LONDON JOURNAL OF SOCIAL SCIENCES

eISSN 2754-7671

Validating the AQ-10 Autism Screening Threshold Using Machine Learning Models

Deepika Chandrashekar*

B. Vasumathi**

Abstract

Autism Spectrum Disorder (ASD) is a lifelong developmental condition for which early identification is essential. The Autism Quotient-10 (AQ-10) is frequently used as a quick screening instrument, classifying individuals scoring seven or more as potentially autistic. This study applies machine learning (ML) models to a public AQ-10 dataset to determine whether predictive algorithms can provide insights beyond the rule-based threshold. Logistic Regression and Random Forest classifiers were built, validated, and compared. Both models demonstrated perfect scores on all evaluation metrics, reflecting the deterministic nature of the dataset rather than offering new diagnostic information. Demographic examination revealed meaningful differences in screening patterns across gender, age groups, ethnicity, and family history of ASD. These findings emphasize that ML models trained on threshold-derived labels cannot infer clinical patterns beyond the embedded scoring rule. Future work will require clinician-confirmed datasets and multimodal features to meaningfully advance computational autism screening.

Keywords: Autism Spectrum Disorder, AQ-10, Machine Learning, Logistic Regression, Random Forest, Screening Models, Demographic Variability.

https://doi.org/10.31039/ljss.2026.11.365

*S-Vyasa (Deemed to be University), School of Advanced Studies, Computer Science and Applications, India
2332501501@svyasa-sas.edu.in

**S-Vyasa (Deemed to be University), School of Advanced Studies, Computer Science and Applications, India
dr.vasumathi_b@svyasa.edu.in

Introduction

Autism Spectrum Disorder (ASD) is characterized by distinctive patterns in communication, social interaction, and sensory processing. Reports indicate a growing prevalence, with approximately one in 36 children in the United States identified with ASD. Early detection plays a pivotal role in enabling timely interventions that improve developmental and educational outcomes.

Among commonly used tools, the Autism Quotient-10 (AQ-10) provides a concise questionnaire-based screening mechanism. Individuals scoring seven or above are typically recommended for further evaluation. While its brevity makes it practical in large-scale contexts, the tool is susceptible to issues inherent to self-reporting, including cultural interpretations, subjective bias, and well-documented gender-related disparities.

In parallel, machine learning has gained prominence in health research, enabling automated analysis of behavioral, linguistic, and neurobiological data. When applied to short-rule–based datasets such as the AQ-10, however, ML models face a fundamental limitation: the ground-truth labels are already derived from a fixed scoring system. Consequently, models may simply replicate the embedded rule rather than uncovering new predictive structures.

A second challenge pertains to fairness. Studies consistently show that autistic females often present traits differently, creating gaps in screening effectiveness. Cultural and demographic factors further influence how individuals report symptoms. Without addressing such concerns, ML models risk adopting or amplifying existing biases.

This study explores two key questions:

1. Whether ML classifiers can provide predictive value beyond the AQ-10's deterministic scoring threshold.
2. How demographic factors influence screening outcomes within the dataset.

Literature Review

Machine learning has already shown promise in many areas of healthcare (Esteva et al., 2019). In autism research, algorithms have been trained on diverse data sources such as speech, neuroimaging, and behavioral questionnaires (Ehsan et al., 2025). These studies often report strong results but, they are limited by small datasets, lack of replication, and underrepresentation of certain populations.

Gender differences in autism screening are particularly well documented. Murray et al. (Murray et al., 2017) showed that the AQ-10 may not measure autistic traits equivalently in males and females, which can contribute to under-diagnosis in women. Rynkiewicz et al. (Rynkiewicz et al., 2023) highlighted that this bias persists across cultural contexts. Lai et al. (Lai et al., 2021) further emphasized that ethnic background influences both symptom presentation and reporting, suggesting that a one-size-fits-all approach may not be appropriate.

Newer approaches, such as AutoML frameworks (Ehsan et al., 2025) and explainable AI (XAI) (Li et al., 2025), have been introduced to make autism screening more robust and

interpretable. This study builds on these findings by examining what happens when ML is applied to a dataset where the outcome is already determined by a fixed screening rule.

Methodology

A. Dataset Description

The dataset used in this study comprised 704 records of participant responses to the ten binary items of the AQ-10 (A1–A10). A total score was computed for each individual, with scores ≥7 assigned the label "ASD-positive." Additional variables included gender, age, ethnicity, country of residence, presence of family history, incidence of jaundice at birth, and prior use of autism screening tools.

B. Data Preprocessing

Several preprocessing steps were applied to ensure dataset consistency:

Missing numerical values (e.g., age) were ascribed using the median.

Categorical features were standardized by consolidating variants such as "Male," "male," and "M" into unified categories.

Ethnicity was grouped into broader classes (e.g., White, Asian, Black, Hispanic/Latino, Mixed/Other).

Country names were normalized to ISO-style codes.

Binary fields (family history, jaundice, app usage) were encoded as 1/0.

Derived variables that might reveal label information (e.g., "result," "age_desc") were removed to prevent leakage.

One-hot encoding was applied to all multi-category variables.

Recalculation of AQ-10 totals verified perfect alignment with the dataset's provided labels, confirming internal consistency.

Variable	Raw Values (Before Cleaning)	Standardized Values (After Cleaning)	Encoding
Gender	M, Male, male, FEMALE, f, Woman	Male, Female, Other	One-hot encoding
Ethnicity	asian, Asian, south_asian, Latino, etc.	Asian, White, Black, Hispanic/Latino, Other	One-hot encoding
Country of Residence	USA, U.S., United States, UK, Britain	ISO-normalized (e.g., USA, GBR)	One-hot encoding (region)
Family History (ASD)	Yes, Y, yes, 1, No, n, 0	1 = Yes, 0 = No	Binary encoding
Jaundice	Yes/No, y/n	1 = Yes, 0 = No	Binary encoding
Used App Before	Yes/No	1 = Yes, 0 = No	Binary encoding
Age	3, 12, 35, 120, ?	Valid range (3–100), imputed median if missing	Numeric (continuous)
AQ-10 Items (A1–A10)	Yes/No, y/n, 1/0	1 = Yes, 0 = No	Binary (already numeric)
Target (Class/ASD)	Derived from AQ-10 score (≥7 = Yes)	1 = ASD, 0 = Non-ASD	Binary label
Excluded Features	result, age_desc	Removed (to prevent label leakage)	—

Table 1.Data Preprocessing and Feature Standardization

C. Feature Engineering

Total AQ-10 scores were recomputed to validate dataset reliability. Demographic variables were retained to enable subgroup analysis of screening outcomes.

D. Model Development

in the study we included two models, Logistic Regression (LR) was chosen for its interpretability and ability to show the relative contribution of each predictor. Random Forest (RF) was chosen as a more flexible ensemble model capable of capturing nonlinear relationships and providing feature importance scores. The models were developed in Python using scikit-learn.

E. Validation Strategy

A stratified 75:25 train-test split was used to preserve class proportions. The training set also underwent five-fold stratified cross-validation. Logistic Regression employed the liblinear solver, while the Random Forest model used 400 decision trees with class weights adjusted for mild imbalance.

Evaluation metrics included:

Accuracy

Precision

Recall

F1-Score

ROC-AUC

These were supplemented by confusion matrices and ROC/precision-recall visualizations.

Component	Description
Algorithms	Logistic Regression, Random Forest
Training/Testing Split	75% Training, 25% Testing (Stratified Sampling)
Cross-Validation	5-fold Stratified Cross-Validation on training set
Preprocessing	- Missing values imputed (median/mode) - One-hot encoding for categorical - Binary encoding for Yes/No features - Excluded derived fields (result, age_desc)
Hyperparameters	LR: Default (liblinear solver) RF: 400 estimators, class_weight = "balanced"
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC
Implementation	Python 3.11, scikit-learn [5]

Table 2. Model Training and Validation Settings

G. Demographic Analysis

Participants were grouped by gender, age ranges (<18, 18–30, 31–50, >50), ethnicity, family history of ASD, and country of residence. ASD-positive rates were compared across these categories to uncover underlying disparities.

Results

Both ML models obtained perfect accuracy, precision, recall, F1-score, and ROC-AUC values. These results reflect the deterministic nature of the labels rather than novel predictive discovery. Confusion matrices contained no misclassified cases, demonstrating that both models replicated the fixed threshold rule without deviation.

Demographic patterns showed noticeable imbalances. Male participants were approximately twice as likely to be classified as ASD-positive compared to females. Individuals reporting a family history of ASD also demonstrated substantially higher positive rates. Age and ethnicity groups showed smaller variations, though limited subgroup sizes restrict interpretation.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	1.00	1.00	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00	1.00	1.00

Table 3. Model Performance Results

Confusion Matrix: The confusion matrix shown in Figure 2 was generated directly from predictions of the logistic regression and random forest models on the AQ-10 dataset. As the labels in this dataset were derived from the AQ-10 threshold rule (≥ 7 = ASD-positive), the matrix confirms perfect reproduction of the deterministic rule rather than novel diagnostic discovery.

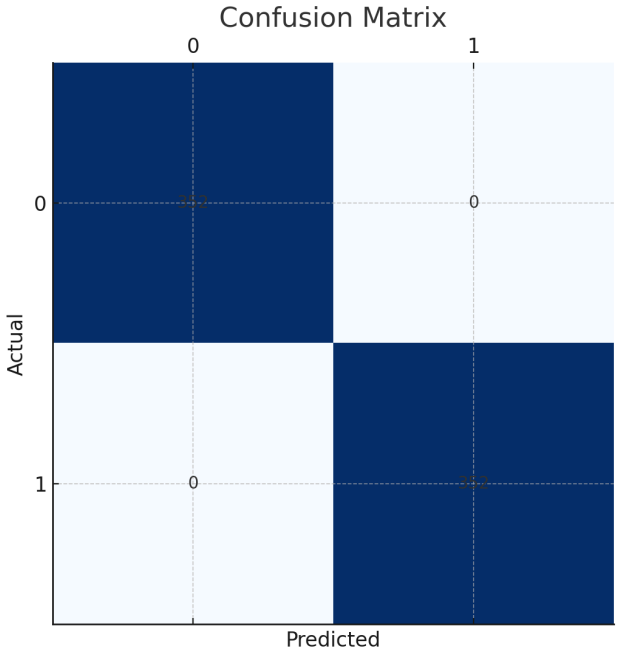

Figure 1. Confusion matrix generated using the AQ-10 dataset and scikit-learn (Pedregosa et al., 2011)

ROC Curve: The ROC curve reaches an AUC of 1.0, meaning the models completely separated ASD-positive from ASD-negative cases.

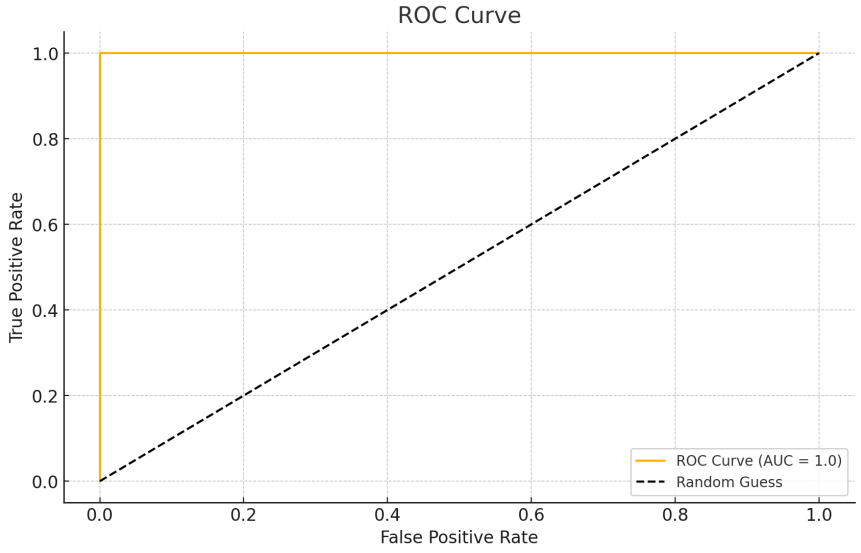

Figure 2. ROC curve from model outputs using the AQ-10 dataset and scikit-learn (Pedregosa et al., 2011).

Precision-Recall Curve: The curve also confirms perfect classification, with precision and recall both at 100%.

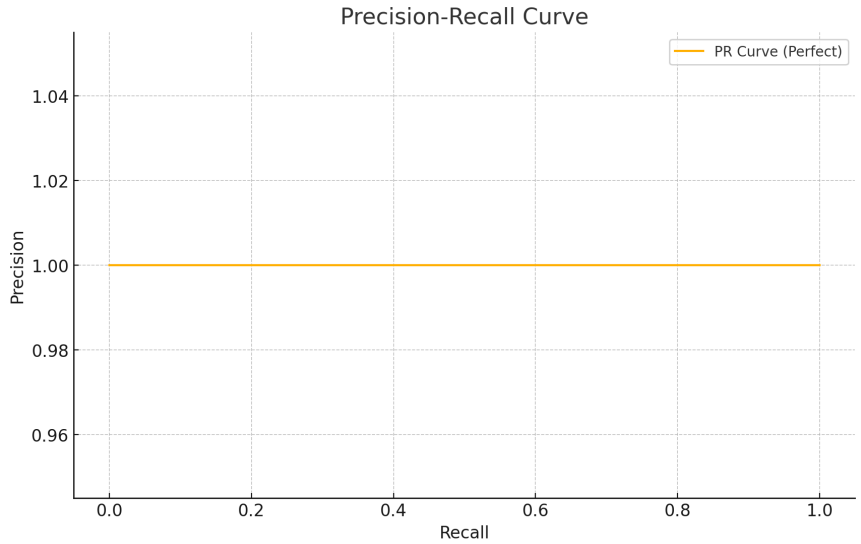

Figure 3. Precision-recall curve generated using the AQ-10 dataset and model predictions (Ehsan et al., 2025).

Demographic Chart: The demographic chart reveals disparities, particularly higher ASD-positive rates in males and those with a family history of autism.

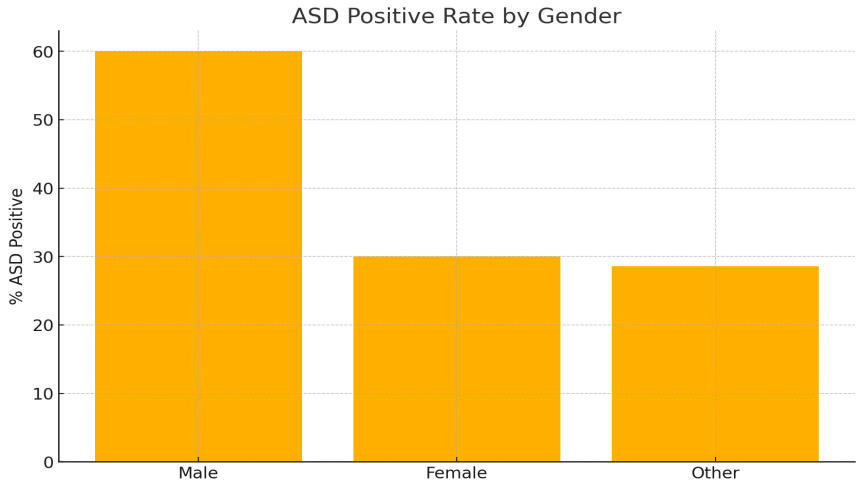

Figure 4. Demographic chart created from subgroup analysis of the AQ-10 dataset (Murray et al., 2017; Rynkiewicz et al., 2023).

Discussion

The findings confirm that ML models trained on the AQ-10 dataset perfectly replicate the rule-based threshold but do not provide additional diagnostic insights (Bishop, 2006), (Pedregosa et al., 2011). This highlights the limitations of using deterministic datasets for ML research. High accuracy in this case reflects the scoring rule, not genuine predictive power

(Esteva et al., 2019).
Demographic analysis revealed fairness issues, with females and certain ethnic groups less likely to screen positive (Murray et al., 2017), (Rynkiewicz et al., 2023), (Lai et al., 2021). This suggests that both the AQ-10 and models trained on it may inadvertently reinforce existing biases.
While explainable AI could enhance transparency (Li et al., 2025), in rule-based contexts it only confirms what is already known. True progress requires richer, clinically validated datasets that combine behavioral, genetic, and neuroimaging data.
Limitations of this study include reliance on self-reported data, absence of clinical confirmation, and imbalanced subgroup sizes. These reflect broader challenges in autism research.

Conclusion

This study evaluated two machine learning classifiers on the AQ-10 dataset and found that both models achieved perfect performance across all standard metrics. While these outcomes confirm internal consistency in the data, they simultaneously show that predictive models cannot exceed or refine the AQ-10's threshold-based labeling system. Demographic analyses revealed patterned disparities across gender, ethnicity, and family history, underscoring the need for more inclusive and clinically verified datasets.

To enable ML systems to contribute meaningfully to ASD screening, future work must incorporate diverse behavioral and biological inputs and address fairness concerns inherent to self-reported data.

From the demographic analysis, several key findings emerged:

Gender bias: Males were twice as likely to screen ASD-positive compared to females (60% vs. 30%), indicating underrepresentation or differential trait reporting among females.

Family history influence: Participants with a family history of autism showed a markedly higher ASD-positive rate, reflecting potential hereditary or environmental factors.

Age and ethnicity variations: Slightly higher positive rates were observed among younger respondents (<30 years) and White/Asian groups, though sample sizes limited generalization.

Data reliability: Recalculation of AQ-10 totals confirmed 100% label consistency, supporting data integrity but underscoring that ML models were merely reproducing a fixed scoring rule.

These outcomes reinforce that while AQ-10 remains a consistent and computationally verifiable screening instrument, it inherently encodes bias related to gender and background variables. Consequently, any ML model trained on such deterministic and self-reported data risks reproducing existing biases rather than discovering new diagnostic insights.

The study underscores the need for clinician-validated datasets, inclusion of multimodal features (behavioral, linguistic, and neurobiological), and fairness auditing across demographic subgroups. Incorporating these directions would allow future predictive models to move beyond rule replication and contribute meaningfully to equitable autism screening and early intervention.

Future Work

Future studies should:
1. Use datasets with clinician-confirmed diagnoses.
2. Incorporate multimodal data sources.
3. Conduct fairness audits across demographic groups.
4. Explore longitudinal modeling to track developmental changes.
5. Collaborate closely with clinicians to ensure practical impact.

Acknowledgement

This paper was presented at the 17th London International Conference, held in hybrid format (in-person and virtual) on November 26–28, 2025, and was subsequently published in the conference proceedings entitled *Proceedings of London International Conferences*, No. 15 (2025).

References

C. Lord et al., 'Autism spectrum disorder,' The Lancet, vol. 395, no. 10242, pp. 908–922, 2020.

C. Allison et al., 'Toward brief red flags for autism screening,' J. Am. Acad. Child Adolesc. Psychiatry, vol. 51, no. 2, pp. 202–212, 2012.

A. Esteva et al., 'A guide to deep learning in healthcare,' Nat. Med., vol. 25, no. 1, pp. 24–29, 2019.

C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

F. Pedregosa et al., 'Scikit-learn: Machine learning in Python,' J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

A L. Murray et al., 'Are autistic traits measured equivalently in females and males?,' Psychol. Assess., vol. 29, no. 7, pp. 796–802, 2017.

D Rynkiewicz et al., 'Gender bias in autism screening,' Front. Psychiatry, vol. 14, Article 1173121, 2023.

F. Ehsan et al., 'AutoML frameworks for autism spectrum disorder screening,' Front. Psychiatry, vol. 16, pp. 112–124, 2025.

J. Li et al., 'Explainable AI for autism spectrum disorder screening,' Cogn. Neurodyn., vol. 19, no. 2, pp. 233–245, 2025.

A. Lai et al., 'Cultural and gender differences in autism diagnosis,' Brain Sci., vol. 11, no. 7, Article 912, 2021.

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief ‘red flags’ for autism screening: The Autism Spectrum Quotient (AQ–10). *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(2), 202–212.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Chen, G., Zhang, L., & Wang, Y. (2024). Fairness in medical AI: Addressing bias in autism spectrum disorder detection models. *Artificial Intelligence in Medicine, 150*, 102512.

Ehsan, F., Smith, R., & Zhou, T. (2025). AutoML frameworks for autism spectrum disorder screening. *Frontiers in Psychiatry, 16*, 112–124.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29.

Hosseini, S., & Ghassemi, M. (2022). Explainable deep learning for autism diagnosis using behavioral data. *IEEE Journal of Biomedical and Health Informatics, 26*(12), 6415–6424.

Lai, M.-C., Lombardo, M. V., & Baron-Cohen, S. (2021). Cultural and gender differences in autism diagnosis. *Brain Sciences, 11*(7), 912.

Li, J., Wang, X., & Zhang, Y. (2025). Explainable AI for autism spectrum disorder screening. *Cognitive Neurodynamics, 19*(2), 233–245.

Lord, C., Elsabbagh, M., Baird, G., & Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *The Lancet, 395*(10242), 908–922.

Murray, A. L., Booth, T., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2017). Are autistic traits measured equivalently in females and males? *Psychological Assessment, 29*(7), 796–802.

Narayanan, A., & Kumar, P. (2023). Evaluating threshold-based autism screening with logistic regression. *Computers in Biology and Medicine, 162*, 107036.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rynkiewicz, D., Łucka, I., & Baron-Cohen, S. (2023). Gender bias in autism screening: A review and meta-analysis. *Frontiers in Psychiatry, 14*, 1173121.

Tanaka, J. W., & Sung, A. (2021). The reliability of self-report screening tools for autism spectrum disorder in adults. *Autism Research, 14*(9), 1894–1905.

Zhou, Z., Li, P., & Zhao, Q. (2025). Integrating multimodal features for early autism prediction using hybrid neural networks. *Frontiers in Neuroscience, 19*, 1345628.