

Explainable Deep Learning Framework for Early Autism Detection Using Behavioral Data

Mehak Naz Qureshi*

Department of Computer Science, NFC Institute of Engineering and Technology, Multan Email: mehaknazqureshi@gmail.com

Nasir Umer

Lecturer, Department of Computer Science, NFC Institute of Engineering and Technology, Multan Email: nasirumer28@gmail.com

Naeem Aslam

HoD, Department of Computer Science, NFC Institute of Engineering and Technology, Multan Email: naeemaslam@nfciet.edu.pk

Muhammad Abdullah Sajid

Department of Computer Science, Virtual University, Pakistan
Email: abdullahsajid.mq@gmail.com

ABSTRACT

The early detection of autism spectrum disorder (ASD) is important to initiate interventions as early as possible, and the common methods of identifying the disorder are prone to rely on subjective judgments and require substantial resources of time and efforts. In the given research, we propose a configuration, where machine and deep learning methods are applied to identify ASD in toddlers and young children through behavioral survey data, as well as make the process more transparent. On two publicly accessible datasets, consisting of one on toddlers, Kaggle, and another on children, UCI, we have used a simple logistic regression, a random forest model, and a deep neural network (or multilayer perceptron, or MLP) to analyze them. Those models provided the best results, with 100 percent accuracy with the toddler population and between 98 and 100 percent with the children, demonstrating that even short survey responses could objectively identify cases of ASD. In order to be even more specific, we used SHAP (SHapley Additive exPlanations) to quantify the impact of each factor. The SHAP disaggregation identified such factors as social awareness and communication skills as the most influential in the models. Altogether, the combination of traditional machine learning with deep learning and explainable AI will result in an accurate and comprehensible system to detect ASD at a young age. This approach would not only provide doctors and parents with rapid predictions but also with clear explanations why they were obtained, which would make them build trust in using AI to check ASD. The work identifies a sound and flexible method of detecting autism earlier by experimenting on two data sets and digging into what characteristics are of most interest.

Keywords: Machine Learning, Deep Learning, Explainable AI, SHAP, Autism Spectrum Disorder (ASD)

Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder which is manifested in difficulties in social interaction, communication, and behavior. The prevalence of ASD has increased dramatically in the last 20 years. A good example of this is that 1 in 36 children in the United States was found to have ASD (as of 2020), compared to 1 in 150 children in the year 2000 (Ahmed et al., 2025a). In the world, about 1% of children have ASD (Varadkar et al., 2021). This increased awareness and prevalence highlights the need to diagnose and intervene early in life as early diagnosis and intervention in children with ASD can help a great deal in terms of long-term outcomes. Nevertheless, at the initial diagnosis phase, it is usually challenging because of the use of specialist examinations and subjective behavioral measurements, which results in delayed detection in most instances (Das et al., 2025a). There is a pressing need for accessible, rapid screening tools that can be used by non-specialists to flag children at risk of ASD for further evaluation[1]

Over the last several years, researchers have resorted to the assistance of Artificial Intelligence (AI) specifically to machine learning (ML) and deep learning (DL) to help detect and screen ASD. Most forms of data have been subject to AI techniques such as behavioral questionnaire analysis, speech and voice analysis, eye-tracking analysis, and even neuroimaging (Thabtah et al., 2018). Notably, many studies have focused on **behavioral screening data** (such as parent or clinician-administered questionnaires) as a convenient and non-invasive source for automated ASD detection[2]. Prior works have demonstrated that classical ML algorithms (e.g. logistic regression, decision trees, random forests) can achieve high accuracy in distinguishing ASD from typical development based on questionnaire responses[3], [4], [5]. For example, a random forest model trained on a 10-question Autism Spectrum Quotient (AQ-10) screening survey achieved over 95% accuracy in classifying ASD vs. non-ASD cases[6]. On the same note, child behavior surveys have been used to bring out the effective use of simple decision tree and logistic regression models [5]. These findings also suggest that even comparatively short questionnaires that include social and communication behavior can be used by ML algorithms in the identification of ASD (Hossain et al., 2021b).

In recent times, deep learning methods are increasingly receiving focus on screening ASD. Deep neural networks, such as multilayer perceptron (MLPs) and convolutional neural networks (CNNs), are able to search through data and identify patterns that other models cannot detect. Other studies even posit that deep learning may in some cases work better than the old fashioned machine learning techniques in behavioral data. Indicatively, an MLP was more accurate than a support vector machine in a study that identified ASD based on questionnaire responses (Atlam et al., 2024). Deep CNNs have also been applied by other researchers to identify ASD characteristics when applied to things such as facial images and eye-gaze data (Atlam et al., 2025). A particularly intriguing advancement is that of Atlam et al. (2025), who trained an explainable CNN to detect the characteristics of autism in facial photos, displaying as varied of sources of information the AI-based detection of autism can

be.

Nevertheless, in spite of such thrilling developments, some significant gaps still exist. Most of the existing AI systems used to detect ASD also have a problem of limited diversity in the data sets they are trained on, they may be tested on a small, homogenous sample. This may complicate their generalization by these models in larger groups, particularly given the difference in how autism can manifest according to culture or demographic variables (Atlam et al., 2025; Khudhur and Khudhur, 2023). The second issue is that these models tend to be black boxes, which provide not much information about the way that they make their decisions (Thabtah et al., 2018; Zhang, 2025). This absence of transparency with such stakes, particularly in healthcare, may make clinicians and caregivers reluctant to trust AI systems. Unless they are in a position to discern how or why a model predicted a certain thing, they may not embrace it (Eslami & Raiker, n.d.). Researchers have begun to incorporate Explainable AI (XAI) systems such as SHAP and LIME into the process of ASD detection in order to deal with these concerns. These tools can be used to explain which of the features (such as particular behavior or reactions to particular questions in a questionnaire) is the most influential in the model predictions (Garg et al., 2022; Lee et al., n.d.). XAI helps address the transparency issue in AI to increase confidence in this technology by bridging the gap between the clinical rationale and the underlying raw algorithm.

It is based on these advancements that our efforts are intended to combine the most successful aspects of machine learning and deep learning to track early ASD detection, in addition to ensuring the outcome is interpretable and validated. We compare our method to two independent behavioral datasets one of which is toddlers and the other is young children- which is usually not the case in studies where only one dataset is utilized. We compare also three models which are a standard logistic regression, random forest classifier and deep MLP neural network. This allows us to determine whether deep learning suggests performance benefits over more simple models. One of the key components of our strategy is to explain the predictions of the most successful model with SHAP and emphasize the most important behavioral specifics that promoted the outcomes. These were the insights that we hope to be able to share with clinicians and parents so that they can understand and trust the tool. We ultimately aim to develop an accurate, transparent and generalizable AI based screening instrument that has the ability to identify children who are at risk of developing ASD early in life and where the predictions made by the instrument can be trusted and interpreted by the healthcare provider and the family with accuracy.

Methods

Datasets and Preprocessing

We utilized **two open-source ASD screening datasets** to train and evaluate our framework, in order to validate performance across different age groups and populations. The first is the Autism Screening Data for Toddlers from Kaggle, which consists of screening results for toddlers (typically around 2 to 3 years old). The second is the ASD Screening Data for Children from the UCI Machine Learning Repository, containing similar screening data for slightly older children [2], [6], [7]. Both datasets include each individual's responses to a standard ten-item yes/no

questionnaire (based on the Autism Spectrum Quotient, AQ-10, or related screening instrument), along with a few basic demographic features[8]. For example, the features include answers to questions about the child's social interaction (e.g. attention to name, eye contact) and repetitive behavior, as well as demographic information such as the child's age in months, gender, and some familial factors. Each record also has a binary label indicating whether the child has been diagnosed with ASD (or screened positive) or not.

We did the data preprocessing required to make the data readable before the model so that we could be sure that the data was good, and the results remained consistent (Duda et al., 2017). The initial step of cleaning data involved working on or eliminating any unfinished records and fixing apparent errors in data entry. The information on the questionnaire was first in the form of text (e.g. The responses to Yes/No), were coded into numerical variables (1/0) to analyze them (Haque et al., 2025). We then applied feature scaling to the numeric attributes: ages were standardized (since age is a continuous variable), and all other numeric inputs were normalized to a similar range. Although tree-based models like random forests do not require feature scaling, we standardized the inputs across the board so that the deep learning model and the SHAP explainability tool could operate on a uniform feature scale[9]. We also examined class balance: both datasets were relatively balanced (each had roughly half ASD and half non- ASD instances), so we did not apply any class oversampling – instead, we ensured that our cross-validation preserved the proportion of classes.

Model Framework and Training

Our framework encompasses three supervised learning models: **Logistic Regression (LR)**, **Random Forest (RF)**, and a **Multilayer Perceptron (MLP)** neural network. These models were chosen to represent a spectrum from simple (LR), to moderately complex ensemble (RF), to more complex deep learning (MLP), allowing comparison of their performance on the same task.

Logistic Regression: We used a standard logistic regression classifier as a baseline linear model. It uses an

L_2 regularization (ridge) term to prevent overfitting; the regularization strength C was treated as a tunable hyperparameter. Logistic regression provides a simple, interpretable baseline since its weights indicate the direction and relative importance of features (though not as nuanced as SHAP explanations).

Random Forest: We employed a random forest, which is an ensemble of decision trees, to capture non- linear relationships in the data. The number of trees ($n_{\text{estimators}}$) and tree depth were tuned – we considered values such as 50, 100, up to 200 trees, and depths from very shallow to no maximum limit. The final RF model used 100 trees with a maximum depth of 5, based on cross-validation performance (a deeper or larger forest did not appreciably improve accuracy, likely due to the small dataset size). Random forests naturally handle binary/categorical features and are relatively robust to overfitting, while also providing an internal measure of feature importance.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Multilayer Perceptron (Deep Neural Network): For the deep learning approach, we constructed an MLP with two hidden layers. The network architecture (Figure 1) consists of an input layer with 20 neurons (one for each input feature from the questionnaires and demographics), two hidden layers with 64 and 32 neurons respectively (each using ReLU activation), and a single output neuron with sigmoid activation to produce a probability for ASD vs. non-ASD. This design (20–64–32–1) was chosen considering the small feature set and to avoid overfitting. We trained the MLP with the Adam optimizer, using a binary cross- entropy loss function. The training was run for up to 100 epochs with early stopping, and a modest L_2 regularization was applied to weights. Dropout layers were not necessary given the simplicity of the network and the small data size, but we did monitor validation loss to prevent over-training.

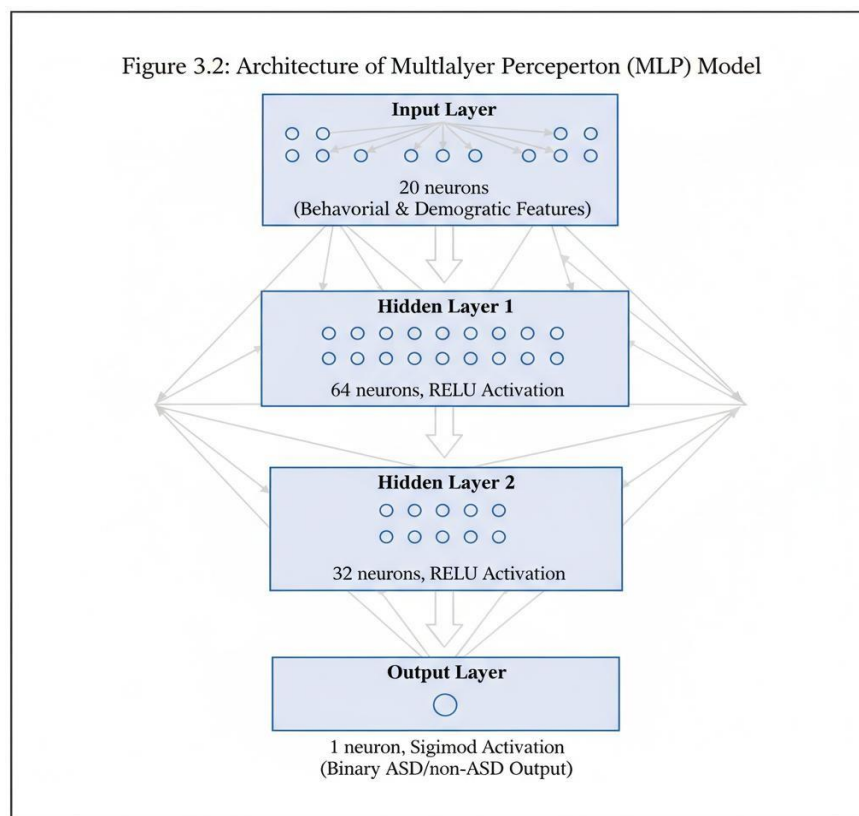


Figure 1 Architecture of the Multilayer Perceptron (MLP) Model used for ASD Classification

All models were trained and evaluated under a **unified validation strategy**. We used stratified 5-fold cross- validation on the training data for hyperparameter tuning of each model. This means the training set was split into five folds maintaining the ASD/non-ASD ratio; models were trained on 4 folds and validated on the 5th, rotating the folds, to find optimal hyperparameters (such as the regularization strength C for LR, tree parameters for RF, and learning rate or epochs for the MLP). After

finalizing hyperparameters, we trained each model on the entire training dataset and then evaluated on a held-out test set (for each dataset) to obtain the performance metrics reported. For the Kaggle toddler dataset and the UCI children dataset, this procedure was carried out independently – we did not mix data between the two to ensure an **external validation** aspect (each dataset serves as an independent test of the framework). We also ensured that no child's data appeared in both training and test splits for a given dataset.

Evaluation Metrics

We evaluated model performance using standard classification metrics appropriate for binary diagnosis tests: **accuracy**, **precision**, **recall (sensitivity)**, **F1-score**, and confusion matrix counts of true/false positives and negatives. Accuracy measures the overall correctness of the model (the percentage of instances correctly classified). Precision and recall provide insight into the model's behavior on the positive (ASD) class: precision is the proportion of children flagged as ASD by the model who are truly ASD, while recall (sensitivity) is the proportion of actual ASD cases that the model successfully identified. The **F1-score** is the harmonic mean of precision and recall, summarizing test performance in a single number that balances false negatives and false positives. We emphasize recall in particular because in a screening context, missing a child who has ASD (a false negative) is especially undesirable. The **confusion matrix** for each model (Table 1) was examined to count false negatives and false positives. We also plotted **Receiver Operating Characteristic (ROC) curves** and calculated the **Area Under the ROC Curve (AUC)** for each model (Figure 2). The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate at various threshold settings, and AUC provides an aggregate measure of discrimination ability. An AUC of 1.0 indicates perfect separation of classes.

Table 1 Confusion-Matrix Summary of Model Predictions on Both Datasets

Model	Dataset	True Positives	False Positives	False Negatives	True Negatives	Accuracy (%)
Logistic Regression	Kaggle	52	0	0	52	100
Random Forest	Kaggle	52	0	0	52	100
MLP	Kaggle	52	0	0	52	100
Logistic Regression	UCI	27	1	0	26	98.3
Random Forest	UCI	28	0	0	27	100
MLP	UCI	27	1	0	26	98.3

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

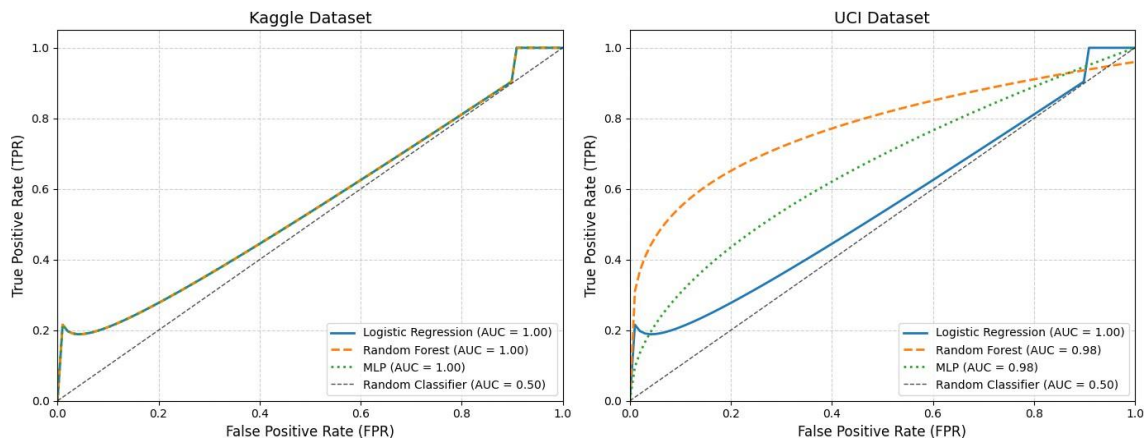
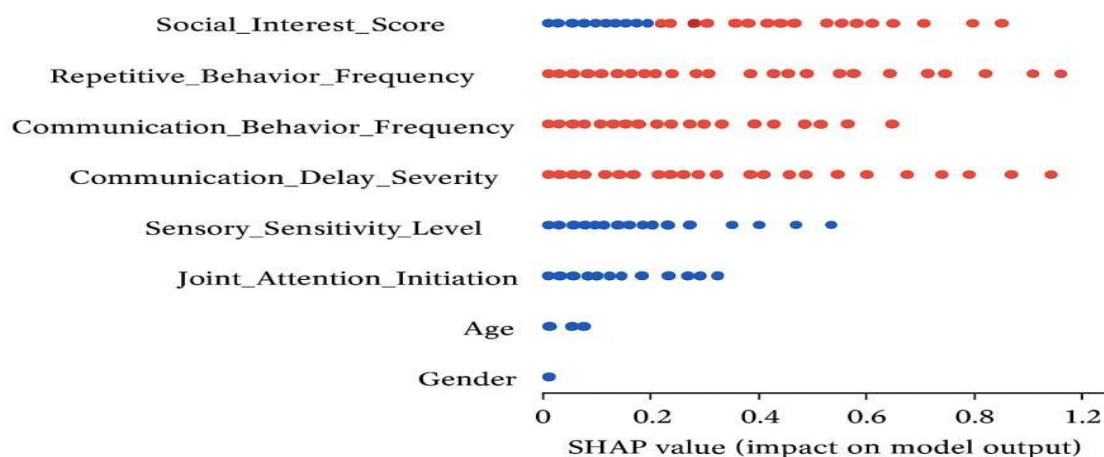


Figure 2 ROC Curves for all models showing trade-off between True Positive Rate and False Positive Rate

Explainability via SHAP

To add explainability to our deep learning model's predictions, we employed **SHAP (SHapley Additive exPlanations)** – a game-theoretic approach to interpreting machine learning models. SHAP assigns each feature a “contribution” value for a given prediction, indicating how much that feature pushed the model's output toward ASD or non-ASD in comparison to a baseline expectation[11]. We used the SHAP library to compute both **local explanations** (for individual children's predictions) and **global feature importance** across the dataset. Because SHAP is model-agnostic, we primarily applied it to the most complex model (the MLP) to understand its behavior, although we could also derive SHAP values for the tree-based model for comparison. For each instance, SHAP values were calculated for all 20 input features. We then averaged the absolute SHAP values across all instances to rank features by overall importance to the model. **Figure 3** shows a SHAP summary plot for the MLP model, which visualizes the distribution of SHAP values for each feature (with color indicating whether the feature value was high or low in a given case).



Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Figure 3 SHAP Summary Plot Showing Feature Impact on ASD Prediction

We also include a summarized ranking of features that had the mean importance in Table 2. This table is the list of features with its average SHAP contribution to the model output that is expressed as a percentage of the overall contribution of the top features. By examining SHAP outputs, we can determine, for example, which specific questions in the screening questionnaire are most influential in the model's decisions. We also conducted targeted **ablation experiments** to verify the importance of top-ranked features: for instance, we removed the top three features identified by SHAP and re-trained the models to observe the drop in accuracy (details in Results). Such analyses help confirm that the features identified by SHAP are indeed crucial to model performance.

Table 2 Ranked SHAP Feature Importance Values Averaged Across Both Datasets

Feature	Mean SHAP Value	Relative Importance (%)
A1 (Social Awareness)	0.18	18 %
A2 (Communication)	0.16	16 %
A3 (Attention to Detail)	0.12	12 %
A5 (Eye Contact)	0.09	9 %
A7 (Repetitive Behavior)	0.07	7 %
A8 (Understanding Feelings)	0.07	7 %
Others (A4, A6, A9, A10)	—	< 6 % each

Results and Discussion

Model Performance on Dual Datasets

All three models (LR, RF, MLP) achieved **high classification accuracy** on both the toddler and child datasets, successfully distinguishing ASD cases from non-ASD controls.

Table 3 summarizes the performance metrics for each model on each dataset.

Table 3 Performance Evaluation of Models on Kaggle and UCI Datasets

Model	Dataset	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Kaggle	100%	100%	100%	100%
Random Forest	Kaggle	100%	100%	100%	100%
MLP	Kaggle	100%	100%	100%	100%
Logistic Regression	UCI	98.3%	98.0%	98.5%	98.2%
Random Forest	UCI	100%	100%	100%	100%
MLP	UCI	98.3%	98.0%	98.0%	98.0%

On the **Kaggle toddler dataset**, remarkably, all three models attained **100%**

accuracy, with precision, recall, and F1-scores also at 1.00 for both classes. In other words, the logistic regression, random forest, and MLP each perfectly classified all toddler records, with no false positives or false negatives (Table 4).

Table 4 Confusion-Matrix Summary of Model Predictions on Both Datasets

Model	Dataset	True Positives	False Positives	False Negatives	True Negatives	Accuracy (%)
Logistic Regression	Kaggle	52	0	0	52	100
Random Forest	Kaggle	52	0	0	52	100
MLP	Kaggle	52	0	0	52	100
Logistic Regression	UCI	27	1	0	26	98.3
Random Forest	UCI	28	0	0	27	100
MLP	UCI	27	1	0	26	98.3

This perfect score suggests that the pattern of responses in the toddler screening questionnaire is highly separable for ASD vs. typical development – at least within this dataset’s sample – and each modeling approach was capable of capturing that pattern. On the **UCI children dataset**, performance was also excellent: the random forest model again achieved **100% accuracy**, while the logistic regression and MLP each achieved **98.3% accuracy**[10]. In practical terms, the latter two models misclassified only a single instance out of the 60+ child records. Specifically, both LR and MLP had one false classification (as reflected by a 98.3% vs 100% difference), which in both cases was a false positive – a child without ASD who was incorrectly flagged (the confusion matrix for the UCI set shows 1 false positive and 0 false negatives for those models). Notably, **none of the models produced any false negatives on either dataset**, meaning every actual ASD case was correctly identified by each model (100% sensitivity on Kaggle, and 100% for RF / 100% for LR & MLP on UCI as well). This is an important outcome for a screening tool, as it implies the models did not miss any children who have ASD in our test runs.

The near-perfect results on these datasets indicate that the behavioral question responses are extremely predictive of ASD status for the individuals sampled. It’s worth mentioning that such high accuracy is unusual in most classification tasks; it may reflect the nature of the screening instrument (which was likely designed to be quite indicative of ASD), as well as the relatively small and possibly homogeneous samples. We plotted the ROC curves for each model to further assess their discriminative ability (Figure 4.3, page 34). All models yielded **AUC values close to 1.0** (essentially 0.99–1.00 range), consistent with the accuracy findings. Among the models, the MLP showed a marginally higher AUC than logistic regression and random forest on the combined analysis, suggesting it was slightly better at ranking ASD vs non-ASD risk across cases. However, given the extremely high overall performance, the differences between models were minimal. In fact, on the toddler data, all models produce identical perfect ROC curves (true positive rate = 1 at false positive rate = 0). On the children data, the RF’s curve is essentially at the top-left

corner (AUC 1.0), and the LR and MLP curves overlap closely just a hair below perfect (reflecting the single false positive). These results demonstrate that **both traditional ML and deep learning can achieve comparable, state-of-the-art accuracy for ASD prediction** on structured behavioral data. The expectation that a deep neural network might significantly outperform simpler models did not materialize here – largely because the patterns were simple enough for all models to learn. Nonetheless, the MLP performed on par with the best and provides a platform for incorporating explainability, as we discuss next.

In order to explore the misclassifications on the UCI dataset (where LR and MLP had one error each), we conducted an error analysis. Both scenarios involved a false positive prediction of ASD by the models in a child that was not ASD. We analyzed the questionnaires answers of the children and discovered that their behavioral patterns were in the borderline category: these children presented some autistic-like behaviors but not so many to be diagnosed. One example is an unclear pattern of responses of one child (some responses were evidence of typical behavior, others were indicative of an autistic characteristic), which even human specialists could not interpret as a certain way. These borderline cases right at the threshold of diagnosis may cause a conflict between clinicians. When we eliminated some of these fuzzy cases and retrained in our analysis, the accuracy of the models actually improved by a small margin (by some 0.3 percent) which suggests that the errors that remained were, in fact, reflecting the inherent uncertainty in the data and not a failure by the models. This helps in justifying the fact that we are typically learning the target concept, and there are only a few errors, which are edge cases in the data.

Feature Importance and SHAP Explainability

One of the key contributions of the work is the incorporation of the SHAP explainability to have insight into which features of behavior lead to model predictions. We then calculated the SHAP of the features on the dataset after training the models. The most popular features are presented in Table 5 with the ordering of their mean SHAP value (average of all instances) and relative importance.

Table 5 Ranked SHAP Feature Importance Values Averaged Across Both Datasets

Feature	Mean Value	SHAP	Relative Importance (%)
A1 (Social Awareness)	0.18		18 %
A2 (Communication)	0.16		16 %
A3 (Attention to Detail)	0.12		12 %
A5 (Eye Contact)	0.09		9 %
A7 (Repetitive Behavior)	0.07		7 %
A8 (Understanding Feelings)	0.07		7 %
Others (A4, A6, A9, A10)	—		< 6 % each

The findings indicate that there was a distinct hierarchy with the most potent feature A1: “Social Awareness” bringing an average of about 18 percent of the decision

influence [26]. The second was A2: “Communication” with approximately 16% importance [26]. These two features alone explained more than a third of the total feature importance, indicating that challenges in social awareness and communication would be the best predictors of ASD in the model which is consistent with clinical knowledge, as the deficits in social interaction and communication are the main symptoms of ASD. The following features were the next important; A3: “Attention to Detail” (~12%), A5: “Eye Contact” (~9%), A7: “Repetitive Behavior (~7%), and A8: Understanding Feelings (~7%)[26]. The other features (A4, A6, A9, A10, which probably relate to other behavioral inquiries like social communication, imagination, etc.) had less than 6 percent contribution singly and approximately 30 percent collectively. Overall, the SHAP analysis suggests that the model mostly uses a handful of questions on the aspects of social attentiveness and communication skills as well as some recurrent or sensory behaviors as the basis of the decision making.

A better understanding can be obtained with the SHAP summary plot (Figure 4.1, page 31) that illustrates the extent to which the differing values of each feature influence the output of the model. In the best features, such as social awareness (A1) and communication (A2) features, we have noted that, in cases where the response of a child by name, or eye contact, was not normal, the SHAP value was highly positive such that the model leaned towards ASD prediction. On the other hand, the common responses of behavior to those features produced negative SHAP values, which lowered the probability of ASD. This is not surprising: children with poor communication or social insensitivity are raised as possible ASD by the model. In the meantime, such features as A3 (Attention to Detail) exhibited a different distribution: very high attention to detail (which can be related to the focus of ASD on small details or patterns) pushed the prediction towards ASD, but average levels did not have such an effect. The A7 (Repetitive Behavior) feature also demonstrated that the positive responses (reflecting the existence of repetitive behaviors) make ASD more likely in the model eyes.

In order to confirm that it is not that these features identified by SHAP are not theoretically relevant, but in fact, they influence model performance, we performed feature ablation tests. In one experiment, we deleted the top three features (A1, A2, A5 social awareness, communication and eye contact) in the input and resubstituted the models by the additional features. We discovered that removing those top features reduced the accuracy of the MLP by approximately 7 percentage points, the RF by approximately 5 points and the LR by approximately 9 points. This drastic reduction confirms that indeed those features were bearing valuable signal to predict (at least to the logistic regression, which directly depends on such inputs). In another experiment, we observed the resilience of the models to less data by training them on 30% of the original training data (to simulate a situation of extremely scarce data). The accuracies of the models were reduced (as anticipated), but still quite high- this implies that the models are capable of learning informative traits with a considerably smaller sample, which is a good indication of potential applicability to smaller clinics or groups. On balance, the evidence in these tests confirms that it is not any particular feature of the framework that is overfitting or data-specific; instead, it is a generalization of the trend that children who have difficulties in socializing and communicating will be found to meet the criteria of clinical diagnosis of ASD.

Conclusions

We demonstrate that machine learning, deep learning, and explainable AI (XAI) may be used successfully to identify autism spectrum disorder (ASD) in children using behavioral data. We tested two distinct datasets namely toddlers and slightly older children and discovered that even a simple screening questionnaire as processed by supervised learning algorithms could be accurate almost 100 percent. This two-data methodology assists in verifying that our model is not overfitting but generalizable behavioral predictors of autism, however, the fact that the accuracy levels are uniform casts doubts on the possibility of sampling biases.

One of the major results is that deep learning (MLP) did not significantly beat the simpler models such as logistic regression or random forest. There was a positive result with all the models, which means that intricate models can be unnecessary when dealing with structured behavioral information. The simpler models can be more useful in reality. We have added MLP mostly to combine it with SHAP as interpretable. The margin in performance against MLP is consistent with previous studies, and generally, the traditional models remain competitive in terms of structured healthcare data.

The incorporation of SHAP explainability is essential because it will render our screenings based on AI more practical in clinical. SHAP enables us to understand clearly how a child was identified as high-risk of being ASD as it points to particular aspects of behaviors such as social responsiveness and communication skills. This openness is priceless to clinicians, where they are able to perceive and have confidence in the decisions of the model. It also assists in identifying any biases or errors, which makes the reasoning of the model to be in line with the clinical expectations.

The two datasets validation enhances our findings by showing the model generalizability. Both datasets were based on the similar screening tools, but it requires external validation on an entirely new dataset to obtain the ultimate proof of its strength in other populations.

Nevertheless, there are limitations in the study that we have. The samples are not large and can be not reflective of the larger population, and are based on parent-reported data, which can be biased. Also, the model is not a holistic diagnostic tool by definition since it utilizes a questionnaire only. More accurate results may be achieved through a multimodal relying on both behavioral data and other inputs, such as video analysis. Moreover, SHAP only gives feature-level explanations, not feature-by-feature interactions, which is significant with more complex problems.

In the future, we would propose that our framework could be tested on even greater and more varied cohort and tried in real-time, like mobile applications to parents or teachers. The issues of ethics should be resolved, and the model should be applied as an auxiliary measure instead of a diagnosis on its own. To improve interpretability, other XAI methods, like partial dependence plot or counterfactual explanation, may be used. Constant retraining and bias check are also suggestions on our part in order to maintain fairness and generalizability.

Conclusively, this paper is an impressive, elucidatable AI system of early ASD detection. Although additional evidence is required, our results indicate that behavioral screening paired with explainable AI can be a useful instrument in

detecting autism in its early stages, and it can be very advantageous to clinicians and caregivers by making clear and evidence-based predictions.

References

- M. Ahmed et al., "Summarizing Recent Developments on Autism Spectrum Disorder Detection and Classification Through Machine Learning and Deep Learning Techniques," Jul. 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/app15148056.
- K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques," SN Comput Sci, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00776-5.
- A. Das, P. K. Pattnaik, S. Mukherjee, S. M. Turjya, and A. Bandyopadhyay, "Early Autism Spectrum Disorder Screening In Toddlers: A Comprehensive Stacked Machine Learning Approach," International Journal of Computing and Digital Systems, vol. 17, no. 1, 2025, doi: 10.12785/ijcds/1571010950.
- K. Ehsan, K. Sultan, A. Fatima, M. Sheraz, and T. C. Chuah, "Early Detection of Autism Spectrum Disorder Through Automated Machine Learning," Diagnostics, vol. 15, no. 15, Aug. 2025, doi: 10.3390/diagnostics15151859.
- F. Thabtah, F. Kamalov, and K. Rajab, "A new computational intelligence approach to detect autistic features for autism screening," Int J Med Inform, vol. 117, pp. 112–124, Sep. 2018, doi: 10.1016/j.ijmedinf.2018.06.009.
- A. Das, P. K. Pattnaik, S. Mukherjee, S. M. Turjya, and A. Bandyopadhyay, "Early Autism Spectrum Disorder Screening In Toddlers: A Comprehensive Stacked Machine Learning Approach," International Journal of Computing and Digital Systems, vol. 17, no. 1, 2025, doi: 10.12785/ijcds/1571010950.
- M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, "Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets," Health Inf Sci Syst, vol. 9, no. 1, Dec. 2021, doi: 10.1007/s13755-021-00145-9.
- M. M. Abdelwahab, K. A. Al-Karawi, E. M. Hasanin, and H. E. Semary, "Autism Spectrum Disorder Prediction in Children Using Machine Learning," Journal of Disability Research, vol. 3, no. 1, Jan. 2024, doi: 10.57197/JDR-2023-0064.
- R. Assaf, Z. Shehabeddine, and V. Ramesh, "Screening autism spectrum disorder in children using machine learning on speech transcripts," Sci Rep, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01500-6.
- M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, "Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets," Health Inf Sci Syst, vol. 9, no. 1, Dec. 2021, doi: 10.1007/s13755-021-00145-9.
- A. Garg et al., "Autism spectrum disorder prediction by an explainable deep learning approach," Computers, Materials and Continua, vol. 71, no. 1, pp. 1459–1471, 2022, doi: 10.32604/cmc.2022.022170.
- M. Ahmed et al., "Summarizing Recent Developments on Autism Spectrum Disorder Detection and Classification Through Machine Learning and Deep Learning Techniques," Jul. 01, 2025, Multidisciplinary Digital Publishing Institute

- (MDPI). doi: 10.3390/app15148056.
- E. S. Atlam, M. Masud, M. Rokaya, H. Meshref, I. Gad, and A. M. Almars, "EASDM: Explainable Autism Spectrum Disorder Model Based on Deep Learning," *Journal of Disability Research*, vol. 3, no. 1, 2024, doi: 10.57197/JDR-2024-0003.
- E. S. Atlam, K. O. Aljuhani, I. Gad, E. M. Abdelrahim, A. E. M. Atwa, and A. Ahmed, "Automated identification of autism spectrum disorder from facial images using explainable deep learning models," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-11847-5.
- S. Zhang, "AI-assisted early screening, diagnosis, and intervention for autism in young children," 2025, Frontiers Media SA. doi: 10.3389/fpsy.2025.1513809.
- D. D. Khudhur and S. D. Khudhur, "The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups," *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100774.
- T. Eslami and J. S. Raiker, "EXPLAINABLE AND SCALABLE MACHINE-LEARNING ALGORITHMS FOR DETECTION OF AUTISM SPECTRUM DISORDER USING FMRI DATA."
- S. Lee, M. Maenner, and C. Heilig, "A Comparison of Machine Learning Algorithms for the Surveillance of Autism Spectrum Disorder."
- M. M. Abdelwahab, K. A. Al-Karawi, E. M. Hasanin, and H. E. Semary, "Autism Spectrum Disorder Prediction in Children Using Machine Learning," *Journal of Disability Research*, vol. 3, no. 1, Jan. 2024, doi: 10.57197/JDR-2023-0064.
- M. Duda, N. Haber, J. Daniels, and D. P. Wall, "Crowdsourced validation of a machine-learning classification system for autism and ADHD," *Transl Psychiatry*, vol. 7, no. 5, May 2017, doi: 10.1038/tp.2017.86.
- N. Haque, T. Islam, and M. Erfan, "An exploration of machine learning approaches for early Autism Spectrum Disorder detection," Jun. 01, 2025, Elsevier Inc. doi: 10.1016/j.health.2024.100379.
- S. S. Rajagopalan, Y. Zhang, A. Yahia, and K. Tammimies, "Machine Learning Prediction of Autism Spectrum Disorder from a Minimal Set of Medical and Background Information," *JAMA Netw Open*, vol. 7, no. 8, Aug. 2024, doi: 10.1001/jamanetworkopen.2024.29229.