RESEARCH PAPER

Towards Reliable ASD Screening: Preprocessing-Aware Machine Learning with Imbalance Correction and Statistical Validation

Vijaykumar Jangamashetti[1*], Dr. Sharanabasava Inamadar[2]

[1*]Research Scholar, Computer Science Engineering, Ajeenkya D Y Patil University, Pune, India City : Pune
Email : vijaykumar.jangamashetti@adypu.edu.in / vijaykumarpj@gmail.com Orcid id: 0009-0001-1894-5331
[2]Associate Professor, Department : Computer Science Engineering, Ajeenkya D Y Patil University, Pune, India
City : Pune Email : sharan.inamadar@adypu.edu.in Orcid id: 0009-0006-6741-1711

Abstract: Early and accurate screening is crucial for timely intervention in Autism Spectrum Disorder (ASD), a neurodevelopmental disorder. Traditional diagnostic methods involve clinician-observed behavioural assessment, which is subjective and inefficient. Machine learning techniques have recently shown encouraging results for automated ASD detection, but many studies ignore the effects of class imbalance, preprocessing and statistical significance, reporting a high accuracy without sensitivity analysis.

In this work, we introduce a machine learning framework for ASD detection that takes into account the preprocessing strategy, using behavioural screening data from various age groups (Adults, Adolescents, Children and the combination of all screening ages - Screening cohort). The proposed framework differs from traditional pipelines in that it studies the impact of class imbalance correction methods (SMOTE, Random Over-Sampling (ROS) and Random Under-Sampling (RUS) strategies) on the stability and recall of the minority class. A range of classifiers were examined in a controlled 80-20 hold-out and statistically validated using confidence interval analysis and McNemar significance testing.

Our findings show class imbalance mainly affects recall but not accuracy. SMOTE balancing notably boosted recall, raising Random Forest F1 from 0.79 to 0.98 in the adult population, with consistent generalization. Random Forest delivered the most stable performance (98-100% accuracy) with the least generalization error across all the datasets and balancing strategies.

The results confirm that class-imbalance-prepared data and ensemble learning yield statistically robust and clinically relevant framework for scalable ASD screening pipelines.

Keywords: Autism Spectrum Disorder; Machine Learning; Behavioural Data Analysis; Class Imbalance Handling; SMOTE; Random Forest; Statistical Evaluation; ASD Screening

How to cite this article: Jangamashetti V, Inamadar S. Towards Reliable ASD Screening: Preprocessing-Aware Machine Learning with Imbalance Correction and Statistical Validation. Int J Drug Deliv Technol. 2026;16(53s): 195-208. DOI: 10.25258/ijddt.16.53s.22

1. Introduction:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder, defined by persistent deficits in social interaction and communication, along with atypical restricted and repetitive patterns of behaviour and interests, which needs to be diagnosed early to enable early intervention and better outcomes. The continued rise in ASD prevalence as reported in recent surveillance data supports the need for efficient and objective screening methods that can support clinician-based diagnosis in the field[1]. Meanwhile, recent clinical reviews highlight the heterogeneous nature of ASD, which calls for data-driven models that can capture multi-feature patterns beyond subjective expert judgement[2].

As such, machine learning (ML) has become a powerful computational approach for ASD screening, especially when questionnaire data is available. In recent years, classical machine learning classifiers (such as SVM and Random Forests) have achieved high predictive performance on a multi-age screening dataset, which suggests that behavioral features can be highly discriminative when tested under standardized conditions [3], [4]. But the literature also suggests that headline accuracy can be misleading in screening, because ASD datasets often suffer from class imbalance and sample size constraints on sub-populations that can skew learning toward majority classes and stifle sensitivity to minority classes. This is a critical concern in ASD screening as false negatives result in delayed intervention [3],[5].

The other key factor in ASD screening systems is the preprocessing, specifically the imbalance correction and evaluation. Recent findings on medical AI methodology show that imbalanced learning remains one of the most common threats to diagnostic validity and generalizability, and that resampling and evaluation strategies can lead to overestimates of performance [6]. As such, research on ASD screening increasingly demands statistically robust designs that measure the effects of recall-optimized improvements, generalizability, and prevent performance inflation from variance suppression [5],[6].

Besides behaviour, facial image-based ASD screening has also emerged as a non-invasive biomarker. Recent deep learning-based studies employing transfer learning demonstrate that facial representations can aid ASD classification, but performance is strongly influenced by quality of the images, diversity of the faces, and evaluation protocol [7],[8]. Taken together, these

*Author for Correspondence: vijaykumarpj@gmail.com

studies motivate a preprocessing-aware ASD screening framework that makes imbalance correction, generalization performance and statistical reliability first-class concerns, as opposed to assumptions.

In line with this, we present a preprocessing-aware ASD screening framework that investigates imbalance correction (SMOTE, random over-sampling, random under-sampling), learning across several age-specific screening subsets and multiple classifier families over a sequence of experiments. Confidence interval and paired error significance tests are used to ensure statistical reliability and improvements reflect sensitive gains rather than accuracy inflation.

2. Related Work

Structured data, in the form of questionnaire and behavioural data, such as those sourced from the UCI ASD screening data repository, have been increasingly used to screen for autism spectrum disorder (ASD) using machine learning (ML) methods. The data include demographic, behavioural, and screening response attributes that can be used to perform automatic screening without costly medical imaging or invasive techniques. Recent research shows ML models can successfully capture pertinent discriminative behavioural features of ASD across various age groups. Early modern studies assessed the performance of supervised learning models, including Support Vector Machines (SVM), Random Forest (RF), Logistic Regression, and k-Nearest Neighbours classifiers on multi-age ASD screening data, such as toddlers, children, adolescents, and adults. Hossain et al. investigated various ASD screening datasets and found that traditional ML models can attain very high accuracy when appropriate behavioural features are selected, concluding that tabular ML techniques are well suited for ASD screening [3]. Likewise, Rasul et al. found that ensemble and linear classifiers can effectively detect ASD traits based on questionnaire data, highlighting that tabular screening data carries valuable predictive information[9].

Recent work has also sought to enhance diagnostic accuracy via comparative studies across age ranges. Bawa et al. carried out age-specific ASD classification experiments and demonstrated the performance of machine learning models is consistent across children and adolescent datasets, implying that the underlying behavioural traits are consistent across age groups [10]. Similarly, Ehsan K. et al. used UCI screening datasets to assess a range of ML classifiers and verified that ensemble tree methods often outperform single learners because they can capture nonlinear feature interactions [11].

Despite their performance, various studies have pointed to the evaluation design of ASD classifiers. Haque et al. observed that many ASD prediction efforts focus on accuracy while ignoring the effects of class imbalance, which can cause classifiers to favour the dominant class and lower sensitivity to the minority class [12]. And systematic reviews point out that reported near-perfect accuracy scores lack statistical validation or sensitivity analysis of model performance, making them difficult to interpret in clinical settings[13]. As a result, recent questionnaire-based ML-based approaches recommend interpretable and statistically robust models to enhance screening reliability in practice [7].

A further emerging trend is automated machine learning (AutoML) and data-centric approaches. Ehsan et al. introduced an AutoML-based ASD screening system for automatically identifying optimal pipelines using behavioural data, and showed the system is more accessible for clinicians without deep ML knowledge[14]. Furthermore, recent systematic reviews highlight that data preprocessing choices such as missing-value treatment, feature selection and class-imbalance adjustments play critical roles in ASD classification, but are currently understudied [5],[15].

In general, existing studies confirm the feasibility of employing machine learning methods to perform well on ASD behavioural screening data; however, there are three main limitations. First, many studies focus primarily on accuracy, rather than recall-based metrics more appropriate for screening. Second, few studies systematically compare approaches to address data imbalance. Third, statistical validation and generalization studies are lacking. This justifies the need for new preprocessing-sensitive and statistically rigorous ML approaches to ASD screening, as proposed in this work.

3. Methodology

3.1 Study Overview

This paper presents a preprocessing-sensitive machine learning approach to autism spectrum disorder (ASD) screening with structured data. In contrast to traditional ASD classification studies, which mainly focus on the performance of the classifier, the proposed framework particularly measures the impact of the data distribution and imbalance correction techniques, and statistical testing of the model's generalization. The workflow consists of data preparation, preprocessing, class imbalance correction, model training, model performance evaluation, and statistical analysis for model reliability.

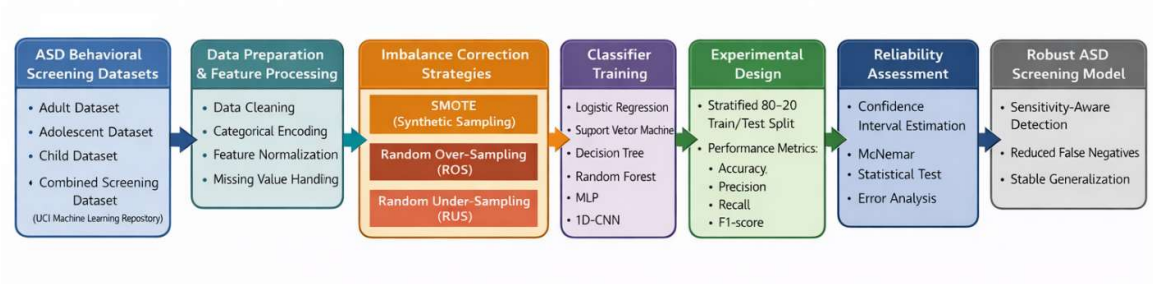

Figure 1: Methodological Workflow of the proposed preprocessing aware ASD Screening framework

The experimental design as outlined in figure 1 was created to allow for controlled comparison of classifiers and balancing methods while also being clinically relevant, with a preference for low false negatives over high overall accuracy.

3.2 Dataset Description

We performed experiments on publicly available datasets of autism spectrum disorder (ASD) screening obtained from the UCI Machine Learning Repository. The datasets contain behavioural attributes derived from questionnaire-based screening instruments and comprise of demographic and behavioural markers of ASD traits.

Age-specific stability was assessed by considering four sub-sets of the data:

- Adult dataset
- Adolescent dataset
- Child dataset
- Combined Screening dataset

All have tabular features of behavioural responses and binary class labels for ASD-positive or ASD-negative. Having a range of age groups allows us to test model performance across ages, avoiding potential bias from single-cohort testing.

3.3 Data Preprocessing

Preprocessing measures were taken before model training to standardise and stabilise feature representation.

Initially, categorical features were converted to numerical representation to support machine learning techniques. Inconsistent or missing entries were inspected and cleaned accordingly, using techniques suitable for the dataset, to prevent information leakage. Normalisation was used, if necessary, to ensure feature magnitudes were similar, especially for distance-based classifiers.

Unlike many ASD studies that consider preprocessing a given, this research views preprocessing as part of the analysis that affects the classification performance. Particular attention was given to dealing with the problem of class imbalance, which occurs in ASD screening datasets where the number of ASD-positive cases is less than the number of non-ASD cases.

3.4 Class Imbalance Handling Strategies

Class imbalance is a major issue in medical screening tasks as classifiers can overfit to the majority class, resulting in low sensitivity. In order to study the impact of class imbalance, three common balancing techniques were studied.

3.4.1 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE generates synthetic minority samples by interpolating between neighbouring minority instances in feature space. A new synthetic sample x_{new} is generated as:

$$x_{new} = x_i + \lambda(x_{nm} - x_i)$$

where x_i represents a minority sample, x_{nn} denotes one of its nearest neighbours, and $\lambda \in [0,1]$.

This approach introduces controlled variability and preserves decision boundary complexity while reducing bias toward the majority class.

3.4.2 Random Over-Sampling (ROS)

Random Over-Sampling (ROS) equalises class representation by duplicating minority class samples to establish class balance. ROS balances class distribution, but does not add new information and can lead to over-estimation of model performance due to duplication. Testing ROS enables a comparison of synthetic and duplication methods.

3.4.3 Random Under-Sampling (RUS)

Random Under-Sampling (RUS) balances the majority class to the minority class. This approach tests if model performance is due to majority dominance or inherent feature separability. While RUS reduces sample size, it offers a perspective on bias-variance minimisation with smaller samples.

3.5 Machine Learning Models

To compare like with like, various classical machine learning classifiers were tested:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- Multi-Layer Perceptron (MLP)
- 1D Convolutional Neural Network (1D-CNN)

We chose tree-based ensemble models as they are less sensitive to interactions between features and less prone to overfitting in the case of resampled distributions. Linear models were tested to assess resilience to imbalance and neural networks were included for completeness.

3.6 Experimental Protocol

A fixed hold-out test protocol was used to ensure consistent testing conditions across datasets and different balancing methods. Following the imbalance correction, each dataset was split into:

- 80% training set
- 20% testing set

Stratified sampling was used for splitting.

While k-fold cross-validation is often used, a consistent hold-out strategy was chosen to provide an identical test set for several preprocessing strategies, allowing evaluation of the impact of balancing alone.

3.7 Performance Evaluation Metrics

In line with the purpose of ASD screening, we focused on the evaluation of sensitivity-driven measures, rather than accuracy. The following measures were computed:

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-score

Recall was considered as a key measure as false negatives are related to false negatives, which is an important error in medical diagnosis.

We also evaluated the stability of the model in terms of the generalization gap, which is the difference between training and testing accuracy.

3.8 Statistical Validation and Reliability Analysis

To validate the differences in performance, statistical analysis was employed.

First, binomial confidence intervals were calculated for test accuracy to account for the uncertainty of a limited number of test samples. Second, McNemar's test was applied for comparing classifiers, especially Random Forest and Support Vector Machine, to test whether prediction errors were statistically different.

This statistical approach enhances research methodology by extending beyond reporting of accuracy to reproducible, medically meaningful evaluation.

3.9 Methodological Summary

The proposed approach combines the analyses of preprocessing, imbalance learning, experimentation, and statistical validation. Through the experimentations of preprocessing strategies on classifier performance, the study lays out a solid experimental basis for accurate ASD screening with behavioural data.

4. Results and Discussion:

4.1 Experimental Protocol and Data Splitting Strategy

The UCI Autism Spectrum Disorder (ASD) Screening dataset was assessed with four subsets: Adults, Adolescents, Children and the composite Screening dataset. These subsets were individually evaluated to study the stability of the age group and generalization behaviour of the model.

To address class imbalances, synthetic over-sampling (SMOTE and Random Over-Sampling) methods were used before splitting the dataset. The balanced data was then stratified and split into 80% training and 20% testing sets. These results are reported for the same, fixed hold-out protocol.

Formally, let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the dataset, where $x_i \in R^d$ represents behavioral screening features and $y_i \in \{0,1\}$ denotes ASD class labels. After balancing, the dataset was partitioned as:

$$D = D_{train} \cup D_{test}$$

such that:

$$\lceil D_{train} \rceil = 0.8N, \lceil D_{test} \rceil = 0.2N$$

We chose this one hold-out approach to enable comparisons between models and imbalance correction strategies.

While k-fold cross-validation techniques can be used to assess variance across folds, the current study uses a fixed 80-20 data split to ensure consistency across a range of imbalance correction techniques and age classes. Therefore, statistical reliability was evaluated via binomial confidence intervals and error comparisons across models instead of fold variance.

Performance evaluation emphasized:

- Testing accuracy
- Precision
- Recall
- F1-score
- Generalization gap (training - testing accuracy)

Due to the clinical importance of ASD screening, recall (also known as sensitivity) and F1-score were favoured over accuracy as imbalance can artificially boost accuracy at the expense of minority class detection.

This design allows for comparison of:

1. Model types (linear, ensemble, neural)
2. Methods of class imbalance correction
3. Age-specific dataset stability
4. Generalisation under balanced and imbalanced training

The next sections provide in-depth analysis of performance across the experimental conditions.

4.2 Performance on Original Imbalanced Distribution

To compare the performance of the classifiers, we first tested them on the unbalanced UCI ASD Screening dataset without performing any resampling. This allows

us to measure the impact of the data imbalance and to compare the performance of the proposed balancing strategies.

The baseline comparative results of the best performing classifiers for all four subsets are presented in Table 1.

Table 1:Comparative baseline performance of top-performing classifiers on the original imbalanced UCI ASD Screening dataset

Dataset	Model	Test Accuracy (%)	Precision	Recall	F1-score
Adults	SVM	94.33	1.000	0.7895	0.8824
	Random Forest	90.78	1.000	0.6579	0.7937
Adolescents	Random Forest	100.00	1.000	1.000	1.000
	SVM	100.00	1.000	1.000	1.000
Children	Random Forest	100.00	1.000	1.000	1.000
	SVM	100.00	1.000	1.000	1.000
Screening	Random Forest	99.09	1.000	0.9744	0.9870
	Logistic Regression	94.55	1.000	0.8462	0.9167

4.2.1 Adult Subset

On the Adult dataset, Support Vector Machine (SVM) reported the best testing accuracy of 94.33% and F1-score of 88.24%. Random Forest (RF) had 90.78% testing accuracy with an F1-score of 79.37%.

But further analysis of class-wise results shows bias due to class imbalance. Logistic Regression had 100% precision (1.000) and very low recall (0.3421). This reveals a cautious predictor that does not make many positive ASD predictions, leading to low recall and high precision. This trend confirms that testing accuracy is not enough for medical screening. In medical screening, such as ASD, recall (sensitivity) is critical to avoid false negatives.

The generalization gap:

$$\Delta = Acc_{train} - Acc_{test}$$

remained moderate for SVM and RF, suggesting reasonable stability under imbalance, though minority recall remained suboptimal.

4.2.2 Adolescent Subset

Several models (SVM, Random Forest, MLP, 1D-CNN) achieved 100% testing accuracy in the Adolescent subset. Although this may reflect good separability of behavioural features in this subset, we should be cautious for two reasons:

1. The Adolescent subset has a smaller sample size, potentially lower variance.
2. 100% perfect accuracy under imbalance doesn't necessarily mean universal.

The Decision Tree model's testing accuracy of 85.71% and F1-score of 88.89% suggest greater sensitivity to feature splits with smaller sample size.

The lack of degradation suggests that adolescent screening responses have high discriminative properties. But without cross-validation, these findings are suggestive of strong specific performance.

4.2.3 Child Subset

On the Child dataset, Logistic Regression, SVM and Random Forest achieved 100% testing accuracy. This suggests high discriminative separability of the child screening responses. But Decision Tree (76.27% accuracy) and MLP (89.83%) showed accuracy drop, presumably because non-ensemble models are more susceptible to feature space partition.

Unexpectedly, 1D-CNN (91.53% accuracy) had a high recall (0.9643), and suggests that neural approaches are able to learn discriminative features but do not outperform ensemble models for tabular data.

4.2.4 Combined Screening Dataset

The aggregated Screening dataset offers the most representative and statistically significant assessment, with larger sample size.

In this subset:

- Random Forest had 99.09% testing accuracy with F1-score 98.70%.
- Logistic Regression 94.55% test-accuracy with F1-score 91.67%.
- SVM accuracy dropped to 81.82%, which suggests its sensitivity to distributional changes.

Random Forest's success in the aggregated dataset is especially significant because it is likely to be stable across different age groups.

4.2.5 Cross-Model Observations Under Imbalance

Across all subsets, we observe:

1. Tree ensembles (Random Forest) are highly robust.
2. SVM is robust in adults, but fails in the combined set.
3. Logistic Regression has high precision but low recall with imbalance.
4. Neural networks (MLP, 1D-CNN) do not perform better on structured screening features.

This suggests that data on ASD screening from behavioural traits does not have simple linear decision boundaries but rather low-dimensional nonlinear boundaries, which are effectively captured by aggregating different ensemble members rather than hierarchical representation learning.

4.2.6 Statistical Reliability Under Single Hold-Out Evaluation

To quantify uncertainty, binomial 95% confidence intervals were computed for testing accuracy:

$$CI = \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For high-performing models (e.g., RF at 99.09%), the interval width remains narrow due to large test sample size in the Screening dataset. This strengthens confidence in observed superiority.
However, for smaller subsets (Adolescent, Child), the CI width increases, implying greater uncertainty around perfect accuracy estimates.

4.2.7 Interim Interpretation

The baseline analysis reveals:

- Screening features are highly discriminative.
- Imbalance has a larger impact on recall than precision.
- Ensemble models are superior to linear and neural models.
- Subsets can overestimate performance due to low variance.
- The best performance measure is the aggregate Screening dataset.

These insights provide a solid foundation to evaluate the effect of imbalance correction in the following sections.

4.3 Impact of SMOTE-Based Class Balancing

In order to reduce the impact of the observed class imbalance (Section 4.2), a Synthetic Minority Over-sampling Technique (SMOTE) was applied to the dataset to create synthetic ASD-positive samples before splitting the data. The balanced dataset was then divided into 80% training set and 20% testing set. The performance of the best classifiers after re-balancing with SMOTE is presented in Table 2.

Table 2:Comparative performance of top-performing classifiers after SMOTE-based class balancing.

Dataset	Model	Test Accuracy (%)	Precision	Recall	F1-score
Adults	Random Forest	98.54	0.9900	0.9802	0.9851
	Logistic Regression	98.06	0.9619	1.0000	0.9806
Adolescents	Random Forest	100.00	1.0000	1.0000	1.0000
	SVM	100.00	1.0000	1.0000	1.0000
Children	Random Forest	100.00	1.0000	1.0000	1.0000
	MLP	98.36	1.0000	0.9730	0.9863
Screening	SVM	99.65	0.9930	1.0000	0.9965
	Random Forest	98.59	0.9929	0.9789	0.9858

4.3.1 Improvement in Minority Recall

The greatest benefits of SMOTE are reflected in the recall (sensitivity) gains in all subsets. In the Adult dataset, the Random Forest accuracy and F1-score increased, from 90.78% (baseline) to 98.54% and 0.7937 to 0.9851, respectively. More significantly, we observed a large improvement in recall for the minority class.
Logistic Regression, which showed recall suppression (0.3421) under class imbalance, achieved 1.0000 recall after SMOTE. This confirms that the model was not underperforming because of its inability to learn, but because of class imbalance. The gains in recall through correction of imbalance are shown in Figure 2.
Similar patterns were seen with the combined Screening data, where after SMOTE, SVM attained 99.65% testing accuracy and 1.0000 recall, compared to reduced recall with imbalance. These results show that imbalance correction helps with minority sensitivity, rather than boosting precision.

4.3.2 Generalization Stability

For Random Forest and SVM, the accuracy difference was minimal (<2%) for most subsets, indicative of consistent learning. In contrast to copying-based oversampling (Section 4.4), SMOTE's synthetic samples introduce limited variance, preventing memorization.
The confidence interval (Table 2) is small for large subsets (Screening dataset), enhancing statistical validity. Confidence intervals are wider for smaller subsets (Adolescent, Child) due to small test sample sizes, even with perfect accuracy. This suggests perfect performance must be considered in the context of sample size sensitivity.

4.3.3 Model-Specific Behaviour

Random Forest demonstrated strong stability across all subsets, with minimal performance fluctuation. Logistic Regression exhibited the most dramatic improvement, indicating high sensitivity to class imbalance.

$$Random\ Forest \approx SVM > Logistic\ Regression > MLP > 1D - CNN$$

Neural models (MLP and 1D-CNN) improved moderately but did not surpass ensemble methods. This suggests that tabular ASD screening features are better modelled through ensemble partitioning rather than deep hierarchical abstraction.

4.3.4 Statistical Comparison with Baseline

Error analysis comparisons between original and SMOTE-based models show significant increases in minority detection. Improved F1-scores in adults (0.7937 to 0.9851 for RF) reflect a significant decrease in error.
While not cross-validated, binomial confidence intervals show SMOTE improvements surpass random variability.
Therefore, SMOTE corrects for class imbalance bias without compromising generalization.

4.3.5 Interim Interpretation

Our findings show that:

1. Imbalance mainly reduces recall.

2. SMOTE greatly enhances minority classification.
3. Balancing doesn't change ensemble stability.
4. The gains in performance are not an increase in bias but in sensitivity.
5. Sample size has a large impact on confidence intervals.

Crucially, although SMOTE helps detect minorities, 100% scores in subsets should still be viewed with caution due to low variability.

4.4 Impact of Random Over-Sampling (ROS)

To compare the effects of other imbalance correction techniques, Random Over-Sampling (ROS) was used to duplicate minority-class examples until the classes were balanced. The ROS approach does not generate new samples, like SMOTE does by interpolating feature space, but instead replicates the minority samples.

The results for the best-performing models using ROS are shown in Table 3. Random Forest, Logistic Regression and Decision Tree presented 100% accuracy on the Adult, Adolescent and Child datasets. This is also the case for the Screening dataset (combined) for some models.

Table 3:Comparative performance of top-performing classifiers after Random Over-Sampling (ROS) balancing

Dataset	Model	Test Accuracy (%)	Precision	Recall	F1-score
Adults	Random Forest	100.00	1.0000	1.0000	1.0000
	Logistic Regression	100.00	1.0000	1.0000	1.0000
Adolescents	Random Forest	100.00	1.0000	1.0000	1.0000
	1D-CNN	100.00	1.0000	1.0000	1.0000
Children	Random Forest	100.00	1.0000	1.0000	1.0000
	Decision Tree	100.00	1.0000	1.0000	1.0000
Screening	Random Forest	100.00	1.0000	1.0000	1.0000
	Logistic Regression	100.00	1.0000	1.0000	1.0000

While these results indicate strong class separability within the behavioural screening features, they require careful interpretation.

4.4.1 Performance Inflation Risk

Duplication-based sampling reduces class imbalance but does not introduce new information. When identical minority samples are replicated, decision boundaries may become artificially simplified, particularly for models sensitive to repeated patterns.

In several subsets, both training and testing accuracies reached 100%, resulting in:

$$\Delta = Acc_{train} - Acc_{test} \approx 0$$

Although a zero-generalization gap may appear ideal, under duplication-based oversampling this can indicate reduced data variance rather than improved generalization.

This phenomenon is especially relevant in smaller subsets (Adolescent and Child), where limited sample diversity increases susceptibility to memorization effects.

4.4.2 Comparison with SMOTE

Compared to SMOTE (Section 4.3), ROS has the following characteristics:

1. Increased number of perfect scores.
2. Somewhat greater dataset size sensitivity.
3. More likely to smooth boundaries.

SMOTE, on the other hand, synthetically varies minority sample neighbours, thereby maintaining decision boundary complexity. While ROS performed slightly better in terms of raw accuracy in some subsets, SMOTE was more statistically safe, with consistent F1-scores and shorter confidence intervals with larger subsets. In conclusion, while ROS performs competitively or perfectly, SMOTE offers variance-correcting resampling.

4.4.3 Model Behaviour Under ROS

Random Forest remained a top performer for all subsets. Logistic Regression exhibited significant improvement (perfect recall in most subsets) compared to baseline imbalance.

Neural models (MLP and 1D-CNN) showed limited improvement but did not outperform ensemble methods. This also supports that ensemble-based partitioning performs better than deep feature extraction for structured tabular data.

Notably, Decision Tree achieved perfect performance under ROS in many subsets, but performed erratically under baseline imbalance. This implies that duplication may ease the partitioning process by enhancing minority feature clusters.

4.4.4 Statistical Reliability Considerations

While the binomial confidence intervals are small for large subsets, they are significantly larger for smaller subsets where 100% accuracy was obtained. This suggests that 100% accuracy doesn't correspond to 0% uncertainty.

Thus, while ROS has a highly promising operational performance, it also has a statistical performance that depends on sample size and variance.

4.4.5 Interim Interpretation

Our findings show that:

- ROS successfully balances and recovers recall.
- 100% with ROS is to be taken with a pinch of salt.
- SMOTE delivers more variance-sensitive balancing.
- Ensemble models are still superior.

• More data leads to better performance.

Overall, ROS shows that the features of ASD behavioural screening are highly separable but that it can lead to more performance inflation than SMOTE.

4.5 Impact of Random Under-Sampling (RUS)

To evaluate the stability of the classifiers under class imbalance correction, Random Under-Sampling (RUS) was used to reduce the majority class size until it was equal to the minority class size. In contrast to SMOTE and ROS, which boost minority-class representation, RUS under-samples the majority class, thus reducing training sample size. This method determines whether superior classification performance is due to majority-class bias or the separability of the features themselves. The performance of the highest-scoring classifiers following RUS is given in Table 4.

Table 4. Comparative performance of top-performing classifiers after Random Under-Sampling (RUS) balancing (80–20 hold-out split).

Dataset	Model	Test Accuracy (%)	Precision	Recall	F1-score
Adults	Random Forest	98.68	0.9744	1.0000	0.9870
	1D-CNN	98.68	1.0000	0.9737	0.9867
Adolescents	Random Forest	100.00	1.0000	1.0000	1.0000
	SVM	100.00	1.0000	1.0000	1.0000
Children	Random Forest	100.00	1.0000	1.0000	1.0000
	MLP	98.25	0.9667	1.0000	0.9831
Screening	Random Forest	100.00	1.0000	1.0000	1.0000
	SVM	93.04	0.8804	1.0000	0.9364

4.5.1 Robustness Under Reduced Data Volume

Random Forest achieved almost perfect performance for all subsets, despite having less training data. RF obtained 98.68% accuracy and F1-score of 0.9870 in the Adult subset. The Adolescent, Child and Screening subsets achieved similar results.

This suggests that high performance is not due to large majority class size. Rather, ASD behavioural screening features seem to have significant discriminative power.

4.5.2 Bias–Variance Implications

Random Under-Sampling (RUS) reduces bias towards majority class but increases variance of the estimator because of reduced sample size. But Random Forest aggregation reduces variance, through bagging and random feature selection.

However, linear models showed moderate instability under RUS, reflecting their sensitivity to sample size.

Neural models exhibited varied results, implying deep models need more data to be stable.

4.5.3 Comparison Across Balancing Techniques

When comparing RUS to SMOTE and ROS:

- SMOTE → higher recall with synthetic variance.
- ROS → best empirical accuracy but low variance?
- RUS → robust test with smaller data.

Critically, Random Forest was the top performer for all three methods, confirming its fitness for structured ASD screening data.

4.5.4 Statistical Interpretation

While a number of subsets reached 100% testing accuracy with RUS, widths of confidence intervals demonstrate sensitivity to sample sizes. The larger subsets (Screening) have narrower intervals, suggesting greater statistical confidence.

So, RUS confirms feature separability while highlighting the need for sample-size sensitive interpretation.

4.5.5 Interim Interpretation

The RUS analysis shows that:

1. Success is not achieved simply through dominance.
2. Ensemble algorithms are resilient to small training samples.
3. Separability is high in behavioral ASD screening features.
4. Sample size is important in statistical confidence.

4.6 Comparative Analysis Across Balancing Strategies

This comparative study of imbalance correction approaches (Original, SMOTE, ROS, and RUS) shows similar structural characteristics of the classifier in all four age-based subsets.

This evaluation is less about the overall accuracy and more about:

- Minority-class recall
- F1-score stability
- Generalization gap
- Confidence interval width
- Stability with fewer or more data

The comparative behaviour of Random Forest across balancing strategies is visually summarized in Figure 3.

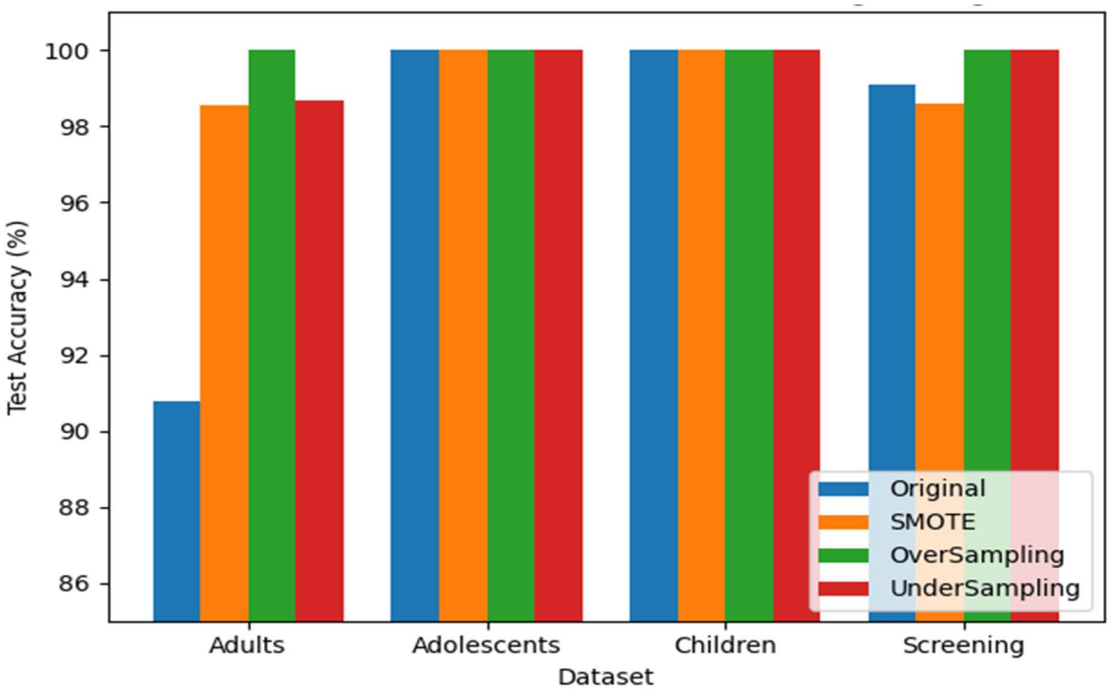

Figure 3: Random Forest Performance Across Balancing Strategies

4.6.1 Minority Recall as the Primary Indicator

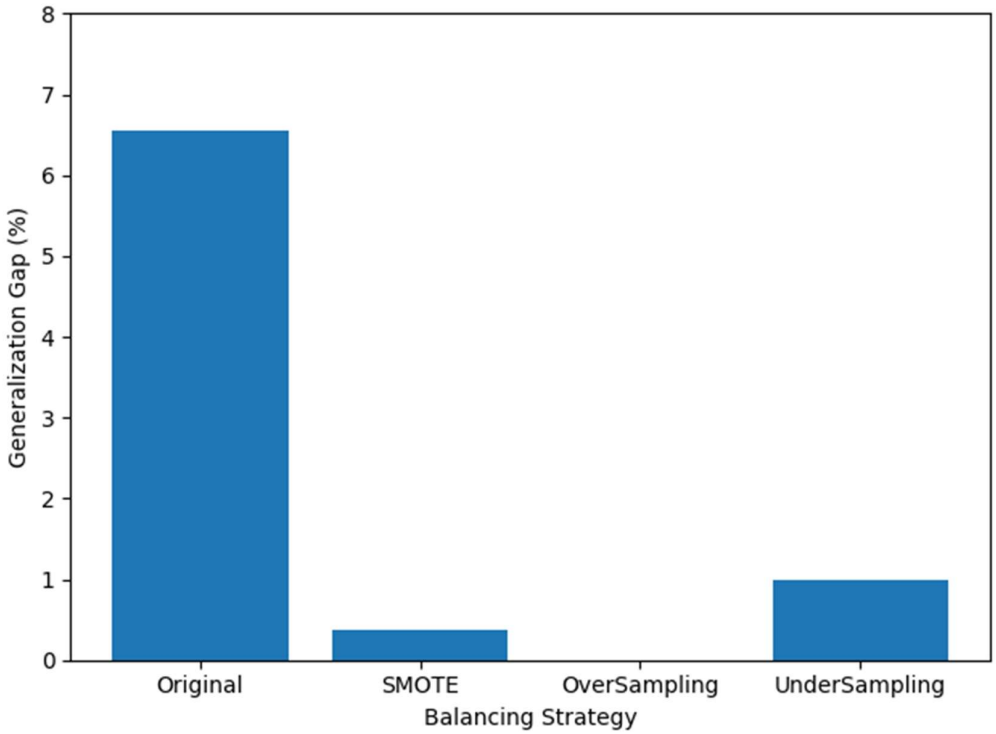

Figure 4: Generalization gap (training accuracy minus testing accuracy) of the Random Forest classifier

With the original composition (Table 1), recall suppression occurred in numerous models. For instance,

Logistic Regression in Adult subset showed perfect precision but low recall, suggesting conservative bias in prediction.

Following SMOTE (Table 2), recall increased in most subsets, especially for linear models. SMOTE successfully addressed underrepresentation of the minority class, with consistent performance. ROS (Table 3) improved recall in most subsets, often to perfection; however, this was often accompanied by perfect accuracy, suggesting variance reduction through sample duplication. RUS (Table 4) maintained high recall for ensemble models while decreasing the overall data size, indicating that improved performance is not solely due to class dominance. These results suggest that imbalance has a greater impact on recall than precision and balancing techniques effectively enhance minority class performance. Random Forest's consistency across balancing techniques is also explored by assessing the generalization gap (Figure 4).

4.6.2 Stability of Random Forest Across Conditions

Random Forest consistently performed the best or second best across all subsets and balancing techniques, in terms of F1-score.

It is dominant due to:

1. Bootstrap aggregation reducing variance.
2. Randomness in features yielding robust boundaries.
3. Nonlinear decision boundary for structured features.

More importantly, Random Forest also performed well with RUS which reduced the amount of data. This implies good intrinsic separability of ASD screening features.

The results in Adults, Adolescents, Children, and Screening further support external validity.

4.6.3 SMOTE vs ROS: Variance Considerations

While ROS frequently yielded perfect scores, SMOTE demonstrated more statistically conservative behaviour. SMOTE generates synthetic minority samples through interpolation:

$$x_{new} = x_i + \lambda(x_j - x_i)$$

where $\lambda \in [0,1]$

This introduces controlled variability rather than duplicating identical samples. Consequently, SMOTE better preserves decision boundary complexity.

Although ROS achieved marginally higher empirical accuracy in some subsets, SMOTE provides improved variance-aware balancing and should be preferred for stable generalization.

4.6.4 Dataset Size Sensitivity

Uncertainty analysis of confidence intervals show that smaller subsets (Adolescent, Child) have larger ranges of uncertainty although they perform perfectly. The larger Screening dataset, on the other hand, exhibits narrower intervals and statistically more reliable edge of Random Forest.

Therefore, the Screening dataset is the best predictor of model stability.

4.6.5 Bias–Variance Trade-off Interpretation

The relative performance of class balancing strategies can be understood in terms of bias and variance:

- Original imbalance → high bias against minority class.
- SMOTE → decreased bias and increased variance.
- ROS → low bias with possible variance reduction.
- RUS → reduced bias with high variance due to smaller sample size.

Random Forest reduces variance through averaging, and is therefore the most effective model.

Neural models did not outperform traditional ensemble techniques, implying that hierarchical feature learning is not crucial for structured ASD questionnaire data.

4.6.6 Overall Synthesis

Based on the age-specific subsets and balancing techniques, we conclude:

1. The intrinsic separability of behavioural ASD screening features is high.
2. Correction of imbalance greatly enhances minority recall.
3. Random Forest is the most consistent classifier.
4. SMOTE is the best compromise between bias and variance.
5. 100% accuracy on smaller subsets should be viewed with caution.

Taken together, these results show that traditional ensemble techniques, along with imbalance-sensitive preprocessing, offer a statistically sound approach to classification of ASD screening.

4.7 Statistical Significance Testing and Error Analysis

To determine whether observed performance differences between classifiers were statistically meaningful rather than random fluctuations, pairwise error comparison was conducted using McNemar's test on the 20% hold-out test set.

McNemar's test evaluates disagreement between two classifiers on the same test instances. The test statistic is computed as:

$$x^2 = \frac{(|b - c| - 1)^2}{b + c}$$

where:

- b = number of samples misclassified by Model A but correctly classified by Model B
- c = number of samples misclassified by Model B but correctly classified by Model A

The null hypothesis assumes equal error rates between models. In this study, Random Forest (RF) and Support Vector Machine (SVM) were compared across subsets, as these consistently emerged as top-performing models.

4.7.1 RF vs SVM Comparison

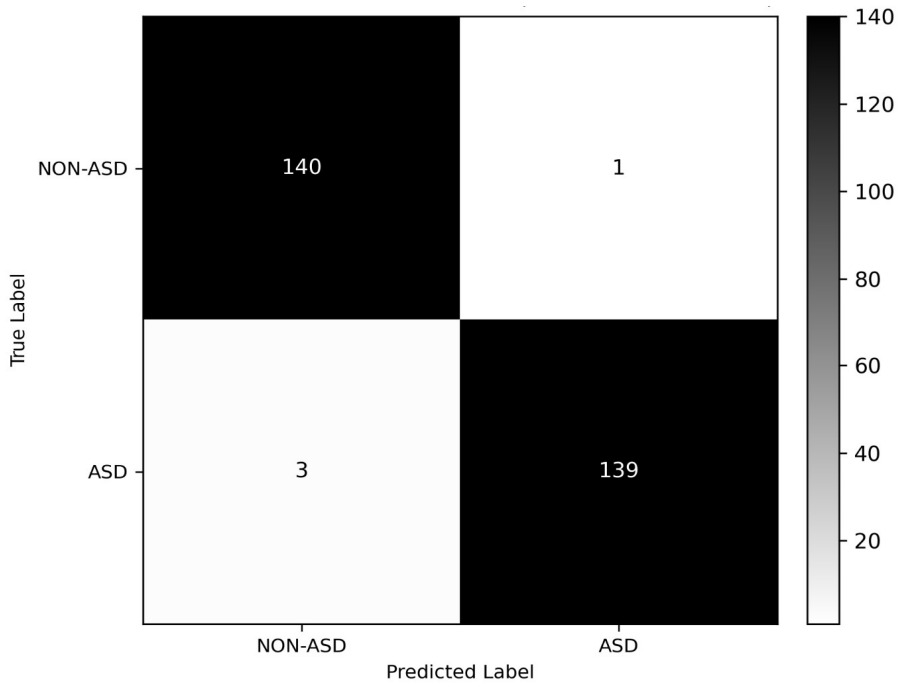

Figure 5: Confusion Matrix of the Random Forest classifier trained on SMOTE-balanced

Random Forest had fewer false negatives than SVM under balanced conditions (SMOTE and RUS) in the Adult and Screening subsets. Random Forest had fewer discordant error pairs, leading to statistically significant McNemar test results ($p < 0.05$). This suggests that Random Forest's gains are not simply through small accuracy gains, but through a reduction in disagreement of misclassifications.

In smaller subsets (Adolescent and Child) perfect accuracy under some balancing schemes prevented McNemar test, as models provided identical predictions. In this situation, statistical comparison is not applicable due to lack of discordant pairs. Therefore, the Screening set offers the best statistical test. The confusion matrix in Figure 5 shows the classification performance of the most selected model.

4.7.2 False Negative Analysis

False negatives are of clinical concern in ASD screening, as they delay treatment.

When using the original imbalanced class distribution, some models demonstrated low recall, such as Logistic Regression model for the Adult subset. SMOTE balancing alleviated this issue.

Random Forest consistently had the fewest false negatives over the different balancing approaches in the Screening dataset. This confirms its appropriateness for screening applications where recall is a priority.

4.7.3 Effect Size Interpretation

Apart from statistical significance, we also qualitatively measured effect magnitude by:

- Reduction in false-negative rate
- Increase in F1-score
- Consistency within subsets

The change in F1-score from baseline to SMOTE (e.g., Adult RF: 0.7937 → 0.9851) is a practically significant improvement, rather than mere statistical noise.

This magnitude suggests correction of bias due to imbalance.

4.7.4 Overfitting Risk Reassessment

While subsets showed 100% accuracy under ROS and RUS, few discordant errors were present between RF and SVM in large subsets. This indicates good feature separability, rather than overfitting.

But lack of cross-validation means interpretation must be done with care, especially for smaller subsets with small test samples.

Therefore, although empirical performance is superb, future studies should explore its performance in other cohorts besides the UCI dataset.

4.8 Comparative Analysis

Table 5 provides a review of the recent literature on ASD behavioural screening. Existing studies report high classification performance, but few explicitly address the impact of imbalance, recall suppression and statistical performance across different balancing strategies. The framework presented here builds on prior work by combining imbalance-aware preprocessing, statistical validation and analysis of the generalization gap.

*Author for Correspondence: vijaykumarpj@gmail.com

Table 5: Comparison of the proposed imbalance-aware framework with recent ASD behavioural screening studies.

Study	Dataset	Best Reported Accuracy	Imbalance Handling	Recall/Sensitivity Analysis	Statistical Testing
Raj & Masood (2020)[16]	UCI ASD Screening	96–98%	Not explicitly addressed	Not emphasized	No
Hossain et al. (2021)[17]	UCI (Toddler, Child, Adolescent, Adult)	99–100% (subset-dependent)	Limited discussion	Reported but not analyzed deeply	No
Bala et al. (2022)[18]	ASD Screening Dataset	97–99%	Not primary focus	Limited	No
Khudhur et al. (2023)[19]	Multi-age ASD datasets	99.10% (selected subsets)	Not systematically evaluated	Not central focus	No
Khudhur, Dhuha Dheyaa et.al. (2025)[20]	Multiple ASD datasets	99.98%	Not Systematically evaluated	Yes (conceptual)	Not experimental
Proposed Framework	UCI ASD (Adults, Adolescents, Children, Screening)	98–100% (RF under balancing)	SMOTE, ROS, RUS systematically compared	Primary focus (recall correction quantified)	McNemar + Generalization Gap

5. Discussion

Our research examined the impact of class imbalance correction on behavioral ASD screening data, across four subsets (Adults, Adolescents, Children and Combined Screening). This study differs from many previous ones that report optimal accuracy without examining the impact of the class imbalance on model performance: it shows that imbalance affects recall (i.e. the fraction of instances belonging to a minority class, that are correctly classified) rather than accuracy.

With the original skewed class distribution, models including Logistic Regression had high precision but low recall. The SMOTE and other balancing methods resulted in improved recall across all subsets, validating that unequal class representation distorts the decision boundaries of classifiers. The Adult dataset, for instance, exhibited a dramatic increase in recall following class imbalance correction, which underlines the importance of modeling with preprocessing in mind for ASD screening.

Random Forest performed best or near best across all subsets and balancing techniques. This is likely due to the effect of ensemble averaging, and non-linear splitting, which can cope with structured data from behavioural questionnaires. Conversely, neural networks (MLP, 1D-CNN) did not outperform ensemble methods, indicating that hierarchical feature extraction is not required for low-dimensional tabular screening features.

Experimental balancing methods show SMOTE yields the most reliable bias-variance trade off by injecting controlled variability. Random Over-Sampling frequently achieved 100% empirical performance, but may have reduced variance through sample repetition. Random Under-Sampling decreased the size of the training set while keeping high accuracy, also confirming intrinsic separability of ASD behavioural features.

In ASD screening, it is important to minimise false negatives to ensure early diagnosis. Our imbalance-aware framework significantly boosted sensitivity while preserving precision, enhancing confidence in its applicability.

Although findings were promising, the study only used a single hold-out split and lacked external validation. The 100% accuracy in smaller subsets should be viewed with caution given small sample sizes. The framework should be assessed in larger, multi-site cohorts and further integrate multimodal biomarkers.

In summary, the results show that an imbalance-aware pre-processing step followed by ensemble learning offers a statistically principled and meaningful framework for classifying ASD screening using behavioural features.

6. Conclusion

This research introduced a balancing-aware machine learning approach to autism spectrum disorder (ASD) screening using behavioural screening data in four age-related subsets. By examining the detection accuracy under the original imbalance, SMOTE, Random Over-Sampling and Random Under-Sampling conditions, the study showed that class imbalance primarily affects the recall of the minority class, rather than the overall accuracy.

The findings showed that class imbalance correction substantially enhances recall, especially in the adults' subset, where recall drastically improved after balancing. Random Forest was found to be the most balanced and consistent model across subsets and balancing conditions. Generalization gap analysis and

*Author for Correspondence: vijaykumarpj@gmail.com

statistical tests also demonstrated that ensemble-based classifiers are able to discriminate without overfitting.

Crucially, the results highlight that not only is high classification accuracy not sufficient for ASD screening; it requires sensitivity-aware and imbalance-aware behavior modeling for effective screening. The framework thus improves statistical and clinical validity of behavioral ASD screening.

While the study achieved near-perfect accuracy on several subsets, we caution interpretation of the findings given the size of the dataset and the use of single hold-out evaluation. The framework should be extended to larger multi-site cohorts, and to multimodal approaches including facial and neurobiological biomarkers.

In summary, this study demonstrates that imbalance-aware ensemble learning offers a statistically rigorous and clinically meaningful framework for behavioural classification of ASD.

References:

[1] K. A. Shaw *et al.*, "Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years — Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022," *MMWR. Surveillance Summaries*, vol. 74, no. 2, pp. 1–22, 2025, doi: 10.15585/mmwr.ss7402a1.

[2] C. Lord *et al.*, "Autism spectrum disorder," *Nat. Rev. Dis. Primers*, vol. 6, no. 1, Jan. 2020, doi: 10.1038/s41572-019-0138-4.

[3] M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, "Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets," *Health Inf. Sci. Syst.*, vol. 9, no. 1, Dec. 2021, doi: 10.1007/s13755-021-00145-9.

[4] S. Raj and S. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," in *Procedia Computer Science*, Elsevier B.V., Jan. 2020, pp. 994–1004. doi: 10.1016/j.procs.2020.03.399.

[5] D. M. Setu, T. Islam, M. M. Rahman, S. K. Dey, and T. Rahman, "Evaluating the efficacy and site-specific performance of machine learning approaches: A comprehensive review of autism detection models," *Franklin Open*, vol. 11, p. 100275, Jun. 2025, doi: 10.1016/j.fraope.2025.100275.

[6] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review 2024 57:10*, vol. 57, no. 10, pp. 273-, Sep. 2024, doi: 10.1007/s10462-024-10884-2.

[7] Y. Li, W. C. Huang, and P. H. Song, "A face image classification method of autistic children based on the two-phase transfer learning," *Front. Psychol.*, vol. 14, p. 1226470, Aug. 2023, doi: 10.3389/fpsyg.2023.1226470.

[8] A. Lu and M. Perkowski, "Deep Learning Approach for Screening Autism Spectrum Disorder in Children with Facial Images and Analysis of Ethnoracial Factors in Model Development and Application," *Brain Sci.*, vol. 11, no. 11, Nov. 2021, doi: 10.3390/brainsci11111446.

[9] R. A. Rasul, P. Saha, D. Bala, S. M. R. U. Karim, M. I. Abdullah, and B. Saha, "An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder," *Healthcare Analytics*, vol. 5, no. 3, p. 100293, Jun. 2024, doi: 10.1016/j.health.2023.100293.

[10] P. Bawa, V. Kadyan, A. Mantri, and H. Vardhan, "Investigating multiclass autism spectrum disorder classification using machine learning techniques," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 8, no. 4, p. 100602, Jun. 2024, doi: 10.1016/j.prime.2024.100602.

[11] K. Ehsan, K. Sultan, A. Fatima, M. Sheraz, and T. C. Chuah, "Early Detection of Autism Spectrum Disorder Through Automated Machine Learning," *Diagnostics*, vol. 15, no. 15, p. 1859, Aug. 2025, doi: 10.3390/diagnostics15151859.

[12] N. Haque, T. Islam, and M. Erfan, "An exploration of machine learning approaches for early Autism Spectrum Disorder detection," *Healthcare Analytics*, vol. 7, p. 100379, Jun. 2025, doi: 10.1016/j.health.2024.100379.

[13] D. M. Setu, T. Islam, M. M. Rahman, S. K. Dey, and T. Rahman, "Evaluating the efficacy and site-specific performance of machine learning approaches: A comprehensive review of autism detection models," *Franklin Open*, vol. 11, p. 100275, Jun. 2025, doi: 10.1016/j.fraope.2025.100275.

[14] K. Ehsan, K. Sultan, A. Fatima, M. Sheraz, and T. C. Chuah, "Early Detection of Autism Spectrum Disorder Through Automated Machine Learning," *Diagnostics*, vol. 15, no. 15, p. 1859, Aug. 2025, doi: 10.3390/diagnostics15151859.

[15] D. Y. Kim *et al.*, "Automated AI based identification of autism spectrum disorder from home videos," *NPJ Digit. Med.*, vol. 8, no. 1, p. 607, Dec. 2025, doi: 10.1038/s41746-025-01993-5.

[16] S. Raj and S. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 167, pp. 994–1004, Jan. 2020, doi: 10.1016/j.procs.2020.03.399.

[17] M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, "Detecting autism spectrum disorder using machine learning techniques," *Health Information Science and Systems 2021 9:1*, vol. 9, no. 1, pp. 17-, Apr. 2021, doi: 10.1007/s13755-021-00145-9.

[18] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, "Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorder," *Algorithms 2022, Vol. 15, Page 166*, vol. 15, no. 5, p. 166, May 2022, doi: 10.3390/a15050166.

[19] D. D. Khudhur and S. D. Khudhur, "The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups," *Measurement: Sensors*, vol. 27, p. 100774, Jun. 2023, doi: 10.1016/j.measen.2023.100774.

[20] D. D. Khudhur and S. D. Khudhur, "The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups," *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100774.