# An Explainable Hybrid Machine Learning Framework for Autism Spectrum Disorder Screening Using Behavioral and Demographic Information

[1]R.prathusha

[1]Assistant Professor, Department of Computer Science and Engineering, Methodist college of engineering and technology,Abids,Hyderabad

[1]Email: prathyusha88r@gmail.com

## Abstract:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder in which early identification and proper assessment can make a great difference. The use of machine learning techniques for classifying and detecting ASD has become one of the hotspots in the medical field. However, it is still challenging to develop highly performing but also interpretable predictive models. This study proposes an explainable hybrid machine learning framework for ASD screening using behavioral and demographic information. The study utilizes a publicly available ASD screening dataset containing 704 records and 21 original attributes. After duplicate removal, missing-value handling, feature processing, and data quality assessment, 699 records were retained for modeling. Logistic Regression, Support Vector Machine, Random Forest, and XGBoost were evaluated, followed by hyperparameter optimization of XGBoost using Optuna. An ablation study was conducted to investigate the contribution of behavioral and demographic feature groups. The XGBoost model demonstrated exceptional performance, achieving an accuracy of 97.86%, precision of 94.74%, recall of 97.30%, and an impressive F1-score of 96%. The area under the ROC curve (ROC-AUC) came up to 99.84%, while the area under the precision-recall curve (PR-AUC) reached 99.59%. Besides, the model's Brier score was only 0.0167, which indicates outstanding calibration. The most informative behavioral features according to the SHAP analysis are A9, A6, A5, A7, and A4. Local SHAP explanations and counterfactual analysis were utilized to generate instance-specific interpretations of the model's predictions. The results indicate that behavioral screening responses offer significantly better predictive insights than demographic characteristics in this dataset. The suggested framework can thus function as a clear computational method for ASD screening, while clinical evaluation continues to be crucial for diagnosis.

**Key words:** Autism Spectrum Disorder, ASD Screening, Machine Learning, XGBoost, Explainable Artificial Intelligence, SHAP, Counterfactual Explanation, Behavioral Features, Demographic Features, Healthcare AI.

## 1. Introduction

Autism Spectrum Disorder, or ASD, is a condition that affects how a person develops. It often involves trouble with social communication. It can also include limited or repetitive patterns of actions. Catching it sooner can help people get better support. To screen and assess ASD, a person usually goes through more than one step [1]. This can include questions to fill out for screening, plus questionnaires. There may also be tests and other evaluations. Specialists can be involved to review the results.

More health and behavior data are now available. That has opened doors for machine learning in tasks tied to screening. These models can look at many signals at once. They can learn how features relate to what happens in the screening results. For autism research, survey answers are a key source. They reflect aspects of social and behavior function [2]. Other details can also matter. Demographic and situational factors can add more context about the person being evaluated. Still, good accuracy alone does not solve the main issue. In health settings, it is not enough to label a case correctly. People also need to know why the model gave that result. This is especially true if the output will be used to guide more checks.

Researchers have tried several machine learning approaches for classifying ASD screening results. Logistic Regression is often used because it is straightforward and easy to explain. Support Vector Machines and some ensemble approaches can model harder patterns in the data. Methods based on trees, like XGBoost, can work well when feature relationships are not linear [3]. Model selection matters a lot. Tuning hyperparameters can change the results in a noticeable way. For that reason, it helps to test more than one model [4]. After that, the chosen model should be optimized. This is a better way to judge how well the predictions will hold up.

A key issue is whether people can understand what the model is doing. Some XAI tools help show why a prediction happened [5]. For instance, SHAP can point to how much each input feature pushes the output. It can also show this across many cases and for a single case. There are also counterfactual explanations. These look for small changes to the input that would lead to a different prediction. When you use both kinds of explanations, the system can feel clearer. That is often better than just showing the final class.

Some earlier work suggests that machine learning can help with ASD screening. Still, there is a gap. Many teams do not put all key steps in one place [6]. Those steps include careful preprocessing, testing models side by side, tuning hyperparameters, checking feature groups, and also producing predictions that can be explained. It also matters to ask a specific question. If the behavioral measures are already strong, does adding demographics help in a real way? In other words, do demographic details add extra predictive power, or do they mostly add noise. For this reason, the study looks at behavioral data on its own. It also looks at demographic data on its own. Then it tests both together. After that, it uses explainability methods to show how the final model reaches its outputs.

This work has several goals. First, it aims to build a machine learning system that can explain its results for ASD screening [7]. Next, it will test more than one classification method and see how they compare. Then, it will tune the best model to get stronger results. It will also look at how much different sets of inputs matter, like behavioral features and demographic features. After that, it will show both broad model reasoning and person-level insights. It will do this with SHAP and counterfactual analysis. The plan is meant for use as a computer based screening tool. It should help guide follow up evaluation. It is not meant to take the place of a clinician’s judgment for diagnosis.

## 2. Related Work

People have looked more and more at machine learning for ASD checks and for predicting whether someone may have Autism Spectrum Disorder. With machine learning, systems can spot links in data made from behavior, basic demographics, and clinical records. This matters because the connections between a person’s traits and ASD results may not be clear at first. In past work, teams tried several common models. They used Logistic Regression, Support

Vector Machine, Decision Tree, Random Forest, and methods that boost results. Across these studies, the findings are fairly consistent. Machine learning can reach useful prediction levels for ASD screening. This is especially true when answers from behavioral questionnaires are used as features.

### 2.1 Machine Learning for ASD Screening

Some research has looked at supervised machine learning for ASD classification. Earlier work has used both standard statistics and common ML methods with questionnaire style data [8]. The goal is to tell apart people who show ASD screening signals from those who do not. Logistic Regression shows up a lot. People pick it for being simple and for giving results that are easier to read. Support Vector Machine is used too, mainly when the links between inputs and the class are not that neat. Random Forest and other tree methods are also common. They can work with curves in the data and with links between variables. This means they do not rely on many strong claims about how the data must be distributed.

People keep returning to ensemble learning since mixing several decision models often leads to more steady classification results. In Random Forest, many decision trees vote together. Boosting works in a different way [9]. It trains a new model step by step, trying to fix mistakes made earlier. XGBoost is a common boosting approach. It has shown good results on many types of classification tasks. In practice, it can learn patterns that are not linear [10]. It also can capture links between input variables. Regularization can reduce the risk of overfitting. Even so, you cannot be sure of the final numbers. What you see can shift if the dataset changes. It also depends on the features you pick. The way you prepare the inputs matters too. The model setup matters as well.

### 2.2 Behavioral and Demographic features in ASD prediction

Behavior traits are a key clue in ASD checks. Most screening forms are set up with items about how a person handles social situations. They also ask about how someone talks or listens. Other questions focus on noticing social signals [11]. Some items look at routines or repeated habits. Answers on these forms are often yes or no, or they fit into set groups. That kind of data can be used right away in common machine learning tasks. A number of papers have said that scores from these behavior lists can help predict outcomes. The reason is that the questions are tied to what the screen is trying to find.

ASD prediction work often includes demographic and setting details too. Researchers may use age, sex, race, family history, and other personal background items. These can help describe someone before any test results [12]. Still, what these factors predict is not the same across studies. One dataset may show a link between a demographic trait and the outcome. Another dataset may show almost no added value once behavior data is included. Because of that, it helps to group features during model building. Feature-group analysis matters when making an ASD screen.

When you mix behavior data with demographic data, you have to ask if it truly helps the prediction. Just adding more variables does not mean the model will get better. Some variables bring noise or repeat what is already known, and they can make the model harder to generalize. Because of that, it can help to test behavior features on their own and then test demographic features on their own. This way, you can see which part of the data is doing most of the work for the predictions.

We looked at this problem using an ablation test with three setups. First, we used only behavioral features. Next, we used only demographic features. Then we used behavioral features plus demographic data. This way, we can see how much each feature set helps, instead of just treating every variable as equally important.

### 2.3 Explainable AI for Healthcare Prediction

Some machine learning models do well at classification tasks. Still, it is not always easy to see what led the model to one specific result. This problem shows up more when the model uses nonlinear ensembles. In medicine, the issue is even more serious. Clinicians and staff often need to know what pushed the prediction in that direction [13]. They ask for that context before they act on the output during the next review step.

SHapley Additive exPlanations (SHAP) is one of the commonly used model explanation approaches. It assigns contribution values to individual features based on their influence on a particular prediction. SHAP can provide both global and local explanations [14]. Global explanations can be used to identify the features that have the greatest overall influence on a model, while local explanations can show why a particular individual was classified into a specific class. This distinction is useful in ASD screening because it allows researchers to examine both general behavioral patterns learned by the model and the reasoning behind individual predictions.

Counterfactual explanations can add to interpretability. Rather than just listing why a model chose its current output, they ask what would need to change in the input to get a different result. In a two-class ASD screening system, this means you can look for a close set of behavior answers that would flip the model from class A to class B. These outputs can help people make sense of what the model is doing.

Explainability is therefore particularly relevant when developing machine learning systems for ASD screening. A model that provides both accurate predictions and understandable explanations can offer more useful information for researchers and screening-support applications than a model that only produces a classification label.

### 2.4 Research Gap

A lot of work mainly reports how well a model can label cases. Many papers also test only a small set of methods and do not look closely at which feature groups drive the final result. On top of that, models like boosting can be hard to read. Without extra tools, it is not clear why the model makes a specific call. So, we need methods that judge prediction quality and show what parts of the input matter. We also need explanations at the level of each person.

Another issue is that people often mix behavior data with demographic data. They do not always check if the demographic part adds real predictive value. A feature group ablation test can help here. It checks behavioral data alone first. Then it checks demographic data alone. After that, it tests the two together. This can show what drives the model. It can reveal whether performance mostly comes from the behavioral screening answers. It can also show if demographic data adds extra useful signal.

This work fills the gaps noted earlier. We build one machine learning setup. It ties together behavior data and demographic cues. We then compare multiple classification results. We also run ablation checks to see what matters most. For clarity, we add an explainability step. For explanations, we use SHAP. It helps show both overall patterns and case by case drivers. We also use counterfactual testing. This gives a second view of why a model makes a specific call for a person. With this design, the study checks more than prediction accuracy. It also looks at which features push the ASD screening output. In addition, we test how one person's predicted result can shift when the input values change.

## 3. Materials and Methods

### 3.1 Dataset Description

The study uses the UCI Autism Screening Adult dataset, which contains behavioral responses and demographic information collected for ASD screening. The original dataset contains 704 records and 21 variables, including ten behavioral screening items (A1–A10), demographic attributes, and the ASD class label.

Once the dataset was cleaned, we removed duplicated rows. We also removed rows with ages that were illogical, such as obviously incorrect figures. Upon completion, we possessed 699 records to utilize. If a category lacked a value, we left the row unchanged. We marked that absent section as Unknown. In the result column, 0 was assigned for non-ASD. For ASD, we utilized 1. Ultimately, there were 512 entries classified as non-ASD. A total of 187 records were classified as ASD.

### 3.2 Data Preprocessing and Feature representation

The ten behavioral variables (A1–A10) were retained as binary features because they represent the responses to the ASD screening questions. Demographic and contextual variables included age, gender, ethnicity, history of jaundice, family history of autism, country of residence, and relation of the respondent.

The result variable was excluded from model training because it is directly derived from the A1–A10 responses and could introduce redundant information. Similarly, variables with little or no useful variation were excluded. Categorical variables were encoded using one-hot encoding, while numerical variables were retained in their appropriate form.

The behavioral score, calculated as the sum of A1–A10, was used for descriptive analysis to examine differences between the ASD and non-ASD groups, but it was not treated as an additional independent predictor.

### 3.3 Exploratory and Statistical Analysis

Exploratory analysis was performed to examine class distribution and the relationship between behavioral and demographic variables and the ASD outcome. The behavioral score distributions were compared between the two classes using the Mann–Whitney U test. Chi-square tests were used to assess associations between individual behavioral responses and ASD status.

For categorical demographic variables, Cramér's V was used to measure the strength of association with the target variable. These analyses provided an initial understanding of which variables contained useful predictive information before model development.

### 3.4 Machine Learning and Optimization

Four machine learning algorithms were evaluated: Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. The data were divided into training and testing sets using an 80:20 stratified split, preserving the class distribution in both subsets.

XGBoost was subsequently optimized using Optuna-based hyperparameter optimization with cross-validation and F1-score as the optimization objective. The optimized XGBoost model was selected as the final predictive model based on its held-out test performance.

An ablation analysis was also conducted using three feature configurations: behavioral features only, demographic features only, and behavioral plus demographic features. This analysis was used to determine the relative contribution of the two feature groups.

### 3.5 Explainability and Performance Evaluation

Model predictions were analyzed using SHAP (SHapley Additive exPlanations) to identify globally important behavioral features and to explain individual predictions. Counterfactual analysis was additionally used to determine the minimal feature changes that could alter a model prediction. These explanations were treated as model-level interpretations rather than clinical recommendations.

Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. A confusion matrix was used to examine classification errors, while calibration and bootstrap confidence intervals were used to provide additional assessment of model reliability.

## 4. Proposed Framework

The proposed framework is designed to provide an interpretable machine learning approach for ASD screening using behavioral responses and demographic information. Rather than relying on a single classification algorithm, the framework evaluates multiple machine learning models, identifies the most effective predictive configuration, and integrates explainability techniques to make the resulting predictions easier to understand.

### 4.1 Framework Overview

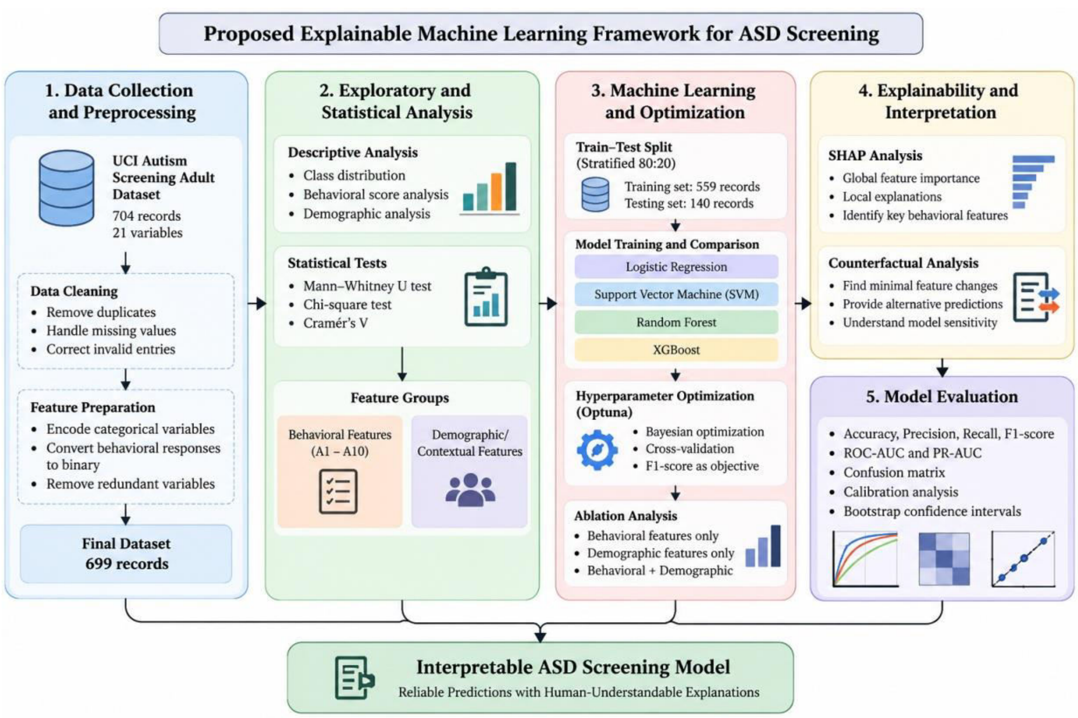


**Figure 1: Proposed Explainable Machine Learning Framework for ASD Screening**

The framework consists of four main stages: data preparation, feature analysis, predictive modeling, and explainability. Behavioral and demographic information is first cleaned and transformed into a suitable representation for machine learning. Exploratory and statistical analyses are then performed to understand the relationship between the input variables and ASD status.

Four classification models—Logistic Regression, SVM, Random Forest, and XGBoost—are trained and compared. XGBoost is further optimized using Optuna to obtain a suitable set of hyperparameters. An ablation analysis is performed to determine whether demographic information provides additional predictive value beyond the behavioral screening responses.

### 4.2 Feature Analysis and Model Development

The framework treats the ten behavioral screening responses as the primary predictive information. Demographic variables are analyzed separately and together with the behavioral variables to measure their contribution to the classification task.

The candidate models are trained using the same training and testing strategy to enable a consistent comparison. After model evaluation, the optimized XGBoost classifier is selected as the final predictive engine because it provides the strongest overall performance on the independent test set. Although a probability-based hybrid combination of multiple models was also investigated, it was not retained as the final model because its test performance did not improve over optimized XGBoost.

### 4.3 Explainability Module

An explainability layer is integrated after model development to examine how the final XGBoost model reaches its predictions. SHAP is used for both global and local interpretation. Global explanations identify the behavioral variables that contribute most strongly to the model, while local explanations show how individual feature values influence a particular prediction.

A counterfactual explanation component is also included. It searches for minimal changes in the input behavioral responses that can change the model's predicted class. This provides an additional perspective on model behavior and helps illustrate the sensitivity of individual predictions.

## 5. Experimental Results

### 5.1 Dataset and Behavioral Analysis

After preprocessing, the final dataset contained 699 records, including 512 non-ASD cases and 187 ASD cases. This corresponds to 73.25% non-ASD and 26.75% ASD samples.

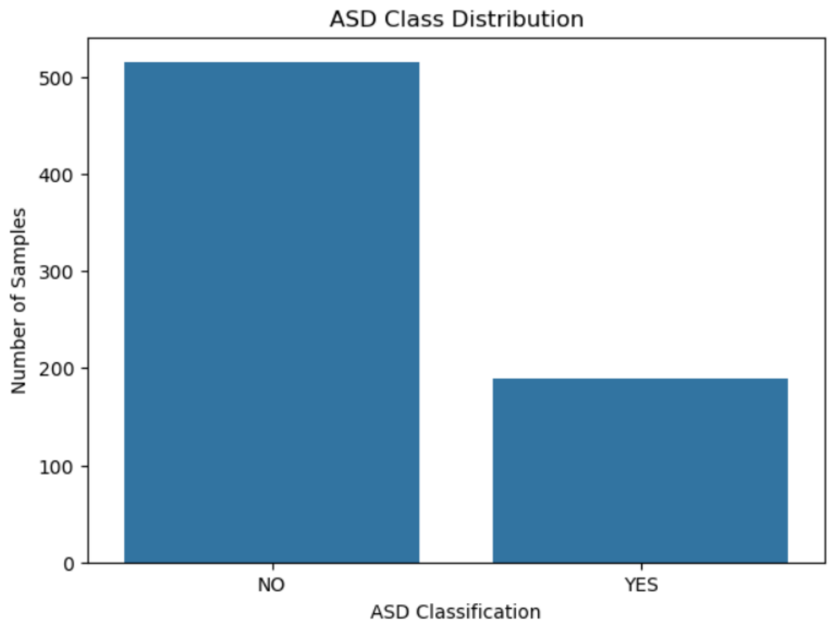


**Figure 2: Distribution of ASD and Non-ASD Classes in the Dataset**

The two groups did not show the same results. For the group without ASD, the mean score was 3.63. For the ASD group, the mean was 8.25, which was higher. We then used a Mann-Whitney U test. That test suggested a noticeable gap between the groups.

Among the ten behavioral screening items, A9, A6, A5, A4, and A3 showed the strongest associations with the ASD outcome based on the chi-square analysis. These results indicate that the individual behavioral responses contain considerably stronger predictive information than the demographic variables.

### 5.2 Comparision of Machine Learning Models

Four classification algorithms were evaluated on the independent test set. The results are presented in Table 1.

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|
| Logistic Regression | 95.71% | 100.00% | 83.78% | 91.18% | 99.82% | 99.49% |
| SVM | 92.14% | 90.63% | 78.38% | 84.06% | 98.98% | 97.06% |
| Random Forest | 92.14% | 93.33% | 75.68% | 83.58% | 98.62% | 96.50% |
| Optimized XGBoost | 97.86% | 94.74% | 97.30% | 96.00% | 99.84% | 99.59% |

**Table 1: Performance Comparison of the Evaluated Machine Learning Models**

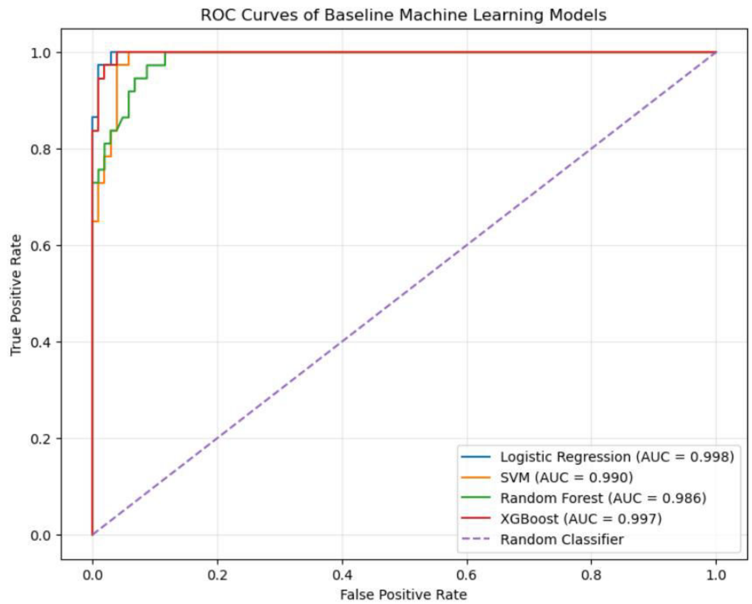


**Figure 3: Receiver Operating Characteristic (ROC) Curves of the Evaluated Models**

The tuned XGBoost setup did best on the test set. About 97.86% of the test cases were labeled right. The F1-score came out to 96.00%. Recall was 97.30%, which means most ASD cases were caught. The ROC-AUC reached 99.84%. The PR-AUC was 99.59%. Together, these scores suggest the classes are separated well.

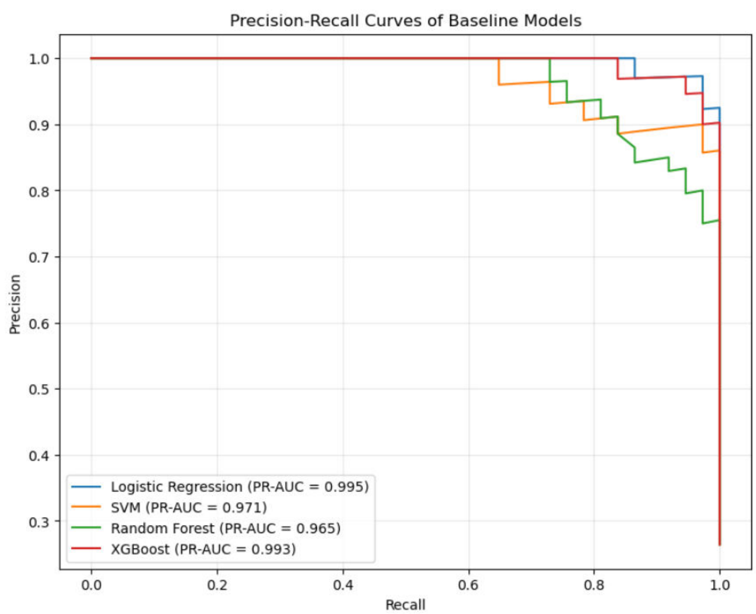


**Figure 4: Precision–Recall Curves of the Evaluated Models**

Logistic Regression also performed strongly, particularly in precision and ROC-AUC, showing that the behavioral screening variables contain highly discriminative information even for a relatively simple linear classifier. However, optimized XGBoost provided a better balance between precision and recall on the independent test set.

### 5.3 Optimized XGBoost Performance

The model correctly identified 101 non-ASD and 36 ASD cases, with only two false positives and one false negative. The resulting specificity was 98.06%, while ASD recall was 97.30%.

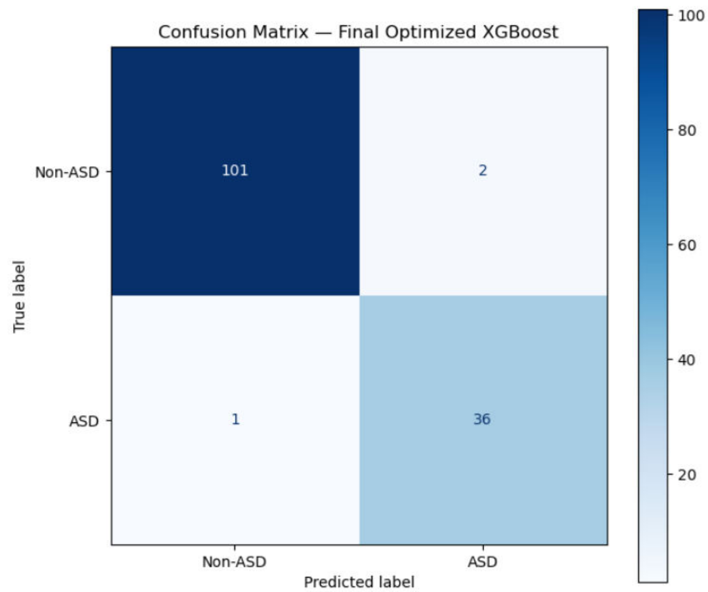


**Figure 5: Confusion Matrix of the Optimized XGBoost Model**

In the bootstrap pass, the discrimination stayed about the same. The ROC AUC 95% band moved from 0.9947 to 1.0000. The PR AUC 95% band was 0.9855 through 1.0000. The Brier score came out to 0.0167. Overall, the predicted probabilities matched the observed pattern pretty well.

### 5.4 Ablation Analysis

To examine the contribution of different information sources, three feature configurations were evaluated, it is clearly mentioned in Table 2.

| Feature Configuration | Accuracy | Precision | Recall | F1-Score | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|
| Behavioral Only | 97.86% | 94.74% | 97.30% | 96.00% | 99.92% | 99.80% |
| Demographic Only | 74.29% | 52.00% | 35.14% | 41.94% | 65.59% | 47.18% |
| Behavioral + Demographic | 97.86% | 94.74% | 97.30% | 96.00% | 99.84% | 99.59% |

**Table 2: Ablation Analysis of Behavioral and Demographic Feature Configurations**

The ablation run shows that behavior cues drive most of the prediction. When we used only demographic data, the results dropped a lot. Then we tried adding demographics on top of the behavior inputs, and nothing really got better. Accuracy stayed about the same, and so did precision, recall, and F1-score. So the last model leans mainly on the behavior screening answers. Demographics still help for simple checks and to make sense of what is in the dataset.

## 5.5 Model Calibration

Model calibration was evaluated by comparing the predicted probabilities with the observed proportions of ASD cases. The calibration curves provide an additional assessment of whether the probability estimates generated by the models are consistent with the actual outcomes. A model whose calibration curve remains close to the diagonal reference line provides more reliable probability estimates.

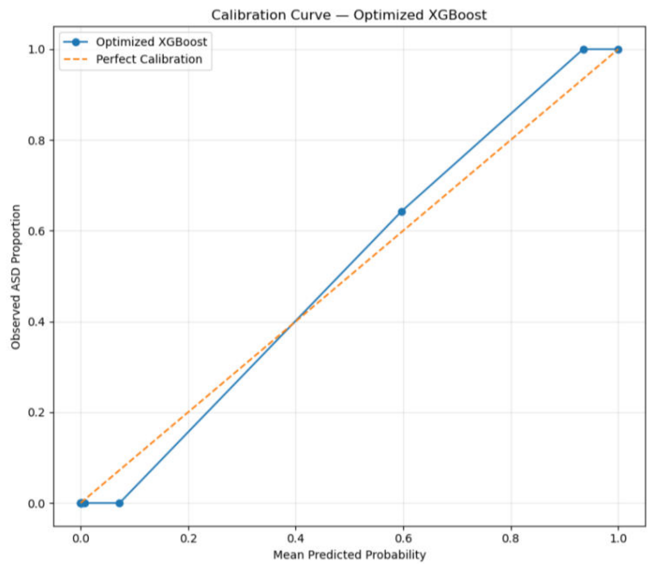


**Figure 6: Calibration Curves of the Evaluated Machine Learning Models**

The calibration results go along with the earlier discrimination scores. ROC-AUC and PR-AUC show how well the models tell ASD apart from non-ASD. Calibration, instead, checks whether the predicted probabilities match what actually happens. For the best XGBoost model, the Brier score was 0.0167. That value points to a small average error in its probability forecasts on the test set.

## 5.6 Explainability Results

SHAP analysis was applied to the final XGBoost model to examine feature importance. The most influential behavioral variables were A9, A6, A5, A7, and A4, followed by A3, A2, A8, A10, and A1.

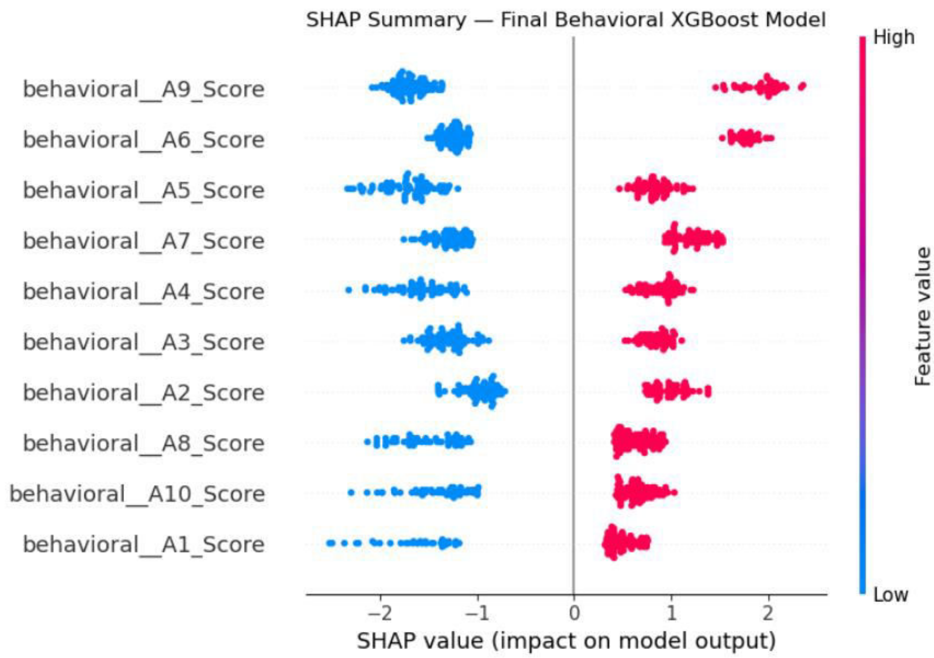


**Figure 7: Global SHAP Feature Importance of the Optimized XGBoost Model**

The SHAP results were generally consistent with the statistical analysis, particularly for A9, A6, A5, A4, and A3. Higher behavioral response values generally contributed toward the ASD prediction, while lower values tended to contribute toward the non-ASD prediction. However, these contributions represent the behavior of the trained model and should not be interpreted as causal or clinical effects.

### 5.7 Counterfactual Analysis

Counterfactual explanations were used to examine how small changes in behavioral responses could alter individual predictions. For one ASD instance, changing a single behavioral response was sufficient to move the model prediction toward the non-ASD class. For a selected non-ASD instance, three behavioral changes were required to obtain an ASD prediction.

These examples demonstrate that the model can provide not only a predicted class but also an indication of how sensitive that prediction is to changes in the input pattern. Counterfactual results are intended to explain model behavior and are not clinical recommendations.

## 6. Discussion

In the study, the experiments show that machine learning can spot signals linked to ASD screening answers in the dataset used. Four models were tested. The best results came from the tuned XGBoost setup on the separate test set. It reached 97.86% accuracy and a 96.00% F1-score. The ROC-AUC and PR-AUC were also high. That suggests the model separated ASD from non-ASD cases in a strong way.

The research indicated that behavioral screening tools are crucial for accurate prediction. When the model utilized solely the behavioral inputs, it achieved identical outcomes to the model that included demographic information as well. The measurements aligned in terms of accuracy, precision, recall, and the F1-score. However, the model that depended solely on demographics performed significantly worse. In general, the findings indicate that behavioral responses are the primary source of predictive signal within this dataset. The demographic characteristics appear to contribute only a minimal amount, if at all, beyond what the behavioral information already encompasses.

The statistical analysis and SHAP results showed similar patterns in feature importance. Variables such as A9, A6, A5, A4, and A3 consistently appeared among the most informative behavioral features. This agreement between statistical analysis and model-based explanation provides additional support for the observed feature patterns. However, these findings should be interpreted as associations learned from the dataset rather than evidence that individual behavioral responses independently cause ASD.

The performance of optimized XGBoost also indicates the benefit of systematic hyperparameter optimization. Although Logistic Regression performed strongly and achieved very high discrimination, XGBoost provided a better balance between precision and recall on the independent test set. The confusion matrix showed only two false-positive and one false-negative prediction, demonstrating the model's strong classification performance on the available test samples.

The explainability component adds an important dimension to the predictive framework. SHAP provides both global and individual-level interpretations, allowing the contribution of each behavioral variable to be examined. Counterfactual analysis provides a complementary perspective by showing how changes in input responses can influence the model's predicted class. Together, these methods make the model's decision process more transparent than a prediction-only approach.

The calibration analysis and low Brier score further support the usefulness of the model's probability estimates. This is relevant for a screening-support system because probability information can provide more context than a simple binary prediction. Nevertheless, probability estimates obtained from this relatively small dataset should not be interpreted as clinical risk probabilities without validation on larger and independent populations.

The results point to an explainable machine learning setup that uses mostly behavioral screening data. It can still do well at prediction and stays easier to interpret. In this work, the method is meant to help with screening, not to act as a full diagnostic tool. Before it could be used in real clinical settings, more tests are needed. These tests should use bigger data sources from different regions and data that has been clinically confirmed.

## 7. Limitations

The results come with a few limits. We used a small public dataset, so the findings may not hold up for other groups. We also tested the approach on a separate split, but we did not bring in outside clinical data to check it. When we looked at demographics, they did not add clear value compared with behavioral features. Any links we saw could be tied to quirks of this dataset. We should also be careful with SHAP and counterfactual outputs. They show how the model acts. They do not prove cause, and they are not meant as medical advice. For the next steps, we should use bigger and more varied datasets. We also need outside validation. And we should test the method on more than one type of population.

## 8. Conclusion

We built an explainable hybrid machine learning method for ASD screening using behavioral data and basic demographic details. We tested four different machine learning models. The best results came from a tuned XGBoost model on the separate test set. It reached 97.86% accuracy, 94.74% precision, 97.30% recall, and 96.00% F1-score. We also checked ROC-AUC and PR-AUC. They were 99.84% and 99.59%, respectively. Next, we ran an ablation study. When behavioral screening responses were removed or limited, performance dropped more than when demographic variables were changed. In other words, behavior carried most of the useful signal. Demographic features added only a small amount of extra gain. We then used SHAP to see what mattered most. The SHAP results pointed to the key behavioral variables. They also made it easier to understand why the model reached its predictions.

The results suggest that machine learning could help with ASD screening. The work also shows clear reasons for the model's choices. This framework is meant to assist with screening only. It is not built to make a final diagnosis. You should also read the findings with care, since they depend on the dataset that was used. In the next phase, the goal should be to use bigger and more varied data. It will also help to test the method on outside groups. More checks on how steady the model is are needed. Only after that should anyone think about real use in clinics.

## References

[1] R. Francese and X. Yang, "Supporting autism spectrum disorder screening and intervention with machine learning and wearables: a systematic literature review," *Complex & Intelligent Systems*, June 2021, doi: 10.1007/s40747-021-00447-1.

[2] F. Thabtah *et al.*, "Autism screening: an unsupervised machine learning approach," *Health Information Science and Systems*, vol. 10, no. 1, Sept. 2022, doi: 10.1007/s13755-022-00191-x.

[3] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," *Scientific Reports*, vol. 13, no. 1, p. 9605, June 2023, doi: 10.1038/s41598-023-35910-1.

[4] M. Liao, H. Duan, and G. Wang, "Application of Machine Learning Techniques to Detect the Children with Autism Spectrum Disorder," *Journal of Healthcare Engineering*, vol. 2022, p. e9340027, Mar. 2022, doi: 10.1155/2022/9340027.

[5] K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques," *SN Computer Science*, vol. 2, no. 5, July 2021, doi: 10.1007/s42979-021-00776-5.

[6] Shyam Sundar Rajagopalan, Y. Zhang, A. Yahia, and Kristiina Tammimies, "Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information," *JAMA Network Open*, vol. 7, no. 8, pp. e2429229-e2429229, Aug. 2024, doi: 10.1001/jamanetworkopen.2024.29229.

[7] K. S. Betts, K. Chai, S. Kisely, and R. Alati, "Development and validation of a machine learning-based tool to predict autism among children," *Autism Research*, Mar. 2023, doi: 10.1002/aur.2912.

[8] H. Lu, H. Zhang, Y. Zhong, X.-Y. Meng, M.-F. Zhang, and T. Qiu, "A machine learning model based on CHAT-23 for early screening of autism in Chinese children,"

*Frontiers in Pediatrics*, vol. 12, Sept. 2024, doi: 10.3389/fped.2024.1400110.

[9] Jana Christina Koehler *et al.*, "Machine learning classification of autism spectrum disorder based on reciprocity in naturalistic social interactions," *Translational psychiatry*, vol. 14, no. 1, Feb. 2024, doi: 10.1038/s41398-024-02802-5.

[10] Wasana Yuwattana *et al.*, "Machine learning of clinical phenotypes facilitates autism screening and identifies novel subgroups with distinct transcriptomic profiles," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-95291-5.

[11] R. Agrawal and R. Agrawal, "Explainable AI in early autism detection: a literature review of interpretable machine learning approaches," *Discover Mental Health*, vol. 5, no. 1, July 2025, doi: 10.1007/s44192-025-00232-3.

[12] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv.org*, Nov. 24, 2017. https://arxiv.org/abs/1705.07874v2 (accessed Sept. 10, 2026).

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *arXiv (Cornell University)*, July 2019, doi: 10.48550/arxiv.1907.10902.

[14] M. Shadan, A. Sai Chaithanya, C. Vaishali, S. Mohammed Musharraf, and M. Gulam Muqeeth, "Hybrid BiLSTM-CNN Model with Attention and XAI (SHAP & LIME) for ECG Arrhythmia Classification Using PTB-XL," *International Journal of Science and Research (IJSR)*, pp. 2242–2249, Dec. 2025, doi: 10.21275/sr251226130813.