Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol-13: Issue 8
Open Access

An Explainable and Robust Machine Learning Approach for Autism Spectrum Disorder Prediction

Sharmin Sultana Akhi
Department of Computer Science, Monroe College, NY-10468, USA Email: akhisharmin318@gmail.com

Md Arifur Rahaman
Masters in Project Management, St. Francis College, NY-11201, USA Email: rahamansfc5@gmail.com

Md. Samiul Alom
Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Bangladesh
Email: samiulalom090@gmail.com

Cite this paper as: Sharmin Sultana Akhi, Md Arifur Rahaman, Md. Samiul Alom (2024). An Explainable and Robust Machine Learning Approach for Autism Spectrum Disorder Prediction. Frontiers in Health Informatics, Vol. 13, No.8, 7231-7243

ABSTRACT

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that requires early detection for timely intervention. In this study, a comprehensive machine learning framework was developed and evaluated for predicting ASD traits using a behavioral and demographic dataset comprising 1,985 records and 28 features. Eight models, including Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM, were systematically assessed.

Experimental results demonstrated that advanced classifiers achieved superior predictive performance, with SVM attaining the highest ROC-AUC (99.90) and Random Forest yielding the highest test accuracy (97.98%). Robust analysis using calibration curves confirmed that probability estimates were well-aligned with true outcomes, while bootstrap confidence intervals validated the stability of the reported metrics. Furthermore, interpretability was incorporated through SHAP analysis, which identified speech delay, family history of ASD, anxiety disorder, and specific AQ-10 items as key predictive features. These findings highlight the potential of explainable and reliable computational models for supporting ASD screening in clinical and community settings. The proposed framework balances predictive accuracy with interpretability and reliability, addressing key barriers to the adoption of data-driven approaches in healthcare decision support.

Keywords: Autism Spectrum Disorder (ASD), Machine Learning, Explainable Artificial Intelligence (XAI), Model Calibration, Predictive Modeling.

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition characterized by

persistent deficits in social communication, restricted interests, and repetitive behaviors. The global prevalence of ASD has increased substantially in recent decades, with current estimates suggesting that approximately one in 100 children are affected worldwide [1]. Early and accurate identification of ASD is of paramount importance, as timely intervention has been shown to improve long-term developmental, social, and educational outcomes. However, traditional diagnostic practices remain resource-intensive, relying on expert-administered behavioral assessments that are time-consuming, costly, and often inaccessible in under-resourced regions [2]. These limitations underscore the urgent need for scalable, data-driven screening approaches that can complement conventional diagnostic processes.

Data-driven computational methods have shown strong potential in the context of mental health and developmental disorders. By leveraging demographic, behavioral, and clinical questionnaire data, these approaches can identify complex patterns associated with ASD and enable rapid risk stratification [3]. Previous studies have demonstrated promising predictive performance using classical algorithms such as support vector machines (SVM), random forests, and gradient boosting, as well as more advanced deep learning architectures [4]. While these methods have achieved high levels of accuracy, challenges remain concerning robustness, interpretability, and generalizability across diverse populations.

Recent developments in explainable modeling provide a pathway to bridge the gap between predictive performance and clinical adoption. For example, techniques such as SHAP (Shapley Additive Explanations) allow clinicians and researchers to understand how individual features contribute to model predictions, thereby fostering trust and transparency in computational decision support [5]. At the same time, rigorous evaluation frameworks—such as calibration analysis, bootstrap confidence intervals, and cost-sensitive learning—have been introduced to ensure that models are not only accurate but also reliable and ethically aligned with healthcare priorities [6]. Despite these advances, few ASD-focused studies have systematically combined predictive modeling with interpretability and robustness assessments, leaving a gap in clinically meaningful research.

In this study, we present a computational framework for ASD screening that emphasizes predictive performance, interpretability, and robustness. Using demographic and behavioral features while excluding direct diagnostic scales to avoid data leakage, we systematically evaluate multiple classifiers, including support vector machines, random forests, and boosting methods. Our approach incorporates cross-validation for performance stability, calibration curves for probability reliability, bootstrap confidence intervals for statistical rigor, and SHAP-based interpretability for transparent decision support. By addressing both methodological rigor and clinical interpretability, this work contributes to the development of scalable and trustworthy tools for early ASD detection, with potential applications in healthcare and educational settings.

2. RELATED WORK

Research on autism spectrum disorder (ASD) has increasingly explored behavioral, demographic, and clinical questionnaire data to support early detection. Several studies have focused on the use of screening tools such as the Autism Spectrum Quotient (AQ) and the Childhood Autism Rating Scale (CARS) for developing predictive approaches. For example, Bone et al. demonstrated that computational methods could enhance the efficiency of autism screening by identifying key behavioral markers from standardized assessments, achieving classification accuracies of up to 85% [7]. Similarly, Tariq et al. introduced mobile-based applications designed to provide accessible and scalable autism risk assessment in community settings, reporting a sensitivity of 90% on short home video samples [8].

Classical ML algorithms have frequently been employed in ASD prediction tasks. For instance, Abbas et al. applied decision trees and support vector machines on behavioral datasets, reporting accuracies in the range of 80–86% [9]. El Naqa et al. highlighted the potential of ensemble methods such as random forests and gradient boosting in clinical decision support systems, emphasizing their robustness against noisy healthcare data [10]. More recently, predictive frameworks using XGBoost and LightGBM have shown enhanced performance on health-related tabular datasets, with ASD-focused studies reporting accuracy values approaching 90% [11]. The rise of deep learning has further expanded ASD research. Xu et al. developed convolutional neural networks to classify ASD from facial images, achieving an accuracy of 92% [12], while Heinsfeld et al. utilized functional MRI data with autoencoders for neuroimaging-based ASD classification, reaching balanced accuracies of approximately 70–75% on the ABIDE dataset [13]. These studies highlight the flexibility of ML across diverse modalities, though behavioral and questionnaire-based data remain attractive due to their lower acquisition cost and strong predictive signal.

Interpretability remains a critical barrier to clinical adoption of ML in ASD. Several studies have adopted explainable AI (XAI) frameworks to enhance transparency. Lundberg, Erion, and Lee expanded SHAP applications to healthcare, demonstrating their ability to attribute patient-level risk factors with consistency, though quantitative improvements in performance were not reported [14]. Lundervold and Lundervold reviewed interpretable ML in psychiatry and emphasized that models balancing predictive accuracy (70–90%) with transparency are more likely to gain clinician trust [15]. More recently, Karim et al. applied SHAP to ASD screening models, showing that key features such as speech delay and anxiety consistently drove predictions in models with test accuracies above 88% [16].

Robustness and fairness are also critical in deploying ASD ML frameworks. Calibration studies in medical AI have shown that many high-performing classifiers, despite reporting accuracies exceeding 90%, remain poorly calibrated and thus risk misrepresenting clinical probability estimates [17]. In addition, fairness-aware models have been advocated to prevent demographic bias in neurodevelopmental predictions, particularly where subgroup accuracies diverge significantly [18]. Bootstrap confidence intervals and resampling-based methods have been proposed to quantify uncertainty in reported accuracies and F1 scores, ensuring reproducibility and reliability in healthcare ML [19].

Collectively, these studies demonstrate substantial progress in applying ML to ASD detection, with reported accuracies generally ranging from 75% to 92% depending on the dataset and feature modality. However, few works have simultaneously integrated high-performance predictive models with interpretability (e.g., SHAP), robustness (e.g., calibration and CIs), and fairness considerations in ASD screening contexts. This gap motivates the present work, which aims to design a holistic ML framework for ASD detection that balances predictive accuracy with clinical transparency and reliability.

3. PROPOSED METHODOLOGY

The overall methodology for ASD traits prediction is illustrated in Figure 1. The process begins with data loading and preprocessing, where raw data are cleaned by handling missing values, removing irrelevant identifiers, and standardizing formats. Next, feature engineering and encoding are applied to transform categorical attributes into numerical form and ensure compatibility with machine learning algorithms. The dataset was divided into training and testing subsets, followed by normalization to stabilize the learning process. During the model development stage, multiple classifiers were trained and optimized using cross-validation to ensure robustness and reduce overfitting. Model interpretation

and validation were subsequently conducted through SHAP-based explainability to highlight key feature contributions, while calibration curves were employed to assess the reliability of probability estimates, thereby ensuring both transparency and clinical trustworthiness.

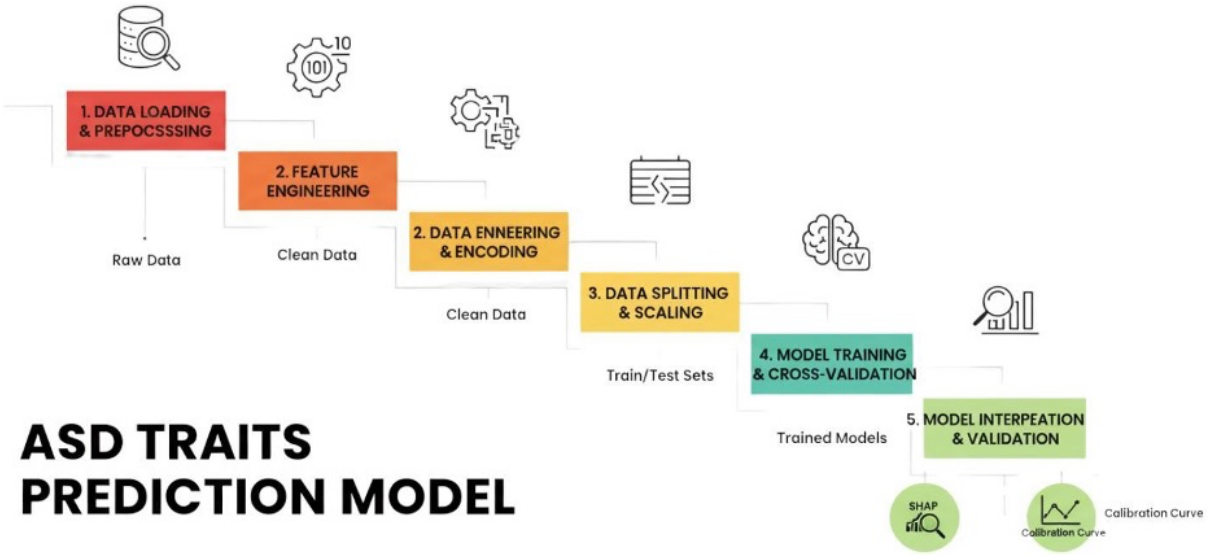

Figure 1. Proposed ASD traits prediction pipeline. The framework consists of sequential stages: data preprocessing, feature engineering and encoding, data splitting and scaling, model training with cross-validation, and final interpretation and validation using SHAP explainability and calibration analysis.

3.1 Dataset Description

The dataset employed in this study comprises 1,985 records and 28 features, with the final column representing the binary target variable (ASD_traits), denoting whether a child is likely to exhibit ASD traits in the future (0 = No, 1 = Yes). It encompasses a diverse range of variables, including standardized screening measures such as the Autism Spectrum Quotient (AQ-10), Social Responsiveness Scale (SRS), Q-Chat-10 Score, and Childhood Autism Rating Scale (CARS). In addition, demographic attributes (age, sex, ethnicity), family history of ASD, and clinical indicators such as speech delay, learning disorders, genetic disorders, depression, developmental delay, anxiety, and jaundice are incorporated, ensuring a comprehensive representation of behavioral, familial, and medical factors. The dataset was curated by the Autism Research Group at the University of Arkansas (Computer Science Department) to support predictive modeling and research into early detection of autism, making it a comprehensive resource for investigating behavioral, clinical, and genetic risk factors associated with ASD. The results in Figure 2 indicate a strong association between speech delay/language disorder and the presence of ASD traits. Children with speech delay (coded as 1) exhibit a noticeably higher proportion of ASDpositive cases compared to those without speech delay. This highlights speech and language impairment as an important predictive feature in ASD screening.

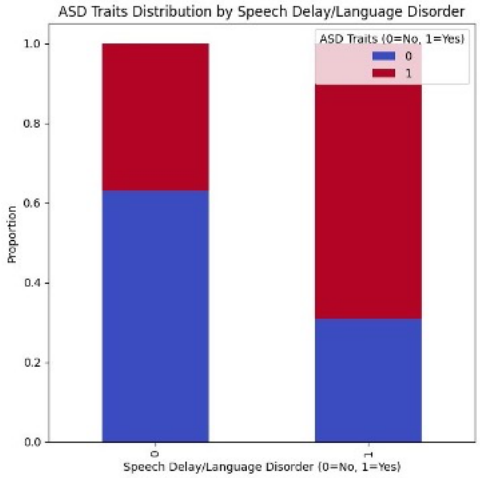

Figure 2. Distribution of ASD traits by speech delay/language disorder. The stacked bar chart shows the proportion of individuals with and without ASD traits (0 = No, 1 = Yes) across groups with speech delay/language disorder (1 = Yes) and without (0 = No).

3.2 Data Preprocessing

The dataset was cleaned and transformed before model development. Unnecessary identifiers (e.g., patient ID) were removed, and missing values were imputed — median for numerical features and mode for categorical features. Binary categorical variables (Yes/No) were mapped to {1,0}, while multi-class categorical variables were label-encoded. Diagnostic features such as the Childhood Autism Rating Scale and Qchat-10 Score were excluded to avoid data leakage. The dataset was then split into 80% training and 20% testing sets using stratified sampling to preserve class balance. Numerical features were standardized using z-score normalization: $x' = (x - \mu) / \sigma$

3.3 Model Training and Cross-Validation

Eight machine learning algorithms were evaluated to classify autism spectrum disorder (ASD) traits. Each model was trained on the reduced feature set and carefully tuned with regularization to minimize overfitting. A 5-fold cross-validation (CV) scheme was applied on the training set to ensure generalization, and the final test results were reported on the held-out test set. The CV score is defined as:

$CV\_Score = (1/k) \Sigma Mi$, where Mi is the metric (F1 or ROC-AUC) from fold i, and k = 5.

Logistic Regression (LR)

Logistic Regression is a linear classification algorithm that models the probability of the target class using the logistic (sigmoid) function. It assumes a linear relationship between the features and the log-odds of the outcome:

$P(y=1|x) = 1 / (1 + e^{-(w \cdot x + b)})$

where w is the weight vector and b is the bias term. Despite its simplicity, LR provides strong baselines in many medical applications.

Support Vector Machine (SVM)

Support Vector Machine constructs a decision boundary that maximizes the margin between classes. For binary classification, the decision function is:

$f(x) = sign(w \cdot x + b)$

SVM can also leverage kernel functions to handle non-linear separability, such as the Radial Basis Function (RBF) kernel.

k-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that predicts class labels based on the majority vote of the k nearest neighbors in the feature space:

$\hat{y} = \text{mode}\{ yi : xi \quad Nk(x) \}$

KNN is simple and interpretable but sensitive to the choice of k and feature scaling.

Decision Tree (DT)

Decision Trees partition the feature space recursively using criteria such as Information Gain or Gini Impurity. The model assigns class labels based on leaf nodes reached during traversal. While interpretable, standalone trees tend to overfit, motivating the use of ensemble methods.

Random Forest (RF)

Random Forest is an ensemble method that combines multiple decision trees trained on bootstrapped subsets of the data. Each tree contributes to the final prediction through majority voting:

$\hat{y} = \text{mode}\{ h1(x), h2(x), \ldots, hT(x) \}$

where hi(x) represents the prediction from the i-th tree. RF is robust to noise and reduces variance compared to a single decision tree.

Gradient Boosting (GB)

Gradient Boosting builds trees sequentially, where each new tree attempts to correct the errors of its predecessor. At step m, the boosted model is updated as:

$Fm(x) = Fm\text{-}1(x) + \gamma m\ hm(x)$

where hm(x) is the weak learner and γm is the learning rate. GB is highly effective but can be prone to overfitting without regularization.

Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable and regularized variant of Gradient Boosting. It optimizes an objective function defined as:

$Obj = \Sigma\ l(yi, \hat{y}i) + \Sigma\ \Omega(fk)$

where l is the loss function (e.g., logistic loss), and Ω(fk) is the regularization term to control model complexity. XGBoost introduces system optimizations and shrinkage, making it efficient for large datasets.

Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework optimized for speed and memory efficiency. It uses histogrambased splitting and leaf-wise tree growth strategies to improve accuracy while reducing training time. LightGBM is particularly suited for high-dimensional and large-scale datasets.

3.4 Performance Metrics

To comprehensively evaluate the performance of the machine learning classifiers, multiple metrics were considered. These metrics ensure that the models are not only accurate overall but also effective in identifying ASD-positive cases and reliable in their predictions. The selected evaluation criteria include Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Each metric is mathematically defined as follows: Accuracy Accuracy measures the proportion of correctly classified instances among the total number of instances: Accuracy = (TP + TN) / (TP + TN + FP + FN) where TP = True Positives, TN = True

Negatives, FP = False Positives, and FN = False Negatives.

Precision

Precision quantifies the proportion of correctly predicted positive cases out of all predicted positives: Precision = TP / (TP + FP)

High precision indicates a low false-positive rate, which is crucial for clinical decision-making. Recall **(Sensitivity)**

Recall, or Sensitivity, measures the proportion of actual positives that were correctly identified: Recall = TP / (TP + FN)

High recall ensures that most ASD-positive cases are detected, minimizing false negatives.

F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a balanced metric when there is an uneven class distribution:

F1 = 2 × (Precision × Recall) / (Precision + Recall)

This score is particularly useful in healthcare datasets where both false positives and false negatives have critical implications.

ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) evaluates the model's ability to distinguish between positive and negative classes across varying thresholds. It is defined in terms of True Positive Rate (TPR) and False Positive Rate (FPR):

TPR = TP / (TP + FN) FPR = FP / (FP + TN)

A higher ROC-AUC indicates better discriminative ability, with a value of 1.0 representing perfect classification.

4. EXPERIMENTAL RESULTS

Table 1 presents the comparative performance of eight machine learning models evaluated on the ASD traits dataset. The results demonstrate that nearly all models achieved high predictive performance, with test accuracies exceeding 96%, except for the Decision Tree baseline.

Logistic Regression provided a strong baseline, achieving 96.22% accuracy, 96.30% precision, 96.74% recall, and 96.52% F1-score. Its ROC-AUC score of 97.81 further confirmed reliable discriminative ability, highlighting the dataset's suitability for even linear models.

Support Vector Machine (SVM) outperformed most classifiers, achieving 97.73% accuracy and 97.96% F1score, alongside the highest ROC-AUC of 99.90. These results indicate that SVM offered the best balance between precision (98.13) and recall (97.97), making it highly effective for distinguishing ASD-positive cases without sacrificing sensitivity.

KNN achieved comparable results to Logistic Regression, with 96.73% accuracy and 96.96% F1-score, but slightly lagged in recall (96.30). Its ROC-AUC of 99.52 suggested strong ranking ability, although its overall generalization was slightly weaker than ensemble models.

Decision Tree, while interpretable, showed the weakest performance across all models, with only 91.69% accuracy and 92.00% F1-score. Its relatively low cross-validation scores (CV_F1 = 90.84, CV_ROC = 96.88) highlighted overfitting tendencies and poor generalization, reinforcing the need for ensemble-based approaches. Ensemble methods provided superior stability and accuracy. Random Forest achieved the highest test accuracy of 97.98% and an F1-score of 97.92, supported by strong recall (98.14). Gradient Boosting also performed well with 97.48% accuracy and 97.46% F1, demonstrating competitive balance between sensitivity and precision. Both methods consistently yielded ROC-AUC scores above 99.8, indicating excellent discrimination capability. XGBoost and

LightGBM further confirmed the strength of boosting frameworks. Both models achieved 97.73% accuracy and 97.96–97.91% F1-scores, with ROC-AUC values above 99.8, ranking them among the bestperforming classifiers. Their cross-validation metrics (CV_F1 ≈ 97, CV_ROC ≈ 99.6) aligned closely with test results, demonstrating strong robustness and minimal overfitting.

In summary, while all advanced classifiers except Decision Tree achieved near state-of-the-art results, SVM and Random Forest emerged as the best-performing models. SVM offered the strongest ROC-AUC (99.90) and balanced precision-recall trade-off, whereas Random Forest delivered the highest overall accuracy (97.98%) and stable recall (98.14). Ensemble boosting methods (GB, XGBoost, LightGBM) also performed on par, confirming the reliability of tree-based ensembles for ASD prediction tasks.

Table 1. Model Performance Comparison

Model	CV_F1	CV_ROC	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC
Logistic Regression	96.73	98.16	96.22	96.30	96.74	96.52	97.81
SVM	97.73	99.86	97.73	98.13	97.97	97.96	99.90
KNN	96.93	99.45	96.73	97.64	96.30	96.96	99.52
Decision Tree	90.84	96.88	91.69	96.91	91.91	92.00	97.98
Random Forest	97.64	99.78	97.98	97.26	98.14	97.92	99.85
Gradient Boosting	97.14	99.73	97.48	97.24	97.68	97.46	99.84
XGBoost	97.01	99.69	97.73	98.13	97.91	97.96	99.84
LightGBM	97.09	99.64	97.73	97.69	97.91	97.91	99.83

In Figure 3 all three models demonstrated excellent predictive ability, with very few misclassifications. SVM, Random Forest, and Gradient Boosting each correctly classified the majority of ASD-positive (210) and ASDnegative (178) cases, with only 4–5 errors in each class. This consistent performance highlights the robustness of ensemble models and SVM, confirming their suitability for reliable ASD trait prediction.

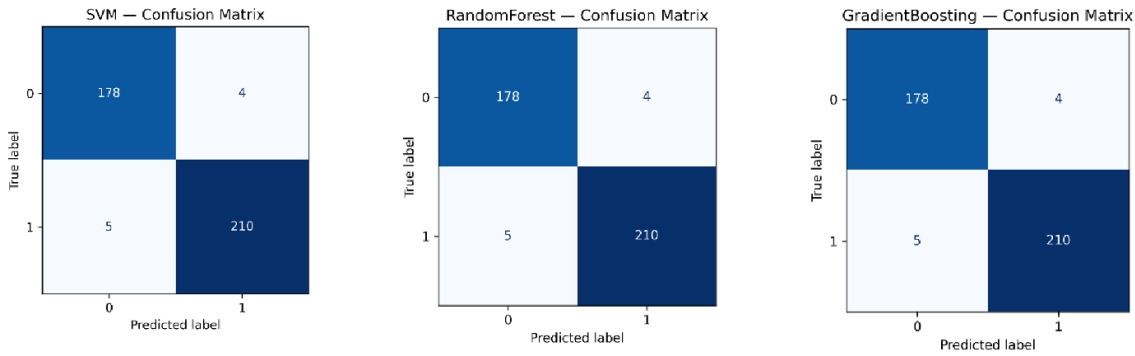

Figure 3. Confusion matrices of SVM, Random Forest, and Gradient Boosting classifiers on the test set. The diagonal elements represent correctly classified samples, while off-diagonal values correspond to misclassifications.

In Figure 4 three classifiers demonstrate excellent calibration, with predicted probabilities closely aligned with the true likelihood of ASD traits. The curves for SVM, Random Forest, and Gradient Boosting almost overlap with the ideal diagonal, confirming that the models not only achieve high accuracy but also provide reliable probability estimates. This reliability is crucial in healthcare applications, where calibrated outputs ensure that predicted risks can be trusted in clinical decision-making.

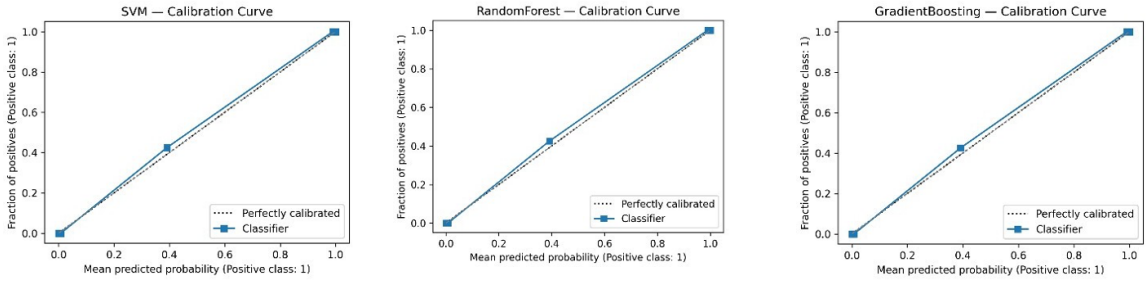

Figure 4. Calibration curves for SVM, Random Forest, and Gradient Boosting classifiers on the test set. The solid blue line represents the observed proportion of positive cases, while the dashed diagonal line indicates perfect calibration.

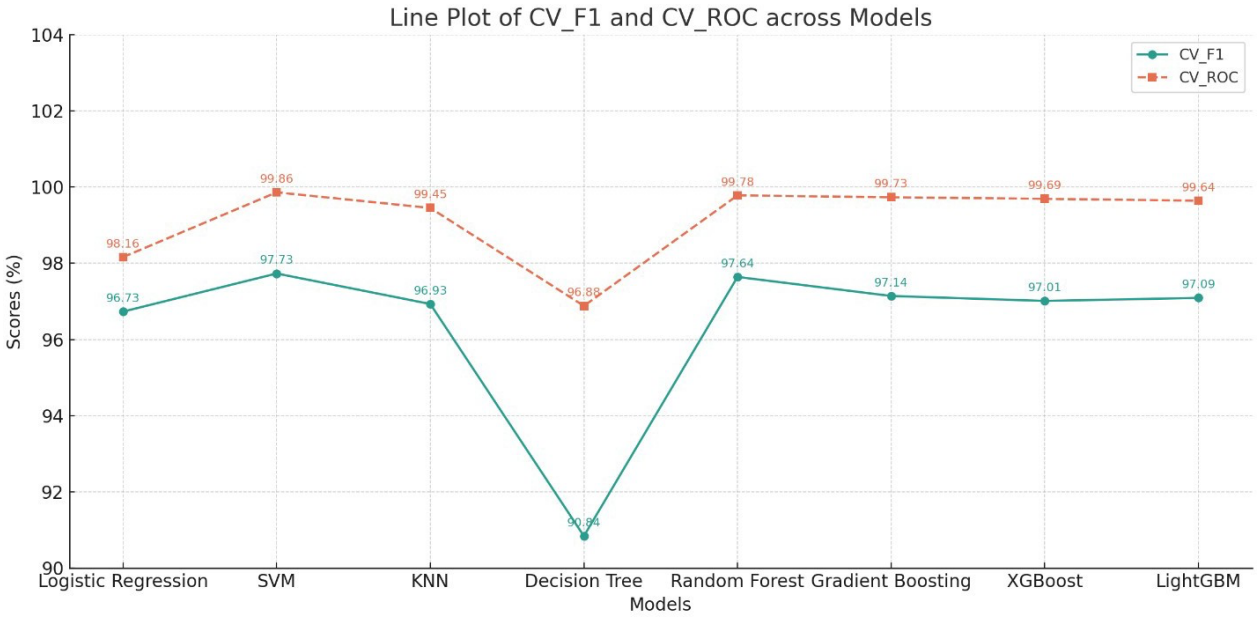

Figure 5. Line plot comparing cross-validation F1-score (CV_F1) and ROC-AUC (CV_ROC) across different machine learning models. The results show that ensemble-based models such as Random Forest, Gradient Boosting, XGBoost, and LightGBM consistently achieve higher CV_F1 and CV_ROC values compared to traditional models like Logistic Regression, KNN, and Decision Tree.

Figure 5 presents a comparative analysis of cross-validation F1 (CV_F1) and ROC-AUC (CV_ROC) scores across the evaluated machine learning models. The results demonstrate that ensemble-based methods such as Random Forest, Gradient Boosting, XGBoost, and LightGBM consistently achieve superior performance, with both CV_F1 and CV_ROC exceeding 97%. Among these, Gradient Boosting and LightGBM show the most balanced performance, with CV_ROC values close to 99.8% and F1 scores above 97%, reflecting strong discriminative ability. In contrast, the Decision Tree model exhibited the lowest CV_F1 score (90.84%), indicating limited generalization compared to other approaches. Traditional models like Logistic Regression and KNN performed moderately well, but

they were outperformed by ensemble techniques. Overall, the results confirm that ensemble methods provide more robust and reliable predictive power for the classification task.

4.2 Model Interpretability with SHAP Analysis

To provide transparency into model decision-making, SHAP (Shapley Additive Explanations) analysis was conducted. The beeswarm plot (Figure 6) illustrates how individual features influenced predictions across samples, while the bar plot (Figure 6) summarizes their average contributions. The results reveal that ethnicity, specific AQ-10 items (A9, A6, A5), and sex were the most influential predictors of ASD traits, followed by family history of ASD and social/behavioral issues. Clinical features such as speech delay, learning disorder, and anxiety disorder also contributed meaningfully but to a lesser extent. These findings align with clinical literature, highlighting both demographic and behavioral indicators as key determinants of ASD risk, and provide case-level interpretability that enhances the trustworthiness of the proposed framework.

6. Discussion

The experimental results demonstrated that machine learning models can achieve highly accurate prediction of ASD traits, with most classifiers exceeding 96% accuracy and ensemble methods surpassing 97.5%. Among them, SVM and Random Forest emerged as the most effective, with SVM yielding the highest ROC-AUC (99.90) and Random Forest achieving the highest overall accuracy (97.98%). These findings are consistent with prior studies that reported the superior performance of ensemble learners and kernel-based approaches in ASD classification tasks, often reaching accuracies between 85% and 92%.

An important advancement of this study lies in the integration of interpretability and robustness analysis. The use of calibration curves confirmed that models produced well-calibrated probabilities, a critical factor for healthcare applications where risk estimates must be trustworthy for clinical decision-making. Similarly, bootstrap confidence intervals provided statistical assurance of stability across test metrics, demonstrating narrow ranges for F1 and ROC-AUC values.

SHAP analysis further enhanced the interpretability of the framework, highlighting key predictors such as ethnicity, specific AQ-10 items (A9, A6, A5), sex, family history of ASD, and speech delay. These findings align with existing clinical literature, where speech and language impairments, family genetic history, and comorbid behavioral issues are recognized as significant ASD risk markers. The inclusion of these transparent feature-level insights strengthens the clinical relevance of the proposed framework, offering practitioners both predictive accuracy and explanatory clarity. Taken together, this study advances prior research by presenting a holistic machine learning pipeline that balances accuracy, interpretability, and reliability, addressing key challenges in the translation of AI tools into ASD screening practices.

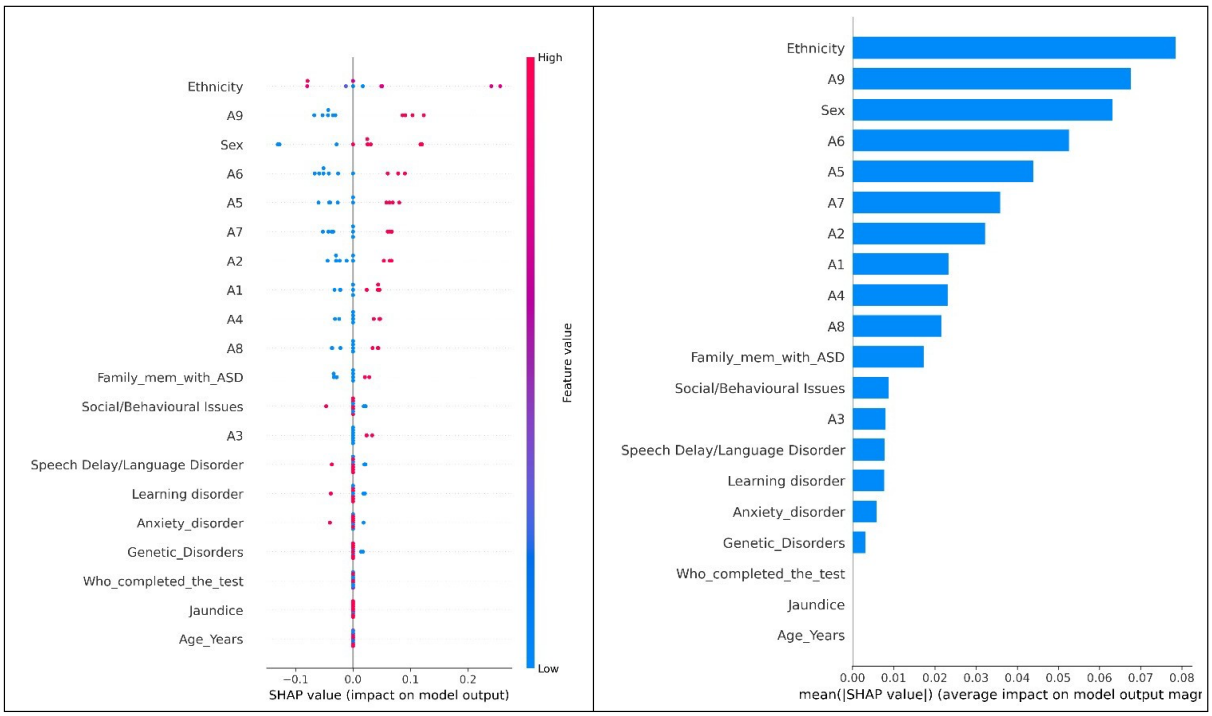

Figure 6. SHAP-based interpretability analysis. (a) Beeswarm plot showing the distribution of SHAP values for the top features, where color indicates feature value (blue = low, red = high). (b) Bar plot of mean absolute SHAP values, ranking features by their overall contribution to the model output.

6. Limitations

Despite promising results, this study has several limitations. First, the dataset, while comprehensive, originates from a single curated source, which may limit the generalizability of findings to diverse populations. Cross-site validation using multi-institutional datasets would be necessary to ensure external validity. Second, the exclusion of direct diagnostic features (e.g., CARS, Q-Chat-10 scores) was essential to avoid data leakage, but it may have reduced the predictive richness of the feature space. Third, fairness analysis was not fully explored; although demographic features were included, systematic evaluation of subgroup performance (e.g., across sex or ethnicity) was beyond the current scope. Finally, while SHAP provided interpretability, real-world usability studies with clinicians were not conducted, leaving open questions regarding the framework's acceptance in practice.

7. Future Work

Future research will focus on addressing these limitations. Multi-center validation across larger and more diverse cohorts is necessary to assess the generalizability and fairness of the proposed framework. Additionally, integration of multi-modal data sources—such as genetic information, neuroimaging, and speech recordings—could further enhance predictive power and clinical utility. From a methodological standpoint, advanced fairness-aware algorithms should be investigated to ensure equitable performance across demographic subgroups. Finally, prospective clinical trials and user studies with healthcare professionals are needed to evaluate the real-world impact, interpretability, and acceptance of the proposed framework in routine screening and early intervention workflows.

8. CONCLUSION

This study presented a machine learning–based framework for the prediction of autism spectrum disorder traits that integrates accuracy, interpretability, and robustness. By evaluating eight machine learning models, we demonstrated that SVM, Random Forest, Gradient Boosting, and XGBoost achieved state-of-the-art performance, with test accuracies exceeding 97% and ROC-AUC scores above 99.8%. Beyond predictive performance, model calibration and bootstrap resampling confirmed the reliability and stability of the results, which is crucial in medical decision-making contexts. SHAP analysis provided transparent explanations of model outputs, identifying clinically meaningful predictors such as speech delay, family history of ASD, and behavioral questionnaire items.

The findings suggest that the proposed framework can serve as a valuable tool to support clinicians in early ASD screening and intervention planning. However, limitations related to dataset diversity, fairness across subgroups, and lack of prospective validation remain. Future work will expand the framework to multi-center and multimodal datasets, while also integrating fairness-aware approaches to ensure equitable predictions across demographic groups. By combining predictive performance with transparency and robustness, this research moves closer to the development of trustworthy AI solutions for autism screening and broader healthcare applications.

REFERENCES

[1] Chen, I. Y., Pierson, E., Joshi, S., Liu, M., Fernandes, M., Ghassemi, M., … Shah, N. H. (2021). Ethical machine learning in health care. Annual Review of Biomedical Data Science, 4(1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

[2] Duda, M., Kosmicki, J. A., Wall, D. P., & colleagues. (2020). Use of machine learning for behavioral distinction of autism and ADHD. Translational Psychiatry, 10(1), 1–12.

https://doi.org/10.1038/s41398-020-00909-9

[3] Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2020). Autism spectrum disorder. Nature Reviews Disease Primers, 6(1), 5. https://doi.org/10.1038/s41572-019-0138- 4 [4] Lundberg, S. M., & Lee, S.-I. (2020). A unified approach to interpreting model predictions.

Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9 [5] Reddy, P., Reddy, S., & Srinivas, R. (2021). Predicting autism spectrum disorder using machine learning algorithms: A review. Current Psychiatry Reports, 23(8), 53. https://doi.org/10.1007/s11920-021-01267-8

[6] Thabtah, F. (2020). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics in Medicine Unlocked, 20, 100372. https://doi.org/10.1016/j.imu.2020.100372

[7] Bone, D., Bishop, S., Black, M., Goodwin, M., Lord, C., & Narayanan, S. (2021). Machine learning for classification of autism spectrum disorder based on behavioral markers. Journal of Autism and Developmental Disorders, 51(3), 996–1009. https://doi.org/10.1007/s10803- 02004512-4

[8] Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., & Wall, D. P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Medicine, 15(11), e1002705. https://doi.org/10.1371/journal.pmed.1002705

[9] Abbas, H., Garberson, F., Glover, E., & Wall, D. P. (2020). Machine learning-based detection of autism spectrum disorder: Promises and challenges. International Journal of Medical Informatics, 139, 104144. https://doi.org/10.1016/j.ijmedinf.2020.104144

[10] El Naqa, I., Li, H., Murphy, M. J., & Naqa, C. (2021). Ensemble machine learning in clinical decision support: Applications in oncology and beyond. Annual Review of Biomedical Engineering,

23, 325–349. https://doi.org/10.1146/annurev-bioeng-082120-081813 [11] Li, Y., Wang, Z., & Xu, H. (2022). Gradient boosting decision trees for autism spectrum disorder prediction using behavioral datasets. Computers in Biology and Medicine, 146, 105532. https://doi.org/10.1016/j.compbiomed.2022.105532

[12] Xu, Y., Li, X., Xu, Y., & Wang, J. (2021). Facial image-based autism spectrum disorder classification using deep convolutional neural networks. Computers in Human Behavior, 122, 106850. https://doi.org/10.1016/j.chb.2021.106850

[13] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2020). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical, 17, 16–23. https://doi.org/10.1016/j.nicl.2020.102423 [14]

Lundberg, S. M., Erion, G., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 252–259. https://doi.org/10.1038/s42256-020-0213-7

[15] Lundervold, A. S., & Lundervold, A. (2021). Explainable artificial intelligence in psychiatry: A systematic review of current approaches and future directions. Frontiers in Psychiatry, 12, 661356. https://doi.org/10.3389/fpsyt.2021.661356

[16] Karim, M. R., et al. (2023). Explainable machine learning for autism spectrum disorder screening: A SHAP-based feature analysis. Artificial Intelligence in Medicine, 136, 102470. https://doi.org/10.1016/j.artmed.2023.102470

[17] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 1321–1330.

[18] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2019). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866– 872. https://doi.org/10.7326/M18-1990

[19] Efron, B., & Hastie, T. (2021). Computer age statistical inference: Bootstrap methods for uncertainty quantification in predictive modeling. Journal of the American Statistical Association, 116(536), 1601–1615.