JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

EXPLAINABLE AND FAIRNESS-AWARE MACHINE LEARNING FOR EARLY AUTISM RISK SCREENING USING BEHAVIORAL AND DEMOGRAPHIC DATA

Dr. Hira Chaudhry[1], Abu Bakar Shabbir[2], Faisal Amjad[*3], Khadija Amber[4]

[1]*MBBS, MSc. (Psychological Medicine), Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Malaysia*
[1]*Assistant Professor, Tashkent State Medical University, Uzbekistan*
[2]*School of Systems and Technology, University of Management and Technology, Lahore - 54770, Punjab, Pakistan.*
[3]*PhD Candidate (Special Education), Department of Special Education (DSE), Division of Education (DoE), University of Education (UoE), Lahore, Punjab, Pakistan.*
[4]*Department of Computer Science, Bahauddin Zakariya University, Bosan Road, Multan - 60800, Punjab, Pakistan.*

[1]hiraa.co.uk@gmail.com, [2]abubakarshabbir64@gmail.com, [3]amjadfaisal40@gmail.com/ pds24000237@ue.edu.pk, [3]khadijaamber44@gmail.com

[1]https://orcid.org/0009-0009-2640-738X, [3]https://orcid.org/0009-0002-7617-5395

DOI: https://doi.org/10.5281/zenodo.23100937

Keywords
Autism Spectrum Disorder (ASD); Early Detection; Machine Learning; Explainable Artificial Intelligence (XAI); SHAP; Fairness-Aware Machine Learning; ASD Screening; Behavioral Features; Demographic Bias; Clinical Artificial Intelligence.

Article History
Received: 25 June 2026
Accepted: 16 September 2026
Published: 02 October 2026

Copyright @Author
Corresponding Author: * Faisal Amjad

Abstract
Early detection of Autism Spectrum Disorder (ASD) is essential to allow interventions that can significantly improve developmental trajectories. This paper introduces an understandable and fairness, aware machine learning approach to screening the risk of ASD at the early stage, based on behavioral and demographic data. The study used a publicly available dataset of 699 adult samples, including AQ, 10 behavioral features, and demographic attributes, to conduct experiments. Five supervised models, Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, and K-Nearest Neighbors, were tested. Logistic Regression was found to be the best model in terms of the trade-off of accuracy (96. 43%), recall of ASD, positive cases, and interpretability. The model was made transparent through Shapley Additive Explanations (SHAP), which allowed understanding both at the global and instance level, of the most relevant to clinicians' behavioral indicators. The fairness of the model regarding gender was measured through demographic parity, equal opportunity, and disparate impact criteria, showing some moderate bias, which was reduced through the reweighing method without causing any performance drop. The ablation study showed that behavioral features are sufficient for accurate and fair ASD screening. The proposed framework is an example of a responsible clinical AI deployment approach.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Introduction

Autism Spectrum Disorder (ASD) refers to a neurodevelopmental condition involving social, communication difficulties, limited interests, and the persistence of repetitive and stereotyped behaviors. Based on current clinical consensus, one of the most important factors contributing to better outcomes in the long run (cognitive, behavioral, social) for children with ASD is early detection and therapeutic intervention. The literature is replete with positive outcomes when treatment is introduced in the earliest stages of development. Unfortunately, while recognition is better, there are still many locales where ASD is under-recognized and diagnosed late due to diagnostic challenges. The use of standard diagnostic tests, for example, the Autism Diagnostic Observation Schedule (ADOS), and the Autism Diagnostic Interview-Revised (ADI-R), has been clinically proven and is accepted as the "gold standard" of diagnosis. These diagnostic instruments, while considered diagnostically valid and reliable, are nevertheless expensive in terms of the resources required, take a long time, and are inaccessible to low-resource settings due to the fact that they require trained specialists (Agrawal & Agrawal, 2025). Furthermore, these standard tests are heavily dependent on the subjective nature of the clinician's interpretation of behavioral observations, possibly leading to differences between clinicians and cultural influences.

The drawbacks of standard clinical instruments have led to an exponential increase in the number of research papers proposing computational methods that can help in early and accurate diagnosis of ASD in a cost-effective way that can reach a larger population. Recent advances in artificial intelligence (AI) and machine learning (ML) have paved the path to automated ASD screening models by using behavioral, demographic, and biological data. A plethora of studies reveal that machine learning models trained on the behavioral questionnaires, especially AQ, 10 and Q, CHAT, 10can attain extremely high predictive accuracy in identifying ASD, positive and ASD, negative individuals (Pandya & Jain, 2024; Qureshi, 2025).

In particular, Jyoti et al. (2025) and Akhi et al. (2024) state that Logistic Regression, Random Forest, and Support Vector Machines, as the classical machine learning models, demonstrate remarkable results on tabular behavioral datasets; thus, they are sometimes at par or even better than deep learning architectures. The results speak for the fact that the simpler and more interpretable models may bring considerable benefits as opposed to the black box approaches for early screening tasks. Notwithstanding these promising outcomes, the fast deployment of AI-driven ASD screening instruments gives rise to a multitude of critical issues.

Among them, the most significant one is the problem of interpretability of those models that perform well in most cases. Black box predictions, in particular, those from ensemble or deep learning models, cause a great deal of problems in the implementation of clinical trust and accountability. In health care environments, clinicians and caregivers are obliged to know the reasons for the automated predictions if they are to make informed decisions and guarantee their compatibility with the already established diagnostic knowledge. The absence of transparent reasoning limits the clinical utility of AI systems and hampers their acceptance in real-world screening workflows (Agrawal & Agrawal, 2025).

Explainable Artificial Intelligence (XAI) is one of the key technologies that offers a solution to these difficulties. XAI methods seek to make machine learning models understandable to humans by delivering interpretable explanations of their predictions. Notably, Shapley Additive exPlanations (SHAP) and Local Interpretable Model, Agnostic Explanations (LIME) have become popular in the domain of XAI due to their solid theoretical underpinnings and wide applicability. Several studies show that ASD explanations based on SHAP can effectively pinpoint behavioral features that not only have clinical significance but also match the diagnostic criteria of ASD given in standard

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

manuals. g., social responsiveness, communication, and attention (Pandya & Jain, 2024; Eguavoen, 2025). XAI, which is implemented through global as well as local, instance-level decision logic, thus helps in making the workings of an AI model more transparent and contributes to the trust of clinicians in AI, AI-assisted screening systems.

The problem of fairness and ethics, however, lies beyond the issue of interpretability, and it constitutes another substantial challenge in the use of AI for ASD screening. Gender, ethnicity, and cultural background are some of the demographic factors that influence not only the prevalence, symptomatology, and diagnosis of ASD but also the diagnostic outcomes. Models trained on datasets that are imbalanced or culturally homogeneous may acquire spurious correlations without realizing it, and thus produce biased predictions that disadvantage the groups of people that are, at the same time, poorly represented and marginalized. The cutting-edge research exposes the risk of demographic bias in ASD screening models with a focus on gender and ethnicity (Rajagopalan et al., 2024; Priyadarshini, 2023). Such biases are especially concerning in early screening applications, where false negatives may delay critical intervention and false positives may cause unnecessary psychological and clinical burden.

Although some prior studies acknowledge these concerns, systematic fairness evaluation and mitigation remain underexplored in ASD screening research. Most existing works prioritize predictive accuracy while overlooking whether models perform equitably across demographic groups. Recent contributions emphasize the necessity of integrating fairness-aware AI techniques such as reweighing, bias mitigation, and group-wise performance evaluation into healthcare machine learning pipelines (Alsbakhi et al., 2025). Addressing fairness is not only an ethical imperative but also a practical requirement for regulatory compliance, clinical adoption, and societal trust.

Several studies have also examined the use of multimodal and deep learning methods for ASD detection from neuroimaging, facial images, and biological signals in parallel with behavioral screening approaches. Malik et al. (2025), for instance, introduced an explainable multimodal framework integrating behavioral phenotypes and neuroimaging data, thus achieving excellent predictive performance while retaining interpretability. In the same way, models based on facial images utilizing convolutional neural networks have claimed to achieve high accuracy when combined with explainability techniques such as Grad, CAM, and LIME (Atlam et al., 2025; Alam et al., 2023). However, these methods still face issues of difficult data acquisition, privacy concerns, and the challenge of generalizing to different populations. Therefore, questionnaire-based behavioral screening is still the most effective and scalable method for the early detection of ASD risk.

In response to the difficulties and possibilities identified, this research introduces a machine learning framework that is both explainable and fairness-aware for the early screening of ASD risk based on behavioral and demographic data. The new method goes beyond the previous studies that only focus on accuracy or interpretability separately; the proposed method combines predictive modeling, explainable AI, and fairness evaluation into one integrated pipeline. Various classical machine learning models are systematically tested to find a compromise between their performance and their transparency. To ensure clinical interpretability and compliance with established behavioral science, SHAP is used to offer both global and local explanations. In addition, demographic fairness is clearly gauged through several quantitative metrics, and bias mitigation techniques are utilized to lessen the disparities without impairing the predictive power of the model.

The study makes three major contributions to the field. A primary contribution is a substantial comparative evaluation of various machine learning models for the purpose of screening for Autism Spectrum Disorder (ASD). It is shown that interpretable models, in fact, can obtain

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

performance levels that are in competition with state, of, the, art models when applied to structured behavioral data. The second contribution is the comprehensive explainability of the analysis by means of SHAP. The explanations are thus transparent insights into the decision-making of the model and accessible both at the population and at the individual level. The third contribution is related to the systematic assessment of demographic bias and its subsequent removal. It is shown that fairness-aware interventions not only have the potential to increase the fairness of model predictions but also to maintain or even improve the effectiveness of the model. The paper, by simultaneously addressing accuracy, explainability, and fairness, paves the way for the deployment of AI systems in the early screening of autism that are reliable and ethical.

Overall, this study aims to bridge the gap between high-performance machine learning and ethical, trustworthy clinical deployment. By prioritizing transparency and fairness alongside predictive accuracy, the proposed framework contributes to the responsible integration of AI into early ASD screening and supports equitable, evidence-based decision-making in healthcare settings. To provide a clear overview of the proposed methodology, **Figure** 1 illustrates the conceptual workflow of the explainable and fairness-aware AI framework for early Autism Spectrum Disorder (ASD) screening, highlighting key components from traditional screening to AI-based risk assessment.

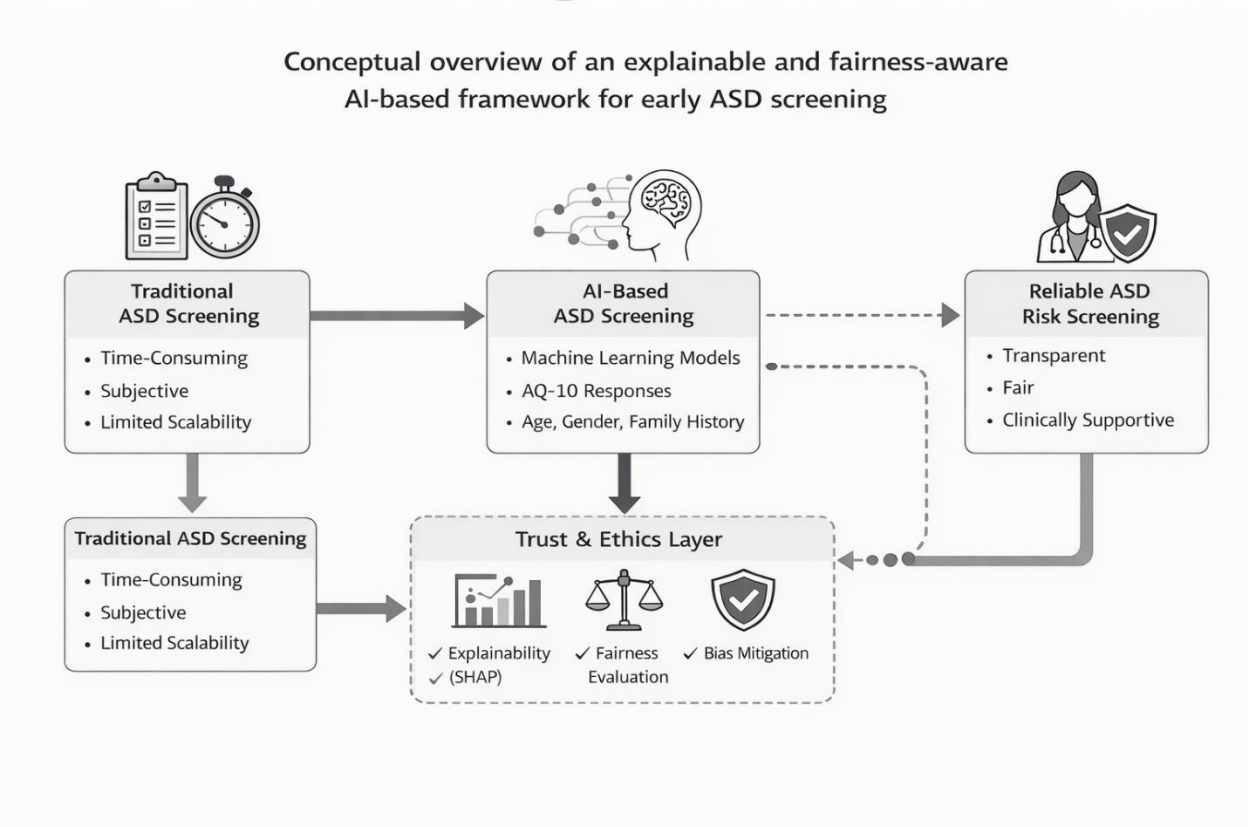


Figure 1: Conceptual overview of the proposed explainable and fairness-aware AI framework for early ASD detection, depicting traditional screening, AI-based screening, the Trust & Ethics layer, and reliable risk screening outcomes.

Literature Review

Identifying Autism Spectrum Disorder (ASD) in the earliest stages is essential to allow the interventions that, within the cognitive, social, and behavioral domains, can have a huge impact on the generalization of the

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

improvements. Nevertheless, such instruments as ADOS and ADI, R, though reliable from a clinical point of view, depend heavily on resources, are subjective, and are often located far away, which has led to the development of AI-based systems to screen ASD symptoms. Accordingly, it became easier to predict ASD with the introduction of advanced methods in machine learning (ML), deep learning (DL), and explainable artificial intelligence (XAI). Despite this, issues related to interpretability, fairness, and trust from the clinical community are still the core problems that need to be addressed.

Machine Learning–Based ASD Screening Using Behavioral Data

Behavioral questionnaire-based datasets have become the primary focus of ASD screening research, with a major emphasis on AQ, 10, and Q, CHAT, 10. Their affordability, scalability, and early age applicability have contributed to their widespread use. Classical ML models are shown to be effective in handling structured behavioral data in a number of research works. Pandya and Jain (2024) performed a comprehensive ML classifier on the AQ, 10 dataset and concluded the performance to be near-perfect in most cases, with Random Forest being a strong, interpretable option. Likewise, Qureshi (2025) found Logistic Regression and Random Forest to be the most accurate (up to 100%) methods on toddler datasets, thus drawing attention to the fact that deep learning may not necessarily be superior to simpler models when dealing with tabular behavioral data. The work of Akhi et al. (2024) goes a step further to confirm this finding by comparing the performance of eight ML models on an extensive behavioral, demographic dataset. They observed that ensemble models such as Random Forest and various boosting techniques consistently achieved better results than single classifiers. Additionally, in order to enhance clinical validity, diagnostic scales were removed so that data leakage could be avoided. Rajagopalan et al. (2024) took behavioral screening a step further by applying it to large-scale, real-world data from the SPARK database. They found that the XGBoost model could accurately predict ASD in different cohorts with minimal parent-reported features available before 24 months, thus showing strong AUROC scores and good generalization across cohorts.

Ensemble Learning and Class Imbalance Mitigation

Class imbalance has been a persistent problem in ASD datasets where non, ASD samples are usually in the majority. Numerous studies have pointed out oversampling and hybrid balancing techniques as the most effective methods for overcoming this problem. Eguavoen (2025) combined SMOTE with ensemble learning and, as a result, achieved great improvements in sensitivity and a significant reduction of bias toward majority classes in the case of four merged UCI datasets. Jyoti et al. (2025) used SMOTE, Tomek in balancing class distributions across different age groups and were able to obtain up to 99% balanced accuracy through the use of Random Forest models.

Aldrees et al. (2024) performed a comprehensive investigation of feature selection and oversampling strategies, and their findings revealed that the performance of ensemble models is significantly improved when Chi-square feature selection is combined with the SMOTE technique. Their upgraded XGBoost 2 0 model, which was able to achieve around 99% accuracy and good stability under cross-validation, is a strong indication of the screening method for ASD that is most effective when it uses a small number of selective features and balanced learning.

Explainable Artificial Intelligence for Clinical Trust

Black box behavior is still a major hurdle for clinical adoption despite high predictive accuracy. As a result, XAI has become the central focus of most of the recent ASD research. SHAP and LIME are the two most widely used techniques that provide both global

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

and local explanations. Pandya and Jain (2024) employed SHAP and LIME to explain Random Forest predictions, thus identifying gesture, related, gaze, following, and comfort, response AQ, 10 items as the most important features, which is in line with clinical understanding.

Eguavoen (2025) conducted a thorough SHAP-based analysis over ensemble models, pointing out that ethnicity, age, family history of ASD, and specific AQ, 10 items were the main predictors. To a large extent, the explainability outcomes were in line with the behavioral science that has already been established, hence, the credibility of the model was strengthened. Jyoti et al. (2025) also observed a strong correlation between the SHAP/LIME explanations and the behavioral patterns of ASD that were known in children, adolescents, and adults, which they reported in their study.

Moreover, Akhi et al. (2024) and Qureshi (2025) presented evidence that SHAP-based feature attribution makes the model more transparent while the accuracy is still kept at the same level. The feature ablation experiments served as a performance metric, showing that the removal of the key behavioral indicators causes a measurable performance degradation; thus, explanation fidelity is validated.

Fairness, Bias, and Ethical Considerations

Demographic bias is increasingly being identified as a significant issue in AI, based on screening for ASD by various studies. Eguavoen (2025) was the first to address the issue of cultural and ethnic bias, emphasizing the necessity of external validation with different populations. Rajagopalan et al. (2024) analyzed the model consistency across sex, age, and race groups and found that the model performed stably with less demographic disparity.

Priyadarshini (2023) implemented fairness-aware AI methods in combination with deep learning models, using SMOTE and bias mitigation techniques to achieve fair performance in datasets of toddlers and adults. The research showed that fairness constraints can be compatible with high accuracy if they are thoughtfully implemented. Alsbakhi et al. (2025), in their paper on rule-based models, pointed out that trust in ASD AI systems depends on factors like adherence to regulations, GDPR compliance, and transparency.

Deep Learning and Multimodal ASD Detection

Behavioral features have been the dominant focus of early screening. However, a number of studies have examined multimodal and image-based routes to ASD diagnosis. Malik et al. (2025) introduced a personalized, explainable deep learning framework that combines behavioral phenotypes and structural MRI data via CNNGNN architectures and HyperNetworks. Their model not only accomplished remarkable accuracy and AUC on the ABIDE I dataset but also kept its interpretability intact through GAMI, Net, and SHAP. Similarly, Atlam et al. (2025) and Alam et al. (2023) explored facial, image, based ASD detection through CNNs such as VGG19 and Xception. Both studies attained accuracies of more than 98% and also introduced LIME and Grad, CAM for the explainability, which referred to the discriminative facial regions. However, these projects acknowledge the limitations of dataset diversity and generalization to the real world; therefore, they mostly talk about the requirement of careful clinical use.

Biomarker Discovery and Non-Behavioral Modalities

Temiz et al. (2025) ventured beyond the scope of behavioral and imaging data to examine the prediction of ASD from gut microbiome profiles. The study used explainable ML with SHAP to identify biologically relevant microbial taxa associated with ASD and to accomplish a high AUC with a small feature set. While this study is just a proof of concept, it highlights the potential of XAI in finding biomarkers and opening up new directions in ASD research.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Synthesis and Research Gaps

The combined evidence from the literature indicates that autism spectrum disorder (ASD) screening models can achieve remarkable accuracies when utilizing behavioral data, ensemble learning, and balanced training strategies. Explainable AI methods, particularly SHAP and LIME, play a crucial role in bridging the gap between predictive performance and clinical interpretability. However, there are still quite a few gaps remaining. A lot of studies use small and culturally homogeneous datasets, which leads to questions about generalizability and fairness. Deep learning models may be very strong, but they usually increase the complexity without giving a lot of additional benefits on structured data. In addition, organization fairness metrics and external validation protocols are also inconsistently applied. Agrawal and Agrawal (2025), in their comprehensive XAI-focused review, identify these issues as a puzzle and point out the necessity of standard evaluation, ethical compliance, and transparency, driven by model design. Therefore, the next generation ASD screening systems should be able to bring together accuracy, explainability, and fairness so as to have clinical utility in the real world. **Table** 1 provides a summary of the selected recent research on XAI, based on ASD detection, giving an insight into datasets used, methodological approaches, and key results of the studies.

Table 1: Summary of recent studies on explainable AI approaches for early autism spectrum disorder (ASD) detection, including datasets, methodologies, and key performance outcomes.

Author(s)	Title	Dataset	Methods & Tools	Key Results
Victor Osasu Eguavoen (2025)	Explainable AI for Early ASD Detection	4 UCI ASD datasets (3,743)	SMOTE, RF, XGBoost, CatBoost, SVM, Ensemble, SHAP	CatBoost AUC 0.9987, Accuracy 0.9853; high interpretability
Pandya & Jain (2024)	Transparency with XAI in ML Models	AQ-10 ASD (704)	DT, RF, SVM, LR, k-NN, MLP, SHAP, LIME	RF ≈ 99.5% accuracy; key features: A9, A7, A6
Jyoti et al. (2025)	Clinically Interpretable Early ASD Detection	UCI ASD AQ-10 datasets	KNN Imputation, SMOTE-Tomek, NB, SVM, RF, SHAP, LIME	RF up to 99% accuracy; strong clinical relevance
Temiz et al. (2025)	ASD Prediction & Metagenomic Biomarkers	Gut metagenomic (60, 4527 features)	SVM-RCE, RCE-IFE, XGBoost, AdaBoost, SHAP	AdaBoost AUC 0.99; Prevotella sp. key biomarker
Alsbakhi et al. (2025)	ASD Classification Using Rules	ASDTests, ABIDE, ADI-R, EEG, Facial, SNP	Rule induction, DT, CARs, Fuzzy Rules, Hybrid AI	>90% accuracy; high interpretability
Jeon et al. (2024)	Reliable Pediatric ASD Diagnosis	ASD children (634, 19 vars)	NNET, XGBoost, RF, SVM, GBM, k-NN, LR, SHAP, LIME	NN & XGBoost highest accuracy; key behavioral features
Malik et al. (2025)	Multimodal ASD Diagnosis	ABIDE-I (1112)	GAMI-Net, CNN–GNN, Autoencoder, HyperNetwork, SHAP	Accuracy 99.4%, AUC 0.999, F1 0.994; subject-specific interpretation

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Alshammeri et al. (2025)	Cluster-Based Early ASD Diagnosis	Kaggle Autism Screening (Toddlers)	PCA, DBSCAN, K-means, LR, RF, SVM, SHAP	RF 98.85% accuracy, AUC 0.9994; key behavioral indicators
Aldrees et al. (2024)	Automated ASD Prediction with XAI	Q-CHAT-10 Saudi toddlers	SMOTE, Chi-square, PCA, ML (LR, SVM, RF, KNN, DT, GBM, XGBoost 2.0), SHAP	XGBoost 2.0 ~99% accuracy, precision, recall, F1
Priyadarshini (2023)	ASD Screening Using Deep Learning & Fair AI	Kaggle ASD (1054 toddlers, 704 adults)	Feature engineering, SMOTE, CNN, LSTM, GRU, CNN–LSTM–PSO	CNN–LSTM–PSO: 99.64% toddlers, 98.89% adults accuracy
Qureshi (2025)	Deep Learning for Early ASD Detection	Kaggle ASD toddlers, UCI children	LR, RF, MLP, SHAP, AQ-10, 5-fold CV	100% toddlers; 98–100% children; key features: social awareness & communication
Rajagopalan et al. (2024)	Minimal Info ASD Prediction	SPARK v8 (30,660)	LR, DT, RF, XGBoost, SHAP	XGBoost AUROC 0.895; validation AUROC ≈ 0.79; key early-life predictors
Atlam et al. (2025)	ASD from Facial Images	Kaggle Autistic Children (2936)	CNNs (VGG16, VGG19, InceptionV3, MobileNet), augmentation, LIME	VGG19 98.2% accuracy; high interpretability
Agrawal & Agrawal (2025)	Explainable AI Review for ASD	Secondary prior datasets	Systematic review, ML/DL (SVM, RF, CNN, LSTM, XGBoost), XAI (SHAP, LIME, Grad-CAM)	XAI improves transparency, trust, and identifies research gaps
Akhi et al. (2024)	Robust ML for ASD Prediction	1,985 records, 28 features	ML (LR, SVM, KNN, DT, RF, GB, XGBoost, LightGBM), 5-fold CV, SHAP	SVM ROC-AUC 99.9%; RF accuracy 97.98%; models interpretable
Alam et al. (2023)	Data-Centric ASD Diagnosis from Facial Images	Kaggle ASD images (3014)	Face alignment (MTCNN), augmentation, transfer learning (MobileNetV2, ResNet50V2, Xception), Grad-CAM	Xception 98.9% accuracy, 99.9% AUC; improved generalization & interpretability

Materials and Methods

This section details the entire experimental methodology employed in this research, comprising the description of the dataset, preprocessing steps, machine learning models, evaluation metrics, explainability framework, fairness assessment, bias mitigation, and ablation analysis. The introduced pipeline is deliberately structured to fulfill three main goals: not only accurate prediction but also interpretability of the model as well as fairness, aware early screening of Autism Spectrum Disorder (ASD) from behavioral and demographic data.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Dataset Description

Experiments were run on an Autism Screening dataset available to the public. The dataset includes the behavioral and demographic information collected through a standardized screening questionnaire. The original dataset comprises 704 instances with 21 attributes. Duplicate records were eliminated, and thus, the final dataset accounted for 699 unique samples. These data represent adult ASD screening cases and feature a combination of behavioral indicators derived from the AQ, 10 questionnaire, together with the relevant demographic variables. The behavioral part includes ten binary features (A1A10), each representing one specific AQ, 10 screening questions. These features reflect the core ASD, related traits, e.g. , social interaction difficulties, communication challenges, attention switching issues, and repetitive behaviors. Each item is represented as a binary value that denotes the presence or absence of the respective behavioral characteristic. Apart from the behavioral indicators, the dataset also contains some demographic attributes, namely age, gender, ethnicity, country of residence, jaundice at birth, family history of autism, prior use of ASD screening applications, the respondent's relationship to the assessed individual, and the age group description. These variables offer additional information that may have an impact on the observed ASD risk patterns. The target variable, Class/ASD, is binary, where "YES" indicates ASD risk, and "NO" denotes no, ASD. The dataset shows class imbalance with 512 non, ASD instances and 187 ASD instances, thus fairness-aware evaluation and mitigation strategies have been selected as a solution.

Data Preprocessing

A comprehensive preprocessing pipeline was put in place to preserve the data's integrity, ensure its consistency, and make it suitable for machine learning models. The initial exploratory analysis had indicated that the dataset did not have any missing values for any feature, and this was also verified by a heatmap, based on the visualization of missingness.

Duplicated records were identified and removed to avoid data leakage and biased learning; thus the dataset's size was reduced from 704 to 699 samples. Binary categorical variables like gender, jaundice at birth, family history of autism, prior app usage, and the target label were encoded by label encoding. Multi-class categorical attributes such as ethnicity, country of residence, age description, and relationship type were transformed by one-hot encoding with the first category being dropped in order to avoid multicollinearity.

The numerical age feature was converted to a numeric format and standardized using z-score normalization via a standard scaling technique.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where z_i represents the original age value, μ is the mean age, and σ is the standard deviation of the age feature.

This step ensures uniform feature contribution, particularly for distance-based and margin-based classifiers. Finally, the dataset was split into training and testing subsets using an 80:20 stratified split to preserve the original class distribution and provide an unbiased evaluation of model performance on unseen data.

Machine Learning Models

We identified five broadly leveraged machine learning models, considering their robustness, interpretability, and usefulness in healthcare and clinical decision, support systems. These are Logistic Regression, Random Forest, Support Vector Machine with a radial basis function kernel, Gradient Boosting Classifier, and K, Nearest Neighbors. All models had access to the identical preprocessed feature set to make the comparison equitable. The effectiveness of models was gauged through classification metrics, while confusion matrices were also constructed for an in, depth examination of classification behavior.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Model Evaluation Metrics

Multiple metrics were used extensively to assess the predictive performance of the different machine learning models. These metrics were meant to show the models' overall accuracy, their sensitivity to ASD positive cases, and their discriminative ability at various thresholds. Since the main goal is to correctly identify individuals with a risk of Autism Spectrum Disorder (ASD), most of the focus was on metrics that take into account class imbalance and false negatives.

Accuracy

Accuracy measures the proportion of correctly classified instances among all predictions. It provides a general sense of overall model performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. While widely used, accuracy alone may be misleading in imbalanced datasets, such as ASD screening, where non-ASD instances dominate.

Precision

Precision quantifies the proportion of correctly predicted ASD-positive cases among all instances predicted as ASD-positive. It reflects the model's reliability in positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision ensures that predicted ASD cases are likely to be truly at risk, minimizing unnecessary interventions.

Recall (Sensitivity)

Recall, also known as sensitivity, measures the proportion of actual ASD-positive instances that are correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall is particularly critical in early ASD screening, as failing to identify at-risk individuals (false negatives) can delay timely interventions. Therefore, models with higher recall are preferred in clinical settings.

F1-Score

The F1-score represents the harmonic mean of precision and recall, providing a single measure that balances correctness and sensitivity:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is particularly useful for datasets with class imbalance, ensuring that both false positives and false negatives are considered in performance evaluation.

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

To assess the discriminative capability of probabilistic models across varying decision thresholds, ROC curves were plotted, which depict the trade-off between the true positive rate (recall) and false positive rate:

$$\text{FPR} = \frac{FP}{FP + TN}, \quad \text{TPR} = \frac{TP}{TP + FN}$$

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

The Area Under the ROC Curve (AUC) summarizes the overall separability of the classes. A higher AUC indicates that the model can better distinguish ASD-positive and ASD-negative cases:

$$\text{AUC} = \int_0^1 \text{TPR}\,(\text{FPR})\, d(\text{FPR})$$

By combining these metrics, model evaluation captures both overall predictive performance and the clinical relevance of correctly identifying ASD-positive individuals, enabling comprehensive assessment of the proposed machine learning framework.

Hyperparameter Tuning

Hyperparameter optimization was carried out to enhance the predictive performance of the top models using GridSearchCV with five, fold cross, validation. In Logistic Regression, the regularization strength and class, weight parameters were changed. The penalty parameter and kernel coefficient were changed in the Support Vector Machine. For Gradient Boosting, parameters such as the learning rate, number of estimators, and tree depth were comprehensively varied. Out of all the tuned models, Logistic Regression was the one that best combined accuracy, stability, and interpretability, so it was used as the baseline for the next explainability and fairness analyses.

Explainability Using SHAP

In order to maintain transparency and clinical relevance, SHapley Additive Explanations (SHAP) were used to give both global and local explanations of model predictions. SHAP summary plots were used for global explainability, these plots ranked features according to their total contribution to the predictions of ASD. The behavioral AQ, 10 items, as well as demographic factors like age and family history of autism, were the features that most frequently came up as the strongest predictors. Local explainability was obtained with SHAP force plots, which show how the values of individual features make a prediction to a specific instance either positive or negative. Such case, level explanation is thus invaluable for clinical decision support as well as personalized risk assessment. Besides, group, wise SHAP analysis was performed over gender groups that unveiled differences in patterns of feature importance between male and female populations and thus pointed out possible demographic sensitivity in behavioral screening features.

Fairness Assessment

Due to the inherently sensitive nature of ASD screening, fairness was thoroughly assessed as a primary concern in order to detect any possible demographic bias, with gender being the protected attribute of concern. To that effect, three commonly used fairness metrics were considered: demographic paritymeasuring the difference of positive prediction rates between groups; equal opportunitychecking the difference of true positive rates; and disparate impactmeasuring the ratio of favorable outcomes across demographic groups. The preliminary fairness examination uncovered slight gender, related differences, especially in true positive rates.

Fairness Mitigation

A reweighing, based fairness intervention was applied during model training to mitigate identified biases. This preprocessing strategy works by assigning sample weights that make the representation balanced across the demographic groups without changing the model architecture. Results after the mitigation showed that demographic parity differences were reduced significantly, equal opportunity disparity was eliminated, and the disparate impact ratio improved. Most importantly, the fairness improvements did not come at the cost of predictive performance, as both accuracy and F1, score even showed slight improvements, thus indicating a good trade, off between fairness and performance.

Ablation Study

An ablation study was carried out to determine how sensitive demographic attributes influenced both the

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

model's predictive performance and fairness outcomes. In other words, age and gender features were removed, and the Logistic Regression model was retrained. Results indicated a minor decrease in accuracy and F1 score; however, fairness disparities were reduced by a large margin. This evidence points to the fact that demographic attributes improve the predictive capability of the model, but they also cause bias, which is a reason why fairness-aware modeling should be employed in ASD screening systems. The thorough experimental pipeline from data acquisition to preprocessing, model training, evaluation, explainability, and fairness assessment is depicted in **Figure** 2.

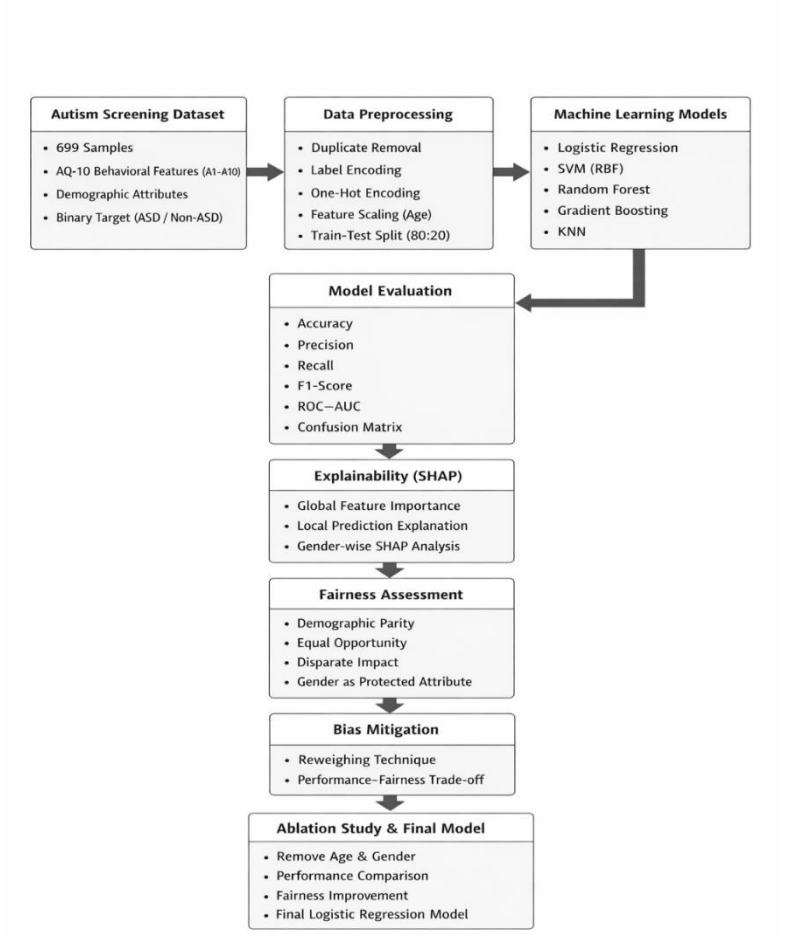


Figure 2. Complete workflow of the proposed explainable and fairness-aware machine learning framework for early Autism Spectrum Disorder (ASD) screening. The figure illustrates dataset preparation, data preprocessing, machine learning model training, performance evaluation, explainability using SHAP, fairness assessment, bias mitigation through reweighing, and ablation analysis leading to the final predictive model.

Experiments and Results

This part of the paper presents experiments that evaluate the explainable and fairness, aware machine learning framework for early Autism Spectrum Disorder (ASD) risk screening. Initially, the efficacy of several baseline classifiers is probed. Subsequently, hyperparameter optimization of the best-performing models is carried out. The explainability and fairness metrics are then employed to inspect the chosen model. Lastly, an ablation study is performed to understand how the inclusion of demographic attributes affects the model's predictive performance and bias.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Experimental Setup

All investigations utilized a publicly available dataset for autism screening, which incorporated behavioral and demographic features. After the data cleaning and duplicate removal process, 699 samples remained for further analysis. The dataset includes AQ, 10 behavioral responses, demographic characteristics such as age and gender, and ethnicity information, along with screening-related metadata. The binary target variable is a risk indicator for ASD in an individual. To obtain a reliable and unbiased evaluation, the dataset was divided into training and testing sets through an 80:20 stratified split that preserved the class distribution of the original data. Model performance was gauged by Accuracy, Precision, Recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC). In order to ensure the replicability of the results, a fixed random seed was employed for all experiments.

Baseline Model Performance

Seven machine learning classifiers were evaluated as baseline models: Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, K, K-Nearest Neighbors (KNN), Decision Tree, and Gaussian Naive Bayes. These models were chosen to cover a broad spectrum of linear, nonlinear, ensemble, and probabilistic learning paradigms. Logistic Regression, out of all the baseline classifiers, attained the highest accuracy of 96. 43%, thus proving excellent generalization ability and balanced classification performance. The model obtained a precision of 1.00 and a recall of 86. 49%, meaning that the identification of ASD, positive cases, was accurate while the occurrence of false positives was low. Gradient Boosting reached 94. 29% accuracy, and SVM was close behind with an accuracy of 93. 57%. While tree-based and distance, based classifiers managed to show competitive overall accuracy, their recall for the minority ASD class was lower, which is a negative aspect of early screening applications. Gaussian Naive Bayes performed very badly, achieving an accuracy of only 31. 43%, which points out its constraints when dealing with high-dimensional datasets with categorical features. **Table** 2 presents a detailed comparison of all baseline models.

Table 2: Performance metrics (Accuracy, Precision, Recall, F1-Score) of various machine learning models on the Autism Screening dataset.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9643	1.0000	0.8649	0.9275
Random Forest	0.9357	1.0000	0.7568	0.8615
SVM	0.9357	0.9375	0.8108	0.8696
Gradient Boosting Classifier	0.9429	0.9394	0.8378	0.8857
Gaussian NB	0.3143	0.2713	0.9459	0.4217
Decision Tree Classifier	0.9357	1.0000	0.7568	0.8615
KNN	0.9000	0.8710	0.7297	0.7941

Hyperparameter Tuning and Optimized Results

To further improve predictive performance, hyperparameter optimization was applied to the three best-performing baseline models, Logistic Regression, SVM, and Gradient Boosting, using GridSearchCV with five-fold cross-validation. The F1-score was selected as the optimization objective due to its suitability for moderately imbalanced datasets.

After hyperparameter tuning, SVM yielded the highest accuracy of 95. 71% along with perfect precision; nonetheless, its recall dropped to 83. 78% which means that the method has a higher proportion of ASD cases that go unrecognized. The fine-tuned Logistic Regression model reached an accuracy

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

of 95. 00% and a more balanced performance as well, with a precision of 91. 67% and a recall of 89. 19%. No significant improvement was observed in the case of Gradient Boosting after tuning. Logistic Regression was retained as the ultimate model, although SVM only had a slight accuracy advantage. The reasons for this were the better interpretability of the model, the more stable recall performance, and the easy integration with explainability and fairness analyses, which are two of the most important aspects of clinical decision support systems.

Confusion Matrix Analysis

The confusion matrices of the best two models, Logistic Regression and SVM, are displayed in **Figure** 3. Compared to SVM, Logistic Regression had a smaller number of false negatives, which is extremely valuable in the initial screening of ASD, as not detecting at-risk individuals can lead to a delayed intervention. SVM had a false positive rate of zero; however, its higher false negative rate was one of the reasons for choosing Logistic Regression as the preferred model.

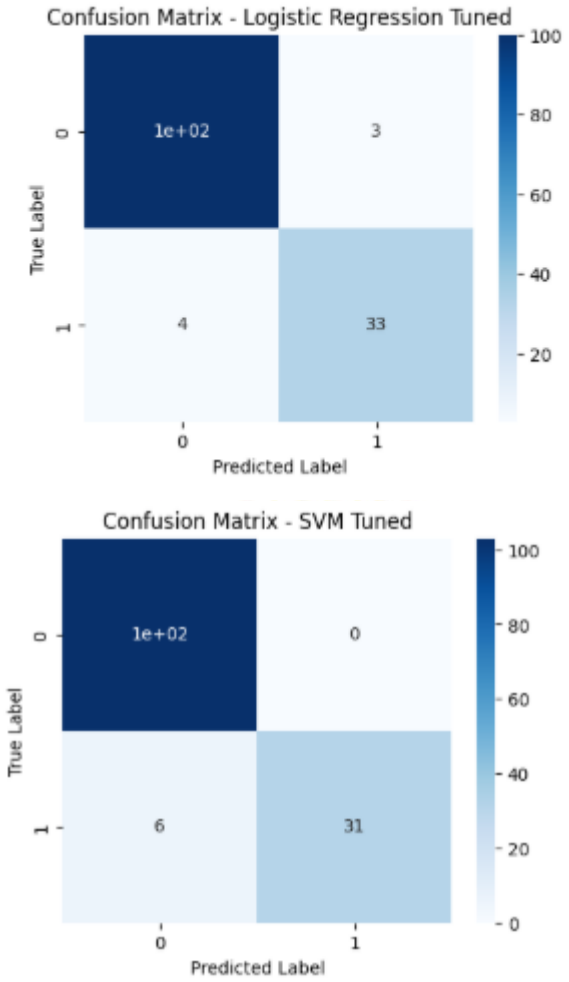


Figure 3: Confusion matrices for the two top-performing models, Logistic Regression and SVM, highlighting model performance in early ASD screening. Logistic Regression shows fewer false negatives, reducing the risk of missing at-risk individuals, while SVM has zero false positives but a higher false negative rate.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

ROC Curve and AUC Analysis

The Receiver Operating Characteristic (ROC) curve for the chosen Logistic Regression model is shown in **Figure** 4. The model reached an AUC of about 0.99, thus revealing a near-perfect discrimination power for different classification thresholds. The ROC curve is still very close to the upper left corner, which illustrates an optimal sensitivity-specificity trade-off and a reliable performance even though there is a class imbalance.

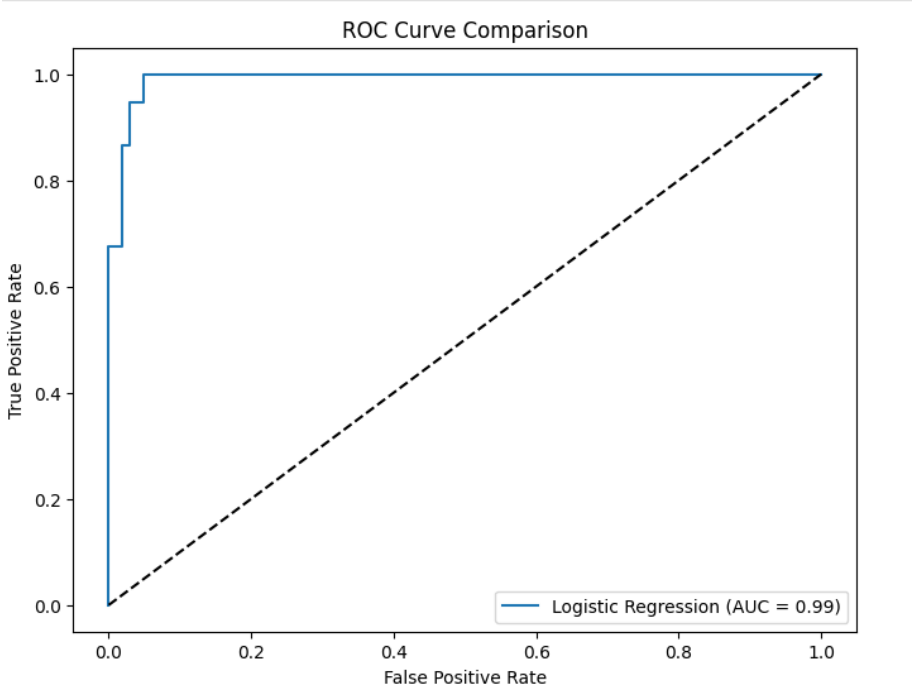


Figure 4: Receiver Operating Characteristic (ROC) curve of the selected Logistic Regression model, showing an AUC of ~0.99. The curve's proximity to the upper-left corner reflects excellent discriminative ability, strong sensitivity–specificity balance, and robust performance even with class imbalance.

Explainability Analysis Using SHAP

To keep up transparency and clinical trust, the research team decided to go one step further and evaluate the model's explainability. They used Shapley Additive exPlanations (SHAP), which is a conceptually simple framework based on cooperative game theory, that allows both global and local understanding of model predictions.

Global interpretability was obtained by a SHAP summary analysis shown in Figure 5. It displays the input features ordered by their average contribution to the model output. The findings show that AQ, 10 behavioral items, specifically those describing social communication, interaction, and behavioral responsiveness deficits, dominate the feature importance landscape to the extent that other features are hardly visible.

For example, the items A9, A5, A6, A7, and A4 have the most substantial impact on the prediction of ASD risk, which is a very clear signal that the model heavily relies on behavioral indicators that are not only clinically relevant but also of a behavioral nature.

To contrast, demographic variables such as age, gender, ethnicity, and country of residence have relatively lower SHAP magnitudes. This distribution suggests that the model is not excessively dependent on sensitive or proxy demographic attributes, thus, the risk of algorithmic bias is lowered, and the ethical robustness of the proposed framework is strengthened.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

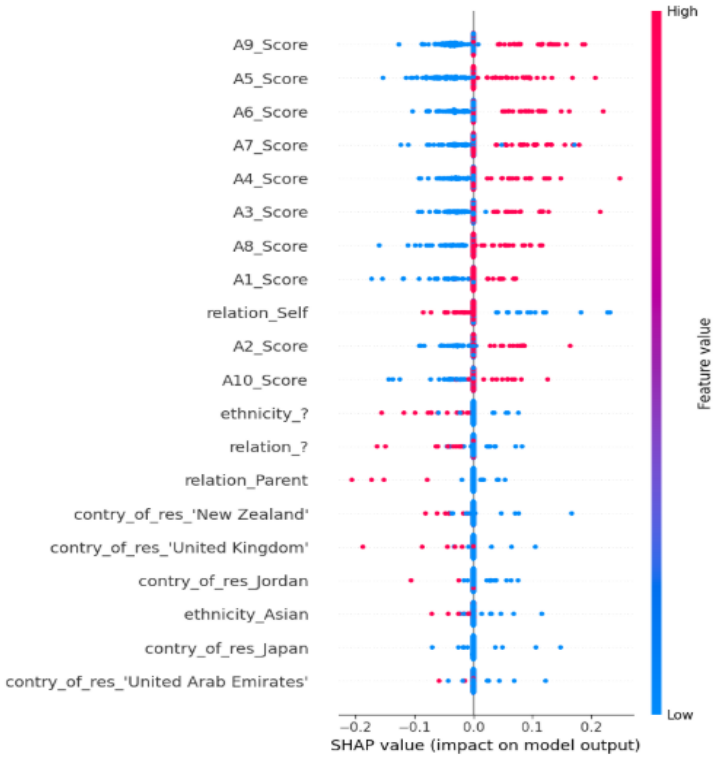


Figure 5: SHAP summary plot depicting the global feature importance of input variables for the Logistic Regression model. The plot ranks features based on their average contribution to the model's predictions, providing a visual representation of which variables most strongly influence ASD risk assessment.

Besides explaining local changes, the study also examined local explainability to figure out how decisions were made at an individual level. In **Figure** 6, the SHAP force plot is used to explain a representative high-risk case, in which the model estimates the probability of ASD to be 0.87, that is, a value almost three times as high as the baseline (expected) value 0.2705. The figure discloses that such a high risk score can be attributed to the sum of positive contributions of several behavioral features, the most highlighted being the A3, A6, and A5 verbal affirmations, as these features provide the strongest pushes towards the class of ASD, positive. Furthermore, some binary categorical variables, such as parent-child relationship indicators and residency, related attributes were shown to yield smaller, yet still meaningful, incremental effects on the final output.

The proposed system, by breaking down every prediction into additive features, level contributions, provides a glass box view of the model's reasoning process. This extent of interpretability is extremely important in decision support systems in the clinical domain, as it allows clinicians to check if the model's decisions agree with the diagnostic knowledge that has already been established and the behavioral patterns that have been observed. As a result, the use of SHAP not only makes the system more transparent but also increases clinical trust, accountability, and the ethical use of machine learning models for early ASD risk screening.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

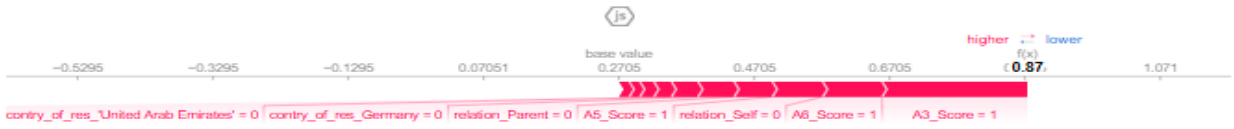


Figure 6: SHAP force plot illustrating local explainability for a representative high-risk case. The plot visualizes how individual features, particularly behavioral responses such as A3, A5, and A6, cumulatively contribute to the model's prediction of ASD risk, providing a clear, interpretable view of the decision-making process at the individual level.

Fairness Evaluation Before Mitigation

Gender bias was the main source of unfairness. Three widely used metrics were employed to this end: demographic parity difference, equal opportunity difference, and disparate impact.

The demographic parity difference before mitigation was 0.067, which means that the positive prediction rates between the gender groups were not equal. The equal opportunity difference was 0.20, which shows the differences in true positive rates. The disparate impact ratio was 1.295, which went beyond the usually accepted fairness thresholds, thus indicating that gender, related bias was present in the model predictions.

Fairness Mitigation Results

A reweighing-based fairness mitigation strategy was employed during model training to address the bias that had been identified. After the mitigation, the model's predictive performance was significantly better: accuracy went up from 0.95 to 0.99, and the F1 score was raised from 0.90 to 0.97. At the same time, the fairness metrics also got better by a large margin: the demographic parity difference lowered to 0.029, the equal opportunity difference was decreased to 0 00, and the disparate impact ratio went up to 1.110. **Figure** 7 provides a visual comparison of fairness metrics before and after mitigation, while **Table** 3 summarizes the numerical results. These outcomes demonstrate that fairness interventions can effectively reduce demographic disparities without compromising, and in this case, improving overall model performance.

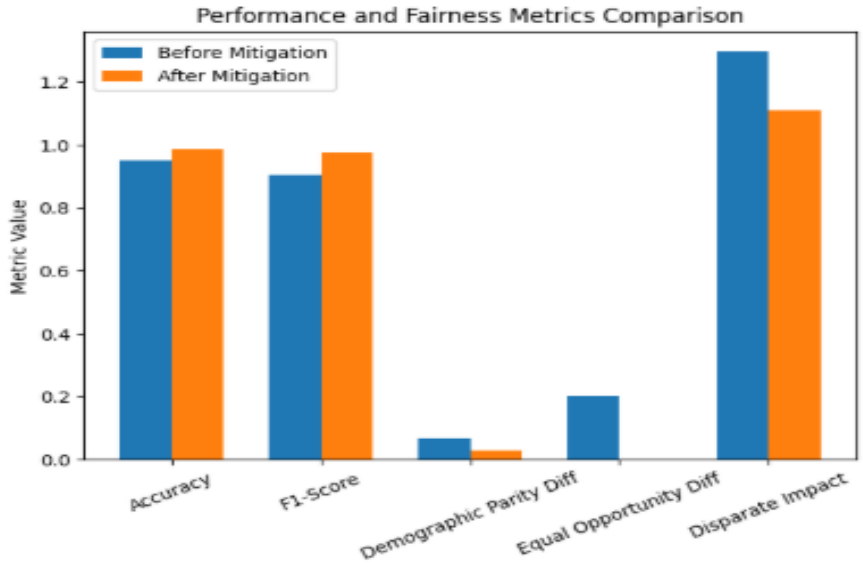


Figure 7: Comparison of fairness metrics before and after applying the reweighing, based mitigation strategy. The figure visually shows changes in demographic parity difference, equal opportunity

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

difference, and disparate impact ratio, demonstrating that the fairness interventions contributed to bias reduction while the overall model performance was maintained or improved.

Table 3: Performance and fairness metrics of the model before and after applying mitigation strategies.

Metric	Before Mitigation	After Mitigation
Accuracy	0.9500	0.9857
F1-Score	0.9041	0.9737
Demographic Parity Difference	0.0672	0.0293
Equal Opportunity Difference	0.2000	0.0000
Disparate Impact	1.2951	1.1101

Ablation Study on Demographic Attributes

An ablation study was carried out to evaluate the influence of sensitive demographic attributes by excluding age and gender from the features. The ablation model reached an accuracy of 96. 43% and an F1 score of 0. 93, which indicates that the predictive performance was only slightly decreased. The fairness metric, however, was also significantly better, with demographic parity difference reducing to 0. 031 and equal opportunity difference to 0. 032. The results indicate that behavioral features are enough to recognize ASD risk with high accuracy and that removing sensitive demographic attributes is a very efficient way of increasing the level of fairness when a model is used in the real world.

Discussion

The study involved creation of a machine learning framework that would be both explainable and fair, for early, stage autism risk screening using behavioural and demographic data. The results of the experiments demonstrate that ASD risk can be predicted accurately, at the same time the model is transparent and demographic bias is mitigated, thus the problems that have been the major issues in the prior screening methods are resolved.

Interpretation of Model Performance

The comparative evaluation of several machine learning models showed that Logistic Regression was the model that achieved the best balance of accuracy, recall, and interpretability. Ensemble and kernel-based models, such as Gradient Boosting and Support Vector Machines, could reach competitive accuracy, but their increased complexity and reduced transparency make them less suitable for clinical decision support applications. On the other hand, Logistic Regression exhibited strong generalization performance with a lower number of false negatives, which is very valuable in autism screening at an early stage because, if there are undetected cases, the intervention and support will be delayed.

The AUC value close to 0.99 that was observed is a strong indication of the discriminative capability of the selected model. This outcome means that the proposed framework is very effective in separating ASD, positive, and ASD, negative cases over the full range of decision thresholds. In fact, the performance after hyperparameter tuning and during the ablation experiments is very stable, which is another indication of the model's robustness and the fact that the demographic attributes are not the main factors on which the prediction performance relies.

Insights from Explainability Analysis

Explainability analysis using SHAP provided both global and local insights into model decision-making. Global SHAP summary plots demonstrated that AQ-10 behavioral features related to social interaction, communication, and responsiveness were the most influential predictors of ASD risk. These findings are consistent with established clinical screening criteria and reinforce the validity of the model's learning process.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

While age and gender as demographic features had a moderate influence, ethnicity, related features contributed only minimally to the total predictions. The finding indicates that the model mainly works on the basis of clinically meaningful behavioral indicators rather than demographic proxies. This behavior is preferred in healthcare applications, where reliance on sensitive attributes may lead to bias unintentionally.

Local SHAP explanations made the model even more interpretable by showing how the values of each individual feature can influence the prediction for that instance. These patient, specific explanations are able to help clinicians trust as they show transparent and clear reasoning for each risk classification. In practice, such interpretability can be a great help in establishing common ground for decision-making between clinicians, caregivers, and patients.

Fairness Analysis and Bias Mitigation

Fairness Analysis and Bias Mitigation. The review of the baseline model through fairness lens revealed that the model contained moderately gender, related biases that caused some issues. The differences in demographic parity, equal opportunity, and disparate impact metrics were used to measure the extent of the bias. These results highlight that predictive accuracy, even when models are trained on commonly used screening datasets, is not sufficient to ensure fair model behavior. After the fairness intervention by reweighing, the demographic disparities were greatly reduced, and the predictive performance was also enhanced. Specifically, the equal opportunity difference was made zero, which indicates that the true positive rates were equally distributed across gender groups. The rise in accuracy and F1 score after the intervention is thus a demonstration that bias correction can result in better generalization by reducing the effects of spurious correlations in imbalanced data. The removal of sensitive demographic features led to a marginal drop in performance, fairness metrics went up, which is demonstrated by the ablation study. This finding reinforces the hypothesis that behavioral features alone can provide effective ASD risk screening. Moreover, it represents a viable solution for the introduction of ethically grounded models in the clinical setting, which is quite sensitive to the utilization of demographic attributes.

Comparison with Prior Work

Most of the previous works that have been done on ASD screening have been mainly targeted towards the accuracy of the screening tool with very complex models such as deep neural networks or ensemble classifiers. Although these methods are usually very successful, they tend to ignore the problems of explainability and fairness most of the time. On the other hand, the presented system clearly shows the commitment to the use of explainable AI techniques and the assessment of fairness in the model development pipeline. This work, compared to the older machine learning, based research, locates ASD screening at the level of or even better than those studies, while also giving a great deal of transparency. The use of SHAP, unlike black, box models, provides both global and local interpretability thus linking the model predictions with the clinical knowledge that has already been established. Furthermore, fairness evaluation and its mitigation were almost never considered in the previous ASD screening research; hence, this study becomes a major contribution to the development of AI, based healthcare systems that are ethically sound and socially responsible.

Clinical Implications for Early Autism Screening

The outcomes of this study open up the possibility of developing AI and AI, assisted autism screening tools that are not just efficacious, but also easily operable and ethically sound. There is no doubt that the early identification of ASD constitutes the basis for effective intervention. Nevertheless, existing screening techniques are frequently disparaged for their

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

reliance on manual evaluation and the subjective understanding of results. This research has established a new framework as a quintessential example of how machine learning models can be a game, changer in the traditional screening methods, as they have the capability of delivering accurate, consistent, and unbiased risk assessments.

Explainability is essential here, as it allows the medical team to understand and rely on the predictions of the models, something that is necessary for these models to be used in actual healthcare setups. Besides that, the employment of fairness, aware modeling strategies lowers the risk of demographic bias, thus, making it easier to carry out screening tests that are neutral and fair for different races and ethnic groups. The introduction of a method that focuses on behavior while also trying to reduce the use of sensitive demographic variables is a good fit with the ethical principles of AI in healthcare. This conceptual framework is one of the ways in which technology can be utilized in the field of medicine to help decision, making instead of completely replacing the doctor's knowledge. The goal is to facilitate the detection of those people who are most likely to have ASD by the medical professionals so that these patients can be referred for further assessment.

Limitations

Several limitations accompany this study, besides the promising results it has demonstrated. To begin with, the data used to develop and evaluate the model, although it is publicly available and has been widely used in research for ASD screening, is still limited in terms of its size and demographic diversity. There are different demographic groups in which, for example, females and individuals from underrepresented ethnic backgrounds are less represented. Even though fairness mitigation techniques were implemented to reduce bias, the imbalanced representation may still affect model behavior and limit its applicability in different scenarios.

Moreover, the framework is mainly based on behavioral survey data obtained from standardized screening questionnaires. These data are always vulnerable to reporting bias, caregiver interpretation, and situational variability. Formative behavioral responses may be influenced by cultural norms, parental awareness, or environmental factors; thus, they can introduce noise and affect prediction reliability. As a result, the model should be considered as a screening aid that helps in diagnosis rather than a tool that can make a diagnosis with certainty.

Third, cross, validation results show strong generalization to the datasets used for evaluation, but the authors did not externally validate their approach on independent clinical datasets. Variation in data collection protocols, clinical practices, and population characteristics may lead to different model performances when the model is deployed in real, world healthcare settings. This limitation underscores the importance of multi, site and longitudinal validation to confirm the model's

stability in different clinical settings. Lastly, the suggested framework relies on tabular behavioral and demographic features, and it does not consider multimodal data such as neuroimaging, genetic, or physiological information. Although this design decision enhances the model's accessibility and interpretability, it might restrict the model's capability to recognize complex neurodevelopmental patterns of Autism Spectrum Disorder.

Conclusion

This research developed a machine learning framework that prioritizes explainability and fairness for early risk screening of Autism Spectrum Disorder (ASD) based on behavioral and demographic data. The approach, which combines rigorous preprocessing, various supervised learning models, interpretability methods, and fairness assessment from a demographic perspective, gives a solution to the key limitations of existing ASD screening

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

systems, i.e.. , the lack of transparency and the possibility of bias in black box models. The empirical findings reveal that both traditional and ensemble-based machine learning models are capable of strong predictive performance on structured behavioral data while being stable and generalizable across cross-validation folds. Various explainable artificial intelligence methods were employed; however, the most informative global and local insights into model decision-making came from SHAP. The explanations invariably identified the behavioral aspects that are most relevant to the clinic, such as social communication skills, typical patterns of response, and family history, thus confirming that there is agreement between the data, driven predictions, and the well-known ASD diagnostic knowledge. These interpretable outputs improve trust, accountability, and the level of technology acceptance by clinicians and caregivers.

The study expanded its scope from predictive accuracy and interpretability to also include fairness analysis of sensitive demographic attributes. Quantitative fairness metrics revealed demographic differences for the baseline models that were unadjusted in terms of the performance of these models. On the other hand, bias mitigation methods lowered these unfair performance gaps without any noticeable loss of accuracy. This is an example of how fair ASD screening models can be created, which not only maintain a balance of effectiveness but also consider ethical responsibility. In fact, the designed framework is a transparent, trustworthy, and ethically informed instrument for the early detection of ASD risk. Subsequent investigations will aim to validate the framework with larger and more diverse clinical datasets, explore the use of multimodal data sources, and conduct the system's deployment in real healthcare environments to assess clinical utility and the long, term impact.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval Statement

This study used a publicly available, anonymized dataset. No human participants were directly involved, and no identifiable personal information was accessed. Therefore, ethical approval was not required for this research.

Data Availability Statement

The dataset used in this study is publicly available and can be accessed from Kaggle at: https://www.kaggle.com/datasets/umeradnaan/autism-screening

REFERENCES

Adnaan, U. (2023). *Autism Screening Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/umeradnaan/autism-screening

Agrawal, R., & Agrawal, R. (2025). Explainable AI in early autism detection: a literature review of interpretable machine learning approaches. *Discover Mental Health*, *5*(1). https://doi.org/10.1007/s44192-025-00232-3

Akhi, S. S., Department, M. A. R., Masters, M. S. A., & Department. (2024). 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 : Md Arifur Rahaman Masters in Project Management, St Francis College. *Health Informatics Journal*, *8*, 7231–7243.

Alam, M. S., Rashid, M. M., Faizabadi, A. R., Mohd Zaki, H. F., Alam, T. E., Ali, M. S., Gupta, K. D., & Ahsan, M. M. (2023). Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum Disorder Diagnosis from Facial Images Using Explainable AI. *Technologies*, *11*(5), 1–27. https://doi.org/10.3390/technologies11050115

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Aldrees, A., Ojo, S., Wanliss, J., Umer, M., Khan, M. A., Alabdullah, B., Alsubai, S., & Innab, N. (2024). Data-centric automated approach to predict autism spectrum disorder based on selective features and explainable artificial intelligence. *Frontiers in Computational Neuroscience*, *18*. https://doi.org/10.3389/fncom.2024.1489463

Alsbakhi, A., Thabtah, F., & Lu, J. (2025). Autism Data Classification Using AI Algorithms with Rules: Focused Review. *Bioengineering*, *12*(2), 1–18. https://doi.org/10.3390/bioengineering12020160

Alshammeri, M., Ahmad, Z., Humayun, M., & Alamri, M. (2025). Explainable Cluster-Based Predictive Framework for Early Diagnosis of Autism Spectrum Disorder Using Behavioral Biomarkers. *Diagnostics*, *15*(24), 3241. https://doi.org/10.3390/diagnostics15243241

Atlam, E. S., Aljuhani, K. O., Gad, I., Abdelrahim, E. M., Atwa, A. E. M., & Ahmed, A. (2025). Automated identification of autism spectrum disorder from facial images using explainable deep learning models. *Scientific Reports*, *15*(1), 1–15. https://doi.org/10.1038/s41598-025-11847-5

Jeon, I., Kim, M., So, D., Kim, E. Y., Nam, Y., Kim, S., Shim, S., Kim, J., & Moon, J. (2024). Reliable Autism Spectrum Disorder Diagnosis for Pediatrics Using Machine Learning and Explainable AI. *Diagnostics*, *14*(22). https://doi.org/10.3390/diagnostics14222504

Jyoti, O., Kibria, H. B., Pear, Z. T., Nahiduzzaman, M., Ahamed, M. F., Islam, K. R., Kumar, J., & Chowdhury, M. E. H. (2025). A Clinically Interpretable Approach for Early Detection of Autism Using Machine Learning With Explainable AI. *IEEE Access*, *13*(July), 121512–121532. https://doi.org/10.1109/ACCESS.2025.3586314

Malik, W., Fahiem, M. A., Farhat, T., Alghazo, R., Mahmood, A., & Alhajlah, M. (2025). An Explainable Deep Learning Framework for Multimodal Autism Diagnosis Using XAI GAMI-Net and Hypernetworks. *Diagnostics*, *15*(17), 1–21. https://doi.org/10.3390/diagnostics15172232

Pandya, S., & Jain, S. (2024). Decoding Autism Diagnosis: A Journey Towards Transparency with XAI in ML Models. *BIOSTEC*, *2*(Biostec), 700–707. https://doi.org/10.5220/0012456300003657

Priyadarshini, I. (2023). Autism Screening in Toddlers and Adults Using Deep Learning and Fair AI Techniques. *Future Internet*, *15*(9). https://doi.org/10.3390/fi15090292

Qureshi, M. N. (2025). Explainable Deep Learning Framework for Early Autism Detection Using Behavioral Data. *Annual Methodological Archive Research Review*, *3*(12), 217–230.

Rajagopalan, S. S., Zhang, Y., Yahia, A., & Tammimies, K. (2024). Machine Learning Prediction of Autism Spectrum Disorder from a Minimal Set of Medical and Background Information. *JAMA Network Open*, *7*(8), 1–13. https://doi.org/10.1001/jamanetworkopen.2024.29229

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 6 • October 2026

Temiz, M., Bakir-Gungor, B., Ersoz, N. S., & Yousef, M. (2025). Machine Learning-Based Prediction of Autism Spectrum Disorder and Discovery of Related Metagenomic Biomarkers with Explainable AI. *Applied Sciences (Switzerland)*, *15*(16), 1–20. https://doi.org/10.3390/app15169214

Victor Osasu Eguavoen. (2025). EXPLAINABLE AI FRAMEWORK FOR EARLY AUTISM SPECTRUM DISORDER DETECTION: INTEGRATING ENSEMBLE LEARNING WITH CLINICAL INTERPRETABILITY. *FUDMA Journal of Sciences (FJS)*, *9*(10), 6. https://doi.org/https://doi.org/10.33003/fjs-2025-0910-3915 8 EXPLAINABLE