequencing platform "[HuGene-2_0-st] Affymetrix Human Gene 2.0 ST Array [transcript (gene) version]". Using the "GB_ACC" column from the platform table, we retained only probes annotated with the prefix "NM", indicating their association with mRNA. This step alone reduced the number of features by 65%, resulting in 12,896 columns. We then ranked the remaining features by variance and selected the top 5,000 (highest variance) for downstream analysis. Thus, the final dataset used in the experiments consisted of 224 samples and 5,000 features. Features were standardized using Scikit-learn's StandardScaler.

2.2. Model development

The methodology for developing the model involved several key steps to prepare the dataset and optimize classifier performance. The core of this work is a two-step ML classifier in order to better learn from the scarce data. The proposed architecture is illustrated in Figure 1: the first step aims to distinguish samples with typical development (TD) from those with atypical development (Non-TD and ASD), while the second step focuses on identifying ASD cases among the non-TD group.

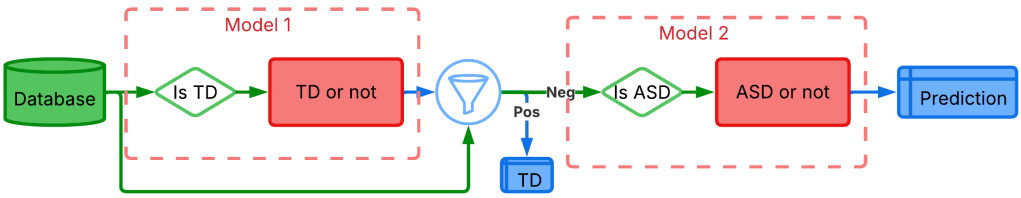

Figure 1. Architecture of the proposed classifier. In green, the dataset transformed for each classification task. In red, the models corresponding to each step (left: *Model 1*; right: *Model 2*), with the output of the first used to filter the input to the second by retaining only instances not predicted as TD. In blue, the final prediction result.

For each classification task, illustrated as a red square in the figure, an ensemble approach was employed, combining eight distinct machine learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LightGBM). This improves robustness, taking advantage of each algorithm's learning capabilities.

Each model was trained on transformed versions of the original three-class target. For the first step, samples labeled as TD were assigned to class 0, while both Non-TD

and ASD were grouped as class 1. For the second step, TD and Non-TD were grouped as class 0, and ASD was labeled as class 1. Although TD samples are not expected in the second step, misclassifications from the first step may still carry them forward, in which case they are simply treated as non-ASD.

After selecting the best hyperparameters, all models were trained on the entire training set, generating eight independent predictions per sample. These outputs were combined using a soft voting strategy, which aggregates the probabilities of the predicted targets by averaging them. Soft voting was chosen for two main reasons: to reduce the chance of ties, since the number of classifiers is even, and to allow classifiers with higher confidence in certain predictions to contribute more effectively. This strategy leverages the strengths of different algorithms, with the potential to generate more robust final predictions.

2.3. Evaluation strategy

The dataset was first divided using an 80/20 stratified holdout, maintaining class proportions in both training and testing sets. This resulted in 179 training samples and 45 testing samples. Model training employed 5-fold stratified cross-validation (CV) on the training set, with a nested 3-fold CV for hyperparameter tuning. Grid search was conducted using algorithm-specific hyperparameter lists, applied identically to both classification steps.

Hyperparameter selection across folds followed these criteria: the most frequent value was chosen for categorical parameters, the median for integers, and the mean for real-valued parameters. Once the optimal hyperparameter configuration was selected for a given algorithm, the model was fitted to the entire training set and subsequently evaluated on the held-out test set.

The F-beta score, with $\beta = 1.5$, was used as the optimization metric during model selection. This metric was chosen due to the class imbalance, particularly the low proportion of ASD cases. Prioritizing both precision and recall, with a higher weight on recall, helps prevent models from defaulting to negative predictions (i.e., predicting most or all cases as non-ASD). To thoroughly assess model performance, multiple metrics were computed at each stage: accuracy, precision, recall, F1 score, F-beta score, and ROC AUC (area under the receiver operating characteristic curve). The results or the metrics not reported here can be seen in the project's repository[1] alongside other supplementary material like the source code and the hyperparameters tested for each algorithm.

3. Results

This section presents and discusses our results. Throughout the text, we refer to the first part of the classification task (i.e., distinguishing TD from non-TD) as *Model 1*, and the second part (i.e., distinguishing ASD from non-ASD within the non-TD group) as *Model 2*. The performance analysis for the grid search conducted during hyperparameters optimization is shown in Figure 2 for *Model 1*, and Figure 3 for *Model 2*. We summarize the mean and standard deviation of the best F-beta score for each individual classifier in the grid search. The individual score at this point of the training is not that illustrative of the final performance, since it was executed on a small amount of data, but was used to

[1] https://github.com/LSpeggiorin/TCC

better tune the models to their respective tasks. Overall, we observe that *Model 2* tends to exhibit greater variability in performance (i.e., higher standard deviations) and generally achieves lower mean F-beta scores compared to *Model 1*.

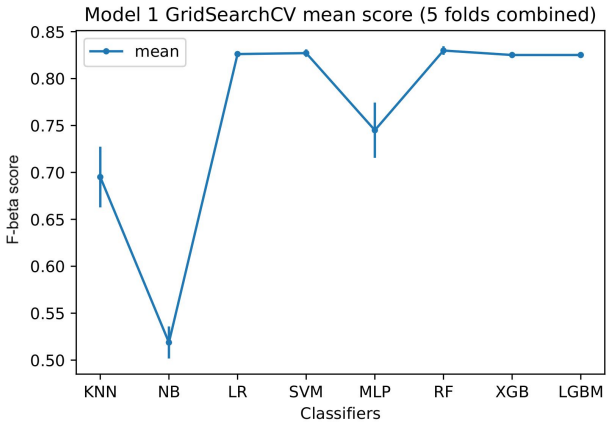

Figure 2. Summary of Model 1 grid search: mean (dots) and standard deviation (bars) of F-beta scores across classifiers in 5-fold CV.

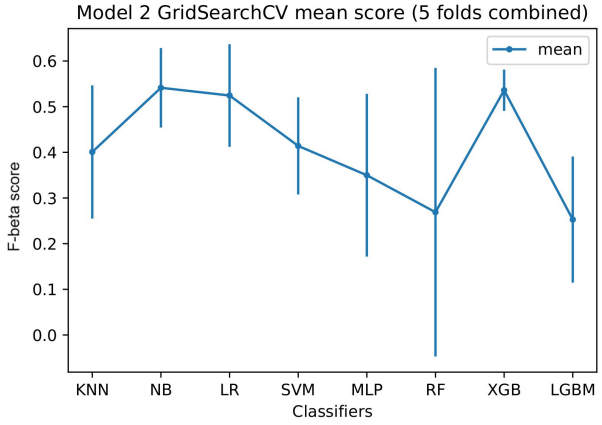

Figure 3. Summary of Model 1 grid search: mean (dots) and standard deviation (bars) of F-beta scores across classifiers in 5-fold CV.

To better understand model performance across the 5-fold cross-validation, we examined the variation in F-beta scores across the different Train/Validation splits. The mean and standard deviation of the F-beta score for both the training and validation folds are shown in Figure 4 for *Model 1* and Figure 5 for M*odel 2*. In addition to the evaluation of individual classifiers, we also show the results for the ensemble-based approach using soft voting (rightmost column). Similar plots were generated for other metrics, but are not shown due to space constraints. However, the overall pattern reveals that *Model 1* exhibits greater robustness, while Model 2 shows a tendency to overfit for several algorithms.

Analyzing the grid search and cross-validation results, we observe that *Model 1* performs well for most classifiers, with the exception of Naive Bayes and KNN. In contrast, for *Model 2*, XGBoost and Naive Bayes stand out as the most promising, considering

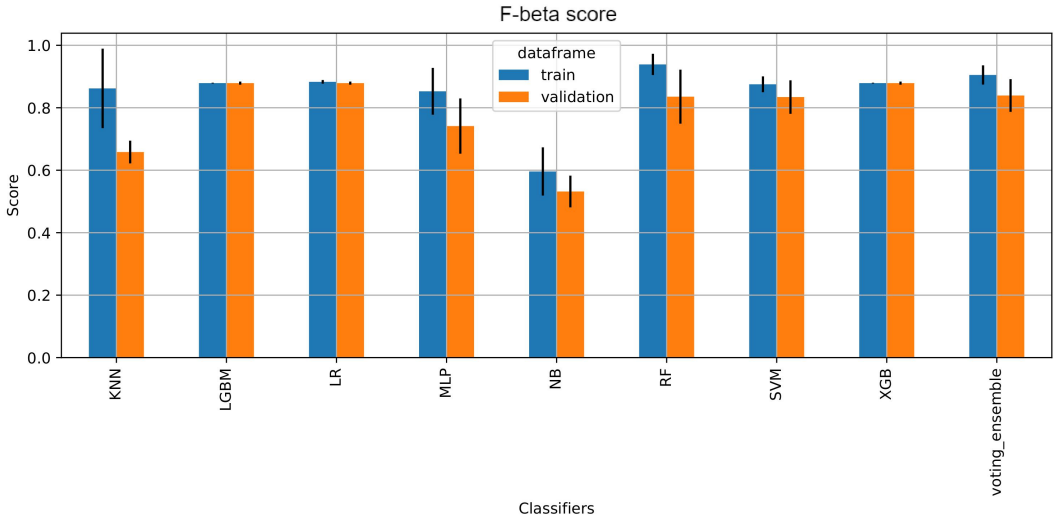

Figure 4. Mean and standard deviation of F-beta scores for each classifier in the cross-validation of Model 1, comparing training and validation performance, including the ensemble model.

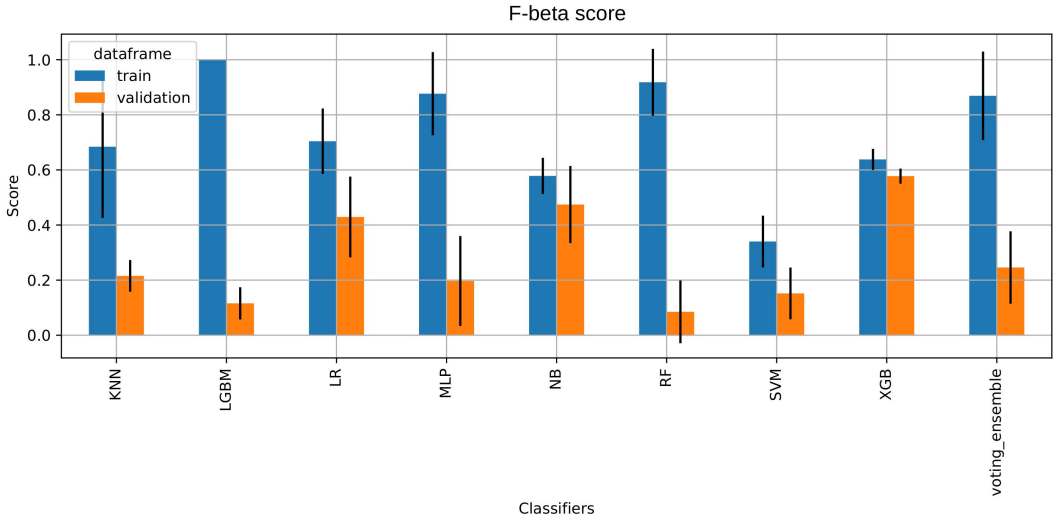

Figure 5. Mean and standard deviation of F-beta scores for each classifier in the cross-validation of Model 2, comparing training and validation performance, including the ensemble model.

both the higher means and the lower gap between train and test performances. Logistic regression also shows competitive performance. These differences highlight how distinct the two classification tasks are in terms of model behavior and learning requirements. Regarding the ensemble model, we note that it achieves strong performance in both classification steps, although signs of overfitting are still present in *Model 2*.

Because of the sequential nature of our two-step classifier, the final classification for ASD is provided by *Model 2*. Thus, we observe the performance of this model for the test set as compared to the train set in Table 1. Due to space constraints, we only show the results for predictions obtained from the ensemble model (i.e., based on the

Table 1. Performance metrics from Model 2 in the training and test sets.

	Accuracy	Precision	Recall	F1 score	**F-beta score**	ROC-AUC
Train	0.98	0.95	0.95	0.95	**0.95**	0.97
Test	0.75	0.5	0.55	0.52	**0.54**	0.68

Table 2. Confusion matrix of Model 2 in the test set.

		Predicted	
		Positive	Negative
True	Positive	6	5
	Negative	6	27

soft voting strategy), since we believe this model provides a good compromise between predictive power and variance. The final model achieved an F-beta score of 0.54 and an accuracy of 0.75, the highest among the reported metrics. The confusion matrix (Table 2) illustrates better the behavior of the model. The high accuracy is due to the True Negative Rate (TNR) of 82%, correctly predicting 27 out of the 33 negative cases. Given that in this problem 75% of all instances are negative, a TNR higher than True Positive Rate (TPR) is expected, which is one of the reasons for using F-beta score as the main metric, training the models to be less skewed to the negative class. This approach was successful, considering the 75% of instances predicted correctly are well divided between classes. The recall is 0.55, meaning 6 of the 11 positive cases were correctly predicted. In the end, given that a patient actually has ASD, the model correctly predicts 55% of these cases. However, we note a performance gap between the training and test datasets, which may indicate overfitting, possibly caused by the limited sample size.

Table 3, which presents the performance of *Model 1* based on the ensemble strategy, shows substantially higher overall performance and, interestingly, very similar results between the training and test sets. *Model 1* achieved an F-beta score of 0.89 and perfect recall, meaning that all Non-TD and ASD instances were correctly identified and passed on to *Model 2*. The precision of 0.61 is better understood by examining the confusion matrix in Table 4.

Since the main priority in *Model 1* was to ensure that all positive cases (Non-TD and ASD) were captured for further classification by *Model 2*, the model was intentionally biased toward the positive class. This was influenced by the use of the same F-beta configuration ($\beta = 1.5$) for both classification tasks, despite the differences in class distribution: in *Model 1*, the positive class represents 67% of the data, while in *Model 2*, it represents only 25%. As a result, *Model 1* reached a lower precision of 0.61 and an extremely low True Negative Rate (TNR) of 12%, introducing a larger proportion of false positives into Model 2. This mislabeling increases the noise in *Model 2*'s input and makes it more difficult for the second model to learn generalized patterns effectively.

Table 3. Performance metrics from Model 1 in the training and test sets.

	Accuracy	Precision	Recall	F1 score	**F-beta score**	ROC-AUC
Train	0.66	0.64	1.0	0.78	**0.90**	0.59
Test	0.62	0.61	1.0	0.76	**0.89**	0.53

Table 4. Confusion matrix of Model 1 in the test set.

		Predicted	
		Positive	Negative
True	Positive	27	0
	Negative	17	1

The predictions of both *Model 1* and *Model 2* are the result of eight individual models trained for each classification task, so it is also relevant to analyze the performance of each classifier separately. Figures 6 and 7 show the F-beta score for each trained classifier, for *Model 1* and *Model 2*, respectively. A pattern observed both in the overall performance metrics (visible in the figures as the "voting_ensemble") and in the individual classifier metrics is that *Model 1* presents very little difference between the training and test sets. This may be related to the model's strong prioritization of the positive class in both datasets. As shown in Figure 6, this behavior is consistent across most of the classifiers used.

Figure 7, on the other hand, shows that the large difference between training and testing performance in *Model 2* is mainly caused by some specific classifiers - KNN, MLP and NB, for example, presented a F-beta score of zero for predicting all cases as negative. Even so, the ensemble model still outperformed most individual classifiers, showing stable behavior overall. The two classifiers that achieved the highest F-beta scores were the ensemble methods XGBoost and LightGBM, with RF, another ensemble method, also among the top performers. This suggests that the data may be too noisy for simpler models, or that this specific problem may require a more tailored architecture composed of a specific subset of classifiers in order to reach precise and robust results.

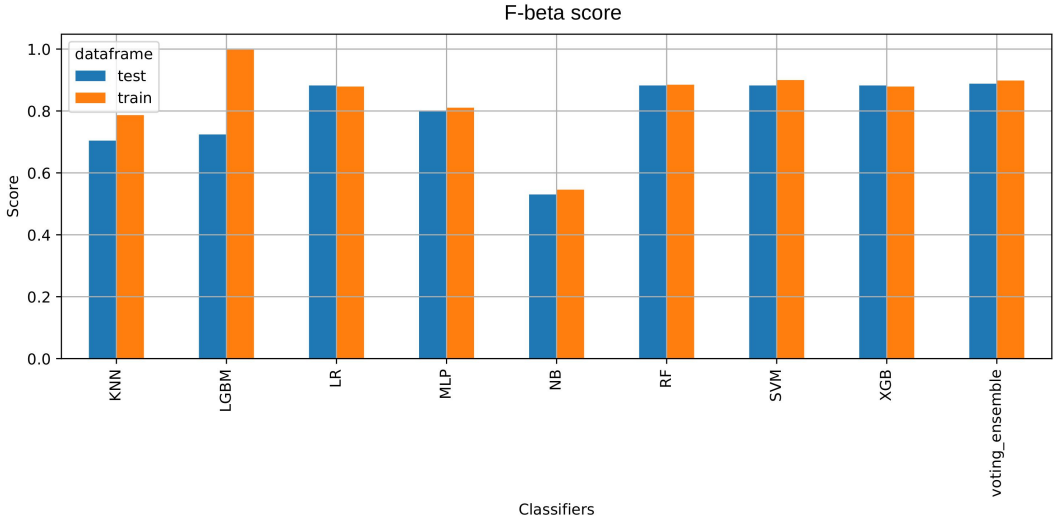

Figure 6. F-beta scores of individual and ensemble classifiers on training (orange) and test (blue) sets, for Model 1.

Interestingly, LightGBM was one of the worst-performing classifiers during cross-validation, suggesting it may require more data to learn effectively. On the other hand, Naive Bayes showed strong performance during cross-validation but achieved an F-beta

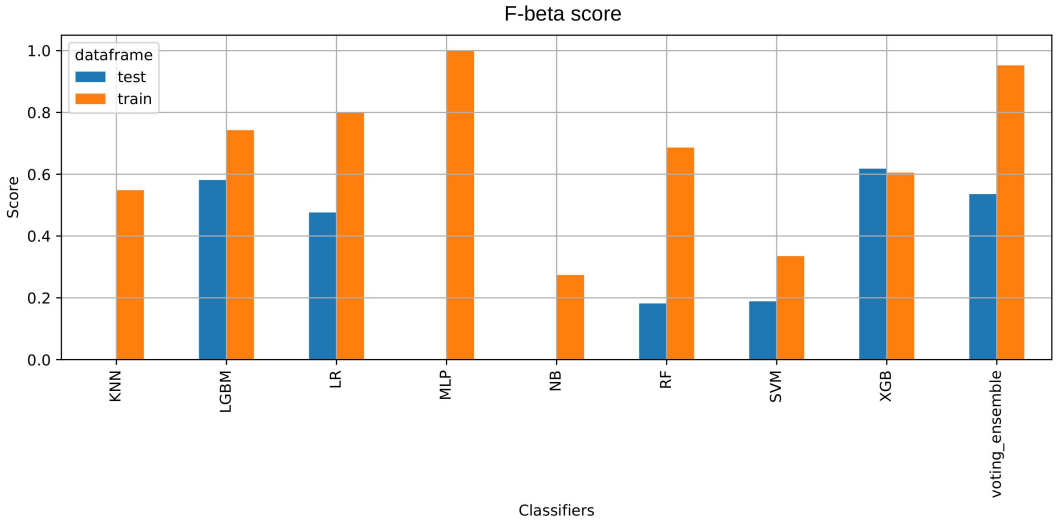

Figure 7. F-beta scores of individual and ensemble classifiers on training (orange) and test (blue) sets, for Model 2.

score of zero on the test set, possibly due to predicting all instances as negative. The ensemble approach was designed to take advantage of each classifier's strengths while mitigating their weaknesses. As seen in Table 1, the final ensemble model performed reasonably well across all metrics when compared to the individual classifiers.

Two additional models worth highlighting are LR, which outperformed RF and was the only other classifier to reach an F-beta score comparable to the ensemble, and SVM with a sigmoid kernel, which achieved the lowest non-zero F-beta score. The only classifiers that managed to identify at least some ASD cases were the ensemble tree-based models and those based on logistic functions.

Finally, to elaborate in the overall behavior of the proposed predictive model taking into account the main problem, and not its parts, Table 5 presents a view of the final classification of each instance from the test set as a multiclass confusion matrix. It can be seen that 55% of ASD subjects were correctly classified, and the other 45% was classified as Non-TD. Interestingly, only 17% of instances classified as ASD were actually TD, meaning most patterns found in *Model 2* may not be exclusive to ASD, but are consistent with atypical development. Of the Non-TD, 75% was correctly implicitly assigned, by not being classified as TD in *Model 1* or ASD in *Model 2*. The focus of this work is exclusively to aid in the diagnosis of ASD, so this classification of Non-TD should not be taken as a pre-diagnosis of any sort, specially considering that 71% of all subjects were predicted as being Non-TD, including 15 out of the 18 TD cases.

Table 5. Confusion matrix for the test set, considering the multiclass problem.

		Predicted		
		ASD	Non-TD	TD
True	ASD	**6**	5	0
	Non-TD	4	**12**	0
	TD	2	15	**1**

4. Discussion

Given the increasing prevalence of ASD and recent findings on its complex etiology, many studies have focused on understanding its genetic and environmental components. This work proposed a two-step ML classifier using gene expression data from umbilical cord blood as a potential tool to support early ASD diagnosis.

The overall performance of the model, particularly in the second classification step, was modest. Only 55% of ASD cases were correctly predicted, and among the predicted ASD cases, just 45% were true positives. Despite these limitations, the model identified patterns that suggest the feasibility of developing such tools. The stronger performance in the first classification step (TD vs. non-TD) may be related to more balanced class distribution and aligns with prior findings from Mordaunt et al. [Mordaunt et al. 2019], who reported subtle expression differences primarily between TD and the other groups.

Three main challenges emerged from our results: limitations of the data, the choice of classifiers, and the inherent complexity of the task. First, the dataset's high dimensionality and small size pose challenges common in gene expression studies. Although dimensionality reduction and task separation strategies were applied, additional data, particularly from postnatal samples, could help validate and improve the approach. Exploring feature selection could also reveal candidate biomarkers for ASD in newborns.

Second, although the ensemble strategy helped balance the strengths of different classifiers, results showed that the two classification tasks may benefit from different modeling choices. For the second task, ensemble methods based on decision trees and models using logistic functions performed best. Future studies could refine the ensemble composition or test alternatives such as AdaBoost, Logistic Model Trees, or deep neural networks.

Lastly, it is possible that traditional ML methods are not sufficient to capture the subtle and complex patterns underlying ASD. More advanced approaches, such as deep learning or graph-based models like gene co-expression networks, may offer better performance, though they also require larger and more structured datasets.

5. Conclusion

While the proposed model is not yet suitable as a clinical tool, it provides a useful proof of concept. An important strength of this study is the use of umbilical cord blood transcriptomic data, which provides biological insights available at birth – long before behavioral symptoms can be clinically assessed. This opens possibilities for earlier interventions and reinforces the value of exploring perinatal biomarkers in future ASD research.

Our findings highlight the potential and challenges of using umbilical cord blood transcriptomic data for early ASD detection. The limitations encountered, such as data sparsity, model choice, and task complexity, point to clear directions for future work, including collecting richer datasets, refining model architectures, and exploring advanced learning algorithms.

Despite the challenges, this work contributes to the growing body of research seeking early, biologically informed ASD diagnostics. It reinforces the importance of

continued exploration in this field, especially when addressing conditions as multifaceted and impactful as ASD.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001; by grants from the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) [Projects No. 21/2551-0002052-0 and No. 22/2551-0000390-7 (CIARS)]; and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [Projects No. 308075/2021-8 and No. 440242/2024-0].

References

Hertz-Picciotto, I., Schmidt, R. J., Walker, C. K., Bennett, D. H., Oliver, M., Shedd-Wise, K. M., LaSalle, J. M., Giulivi, C., Puschner, B., Thomas, J., Roa, D. L., Pessah, I. N., Van de Water, J., Tancredi, D. J., and Ozonoff, S. (2018). A prospective study of environmental exposures and early biomarkers in autism spectrum disorder: Design, protocols, and preliminary data from the marbles study. *Environmental Health Perspectives*, 126(11):117004.

Hodges, H., Fealko, C., and Soares, N. (2020). Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Translational Pediatrics*, 9(Suppl 1):S55–S65.

Liu, W., Li, M., and Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8):888–898.

Lord, C., Brugha, T. S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E. J. H., Jones, R. M., Pickles, A., State, M. W., Taylor, J. L., and Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *Nature Reviews Disease Primers*, 6(1):1–23.

Mordaunt, C. E., Park, B. Y., Bakulski, K. M., Feinberg, J. I., Croen, L. A., Ladd-Acosta, C., Newschaffer, C. J., Volk, H. E., Ozonoff, S., Hertz-Picciotto, I., LaSalle, J. M., Schmidt, R. J., and Fallin, M. D. (2019). A meta-analysis of two high-risk prospective cohort studies reveals autism-specific transcriptional changes to chromatin, autoimmune, and environmental response genes in umbilical cord blood. *Molecular Autism*, 10(1):36.

Newschaffer, C. J., Croen, L. A., Fallin, M. D., Hertz-Picciotto, I., Nguyen, D. V., Lee, N. L., Berry, C. A., Farzadegan, H., Hess, H. N., Landa, R. J., Levy, S. E., Massolo, M. L., Meyerer, S. C., Mohammed, S. M., Oliver, M. C., Ozonoff, S., Pandey, J., Schroeder, A., and Shedd-Wise, K. M. (2012). Infant siblings and the investigation of autism risk factors. *Journal of Neurodevelopmental Disorders*, 4(1):7.

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

Sharma, S. R., Gonda, X., and Tarazi, F. I. (2018). Autism spectrum disorder: Classification, diagnosis and therapy. *Pharmacology Therapeutics*, 190:91–104.

Tylee, D. S., Hess, J. L., Quinn, T. P., Barve, R., Huang, H., Zhang-James, Y., Chang, J., Stamova, B. S., Sharp, F. R., Hertz-Picciotto, I., Faraone, S. V., Kong, S. W., and Glatt, S. J. (2017). Blood transcriptomic comparison of individuals with and without autism spectrum disorder: A combined-samples mega-analysis. *American journal of medical genetics. Part B, Neuropsychiatric genetics: the official publication of the International Society of Psychiatric Genetics*, 174(3):181–201.

Zablotsky, B., Black, L. I., and Blumberg, S. J. (2017). Estimated prevalence of children with diagnosed developmental disabilities in the united states, 2014-2016. *NCHS data brief*, (291):1–8.

Zhang, F., Savadjiev, P., Cai, W., Song, Y., Rathi, Y., Tunç, B., Parker, D., Kapur, T., Schultz, R. T., Makris, N., Verma, R., and O'Donnell, L. J. (2018). Whole brain white matter connectivity analysis using machine learning: an application to autism. *NeuroImage*, 172:826–837.