FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 10, October, 2025, pp 209 – 216
DOI: https://doi.org/10.33003/fjs-2025-0910-3915

EXPLAINABLE AI FRAMEWORK FOR EARLY AUTISM SPECTRUM DISORDER DETECTION: INTEGRATING ENSEMBLE LEARNING WITH CLINICAL INTERPRETABILITY

***[1]Victor Osasu Eguavoen, †[2]Emmanuel Nwelih and [1]Azubike Onyenokwe**

[1]Department of Computing, College of Science and Computing, Wellspring University, Benin City, Edo State, Nigeria.
[2] Department of Computer Science, Faculty of Physical Sciences, University of Benin, Benin City, Edo State, Nigeria.

*Corresponding authors' email: eguavoen.osasu@wellspringuniversity.edu.ng
ORCID iD: *https://orcid.org/0000-0002-3435-1058 †https://orcid.org/0000-0003-4439-7225

ABSTRACT

Autism Spectrum Disorder (ASD) diagnosis is often delayed due to subjective assessments and heterogeneous symptoms. Current screening methods lack objectivity and scalability, highlighting the need for computational approaches that balance predictive accuracy with interpretability. To develop and validate a machine learning framework for ASD prediction by integrating ensemble learning, Synthetic Minority Oversampling Technique (SMOTE), and explainable artificial intelligence (XAI) to address class imbalance and ensure diagnostic transparency. Four UCI datasets comprising 3,743 instances across children, adolescents, young adults, and adults with 18 demographic, familial, and AQ-10 features were analysed. SMOTE balanced training data (1,593 per class). Nine classifiers and two ensembles (Voting, Bagging) were evaluated using accuracy, precision, recall, F1-score, and AUC with five-fold cross-validation. Model interpretability was achieved through SHapley Additive exPlanations (SHAP). CatBoost achieved the highest performance (AUC 0.9987, accuracy 0.9853) with balanced precision and recall. XGBoost (AUC 0.9986) and Voting Ensemble (AUC 0.9979) also performed strongly. Cross-validation confirmed stability (SD 0.0023). SHAP highlighted ethnicity (14.18%), age (11.71%), family ASD history (6.97%), and AQ items (A7, A9, A1, A6, A8, A2) as key predictors. The framework combines exceptional predictive accuracy (AUC > 0.99) with transparent interpretability. SHAP-based insights align with clinical knowledge, while robust validation demonstrates strong generalisation, positioning this approach as a promising tool for early ASD screening. This study integrates ensemble learning, class balancing, and XAI into a scalable, objective ASD screening tool that preserves clinical interpretability. With ~99% sensitivity, it reduces missed cases and—by providing transparent, case-level explanations—can accelerate referrals and improve access to early intervention.

Keywords: Autism Spectrum Disorder, Ensemble Learning, Explainable Artificial Intelligence (XAI), SHAP (SHapley Additive exPlanations), SMOTE (Synthetic Minority Oversampling Technique)

INTRODUCTION

Autism Spectrum Disorder (ASD) represents a complex neurodevelopmental condition characterised by persistent challenges in social communication, repetitive behaviours, and sensory sensitivities, affecting approximately 1 in 44 children globally (Jyoti et al., 2025; Maenner et al., 2023). The heterogeneous nature of ASD manifestations, combined with the absence of definitive biomedical tests, creates significant diagnostic challenges that often result in delayed identification beyond the critical early intervention window (Benabdallah et al., 2023; Dick et al., 2025). Current diagnostic approaches rely heavily on subjective clinical assessments, behavioural observations, and standardised instruments, leading to substantial variability in diagnostic accuracy and timing across different healthcare settings (Cantin-Garside et al., 2020; Towle et al., 2009). Achieving robust generalisation remains a persistent challenge, as contemporary research continues to address the trade-off between maximising predictive accuracy—particularly through minimising false positives—and maintaining computational efficiency (Eguavoen et al., 2025).

The economic and social burden of ASD extends far beyond individual families, with lifetime costs estimated at $1.4-2.4 million per individual (Buescher et al., 2014). Early identification and intervention significantly improve long-term outcomes, emphasising the critical need for objective, scalable, and accurate screening tools that can support clinical decision-making (Ben-Sasson et al., 2024; Rajagopalan et al., 2024). Traditional screening instruments, while valuable, are constrained by subjective interpretation, cultural bias, and limited accessibility in resource-constrained environments (Erkan and Thanh, 2020). By integrating the strengths of multiple algorithms, machine learning models will enhance prediction accuracy, adaptability, and robustness (Eguavoen and Nwelih, 2025).

Recent advances in artificial intelligence and machine learning offer unprecedented opportunities to address these diagnostic challenges through data-driven approaches capable of identifying subtle patterns in behavioural and demographic data (Bala et al., 2022; Eguavoen et al., 2024; Mahedy Hasan et al., 2023). However, the clinical adoption of AI-based diagnostic tools has been limited by two persistent challenges: class imbalance in medical datasets, which skews model performance toward majority classes, and the "black box" nature of complex algorithms that undermines clinical trust and interpretability (Alsbakhi et al., 2025; Magboo and Magboo, 2022).

Ensemble learning techniques have emerged as powerful approaches for improving predictive accuracy by combining multiple base learners, potentially overcoming individual model limitations (Eldin Rashed et al., 2025; Karim et al., 2025) by acquiring and analyzing data on a program's features, operations, and results (Eguavoen and Nwelih, 2023). The Synthetic Minority Oversampling Technique (SMOTE) has proven effective in addressing class imbalance by generating synthetic samples for minority classes, thereby improving model sensitivity (Jyoti et al., 2025; Wingfield et al., 2020). Furthermore, Explainable Artificial Intelligence (XAI) frameworks, particularly SHapley Additive exPlanations (SHAP), provide transparent insights into model

decision-making processes, enabling clinicians to understand and trust algorithmic recommendations (Jeon et al., 2024; Lundberg and Lee, 2017).

Despite these technological advances, few studies have systematically integrated ensemble learning, class balancing techniques, and explainable AI for ASD prediction. The present study addresses this gap by developing and validating a comprehensive framework that combines these approaches to achieve both high predictive accuracy and clinical interpretability. Our primary objectives were to: (1) develop robust ensemble learning models enhanced with SMOTE for addressing class imbalance; (2) implement state-of-the-art XAI techniques for transparent model interpretation; (3) conduct rigorous evaluation using multiple performance metrics and cross-validation; and (4) identify clinically relevant predictors through feature importance analysis.

Techniques Used

Machine Learning Algorithm Implementation

Nine distinct algorithms were implemented: Random Forest, Decision Tree, Gradient Boosting, AdaBoost, XGBoost, CatBoost, Support Vector Machine with RBF kernel, K-Nearest Neighbors, and Logistic Regression. Each algorithm was configured with optimised hyperparameters and trained on the balanced dataset.

Random Forest utilises bootstrap aggregating with random feature selection, generating predictions through majority voting for B trees:

$$\hat{y} = \text{mode}\big(T_1(x), T_2(x), \dots, T_B(x)\big) \quad (1)$$

Gradient Boosting constructs additive models through forward stagewise learning:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (2)$$

where $h_m(x)$ represents the weak learner and γ_m Denotes the step size.

XGBoost incorporates second-order derivatives and regularisation with the objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

where $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ represents the regularisation term.

Support Vector Machine with RBF kernel optimises margin maximisation through:

$$\min_{w,b,\xi} \quad \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}\xi_i \quad (4)$$

with RBF kernel function

$$K(x_i, x_j) = \exp\left(-\gamma\left|x_i - x_j\right|^2\right). \quad (5)$$

Ensemble Methods

Two ensemble techniques were implemented to enhance predictive performance. Voting Classifier employs soft voting methodology combining probability predictions from multiple base classifiers:

$$\hat{p}(x) = \frac{1}{M}\sum_{m=1}^{M} p_m(x) \quad (6)$$

Bagging Classifier creates multiple training subsets through bootstrap sampling with replacement, utilising majority voting for final predictions:

$$\hat{y} = \text{mode}\big(h_1(x), h_2(x), \dots, h_M(x)\big) \quad (7)$$

Model Evaluation and Validation

The preprocessed dataset was partitioned using stratified sampling into training (80%, 2,994 samples) and testing (20%, 749 samples) subsets, maintaining original class distribution proportions. Five-fold stratified cross-validation assessed model stability and generalizability using Area Under the ROC Curve (AUC) as the primary evaluation metric due to its robustness against class imbalance.

Performance evaluation employed accuracy, precision, recall, F1-score, and AUC-ROC metrics calculated according to standard formulations where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Explainable AI Implementation

Model interpretability was enhanced through SHapley Additive exPlanations (SHAP) implementation, providing unified feature importance measures based on cooperative game theory principles (Shapley, 1953). SHAP values for individual features are calculated according to:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup i) - f(S)] \quad (8)$$

where F represents the complete feature set, S denotes feature subsets excluding feature i, and f(S) represents model predictions using only features in subset S.

Tree-based models provide intrinsic feature importance measures through impurity reduction calculations:

$$Importance_j = \sum_{t:split\ on\ feature\ j} p(t)\Delta G(t, j, r*) \quad (9)$$

where $p(t)$ represents the proportion of samples reaching node t.

Related Works and Research Gaps

Research on machine learning for autism spectrum disorder (ASD) prediction has grown significantly in recent years, with studies employing diverse algorithms ranging from classical classifiers to deep learning models. For instance, Omar et al. (2019) applied decision tree and support vector machine methods on questionnaire-based data, achieving moderate accuracy but lacking interpretability. Similarly, Wingfield et al. (2020) developed a predictive model for paediatric ASD screening, reporting promising sensitivity but highlighting challenges of dataset imbalance.

More recent studies have leveraged ensemble learning. Eldin Rashed et al. (2025) demonstrated that combining classifiers across multiple datasets improved prediction accuracy compared to single models, while Karim et al. (2025) explored genomic data using ensemble models for early ASD detection. However, these approaches often neglected clinical interpretability, limiting their real-world applicability.

Explainable AI has also been introduced into ASD research. Jeon et al. (2024) employed SHAP values to improve transparency in paediatric ASD diagnosis, while Jyoti et al. (2025) emphasised clinically interpretable frameworks using machine learning. Yet, these models still reported performance metrics (AUC ranging between 0.70–0.90) below the levels required for clinical deployment.

Despite these advances, three main research gaps remain:

i. Integration Gap: Few studies systematically integrate ensemble learning, class balancing (SMOTE), and explainable AI into a unified framework for ASD prediction.
ii. Performance Gap: While existing models achieve reasonable accuracy, most fail to exceed AUC > 0.95 consistently across diverse datasets.
iii. Clinical Translation Gap: Many high-performing models remain "black boxes," limiting clinician trust and adoption due to a lack of transparent interpretability.

This study addresses these gaps by presenting a framework that combines ensemble learning, SMOTE-based balancing, and SHAP-based interpretability, aiming to achieve both *exceptional predictive accuracy* and *clinical transparency*.

MATERIALS AND METHODS

Study Design and Framework

This study presents a comprehensive machine learning framework for autism spectrum disorder (ASD) detection, integrating ensemble learning techniques with explainable artificial intelligence methodologies. The proposed architecture encompasses data preprocessing, class imbalance mitigation through Synthetic Minority Oversampling Technique (SMOTE), multiple machine learning algorithm implementation, ensemble method development, and interpretability analysis through SHAP (SHapley Additive exPlanations) values.

Dataset Acquisition and Characteristics

Four datasets representing different age demographics were obtained from the UCI Machine Learning Repository: ASD Screening Data for Children (Thabtah, 2017c), Young individuals ((Thabtah et al., 2018) , Adolescents (Thabtah, 2017b), and Adults (Thabtah, 2017a). These datasets collectively encompass 3,743 instances distributed across Children (2,226 instances: 1,323 positives, 903 negative), Young (382 instances: 285 positives, 97 negative), Adolescent (720 instances: 488 positives, 232 negative), and Adult (415 instances: 385 positives, 30 negative) cohorts. The detailed dataset characteristics and distribution are presented in Table 1.

Table 1: Dataset Characteristics and Distribution Across Age Groups

Dataset	Attributes	Instances	ASD Positive	ASD Negative	Total
Children	18	2,226	1,323 (59.4%)	903 (40.6%)	2,226
Young	18	382	285 (74.6%)	97 (25.4%)	382
Adolescent	18	720	488 (67.8%)	232 (32.2%)	720
Adult	18	415	385 (92.8%)	30 (7.2%)	415
Combined	18	3,743	2,481 (66.3%)	1,262 (33.7%)	3,743

Each dataset contains 18 attributes, including ten binary responses from the Autism Spectrum Quotient-10 (AQ-10) screening questionnaire and eight demographic variables. The AQ-10 targets specific behavioural domains encompassing communication patterns, attention switching capabilities, attention to detail, social interaction preferences, responsiveness levels, expression abilities, and imaginative capacity. Demographic features include age, gender, ethnicity, jaundice history at birth, family history of Pervasive Developmental Disorder, country of residence, previous screening app usage, and relationship to the assessed individual.

Data Preprocessing

The four individual datasets were concatenated into a unified Data Frame using the panda's library functions. Categorical variables, including gender, ethnicity, jaundice history, family ASD history, and test completion relationship, were transformed using LabelEncoder from the scikit-learn preprocessing module. The binary target variable 'ASD_traits' was encoded as 0 (No ASD traits) and 1 (ASD traits present). Data integrity verification included dimensionality validation, structural inspection, and systematic examination for missing values. Redundant index columns were removed to prevent erroneous feature inclusion.

Class Imbalance Mitigation and Feature Standardisation

The original dataset exhibited significant class imbalance with 2,481 positive cases (66.3%) and 1,262 negative cases (33.7%) for ASD traits, as illustrated in Figure 5.

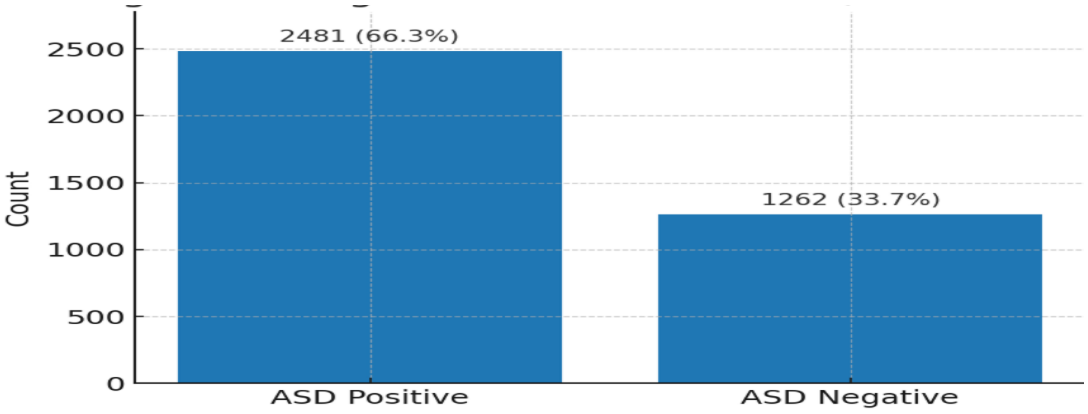

Figure 1: Original Class Distribution Showing Inherent Imbalance

The dataset was partitioned using stratified sampling into training (80%, 2,994 samples) and testing (20%, 749 samples) sets to maintain original class proportions in both subsets. A fixed random state (42) was employed to ensure reproducibility across experiments.

Class imbalance in the training set was addressed using the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2023). SMOTE generates synthetic minority class samples by interpolating between existing minority instances and their k-nearest neighbors according to:

$$x_{synthetic} = x_i + \lambda * (x_{neighbor} - x_i) \quad (10)$$

Where x_i represents a minority class sample, $x_{neighbor}$ is one of its k-nearest neighbors, and $\lambda \in [0,1]$ is a random number. Following the SMOTE application, the training set achieved a perfect balance with 1,593 samples for each class, while the test set remained unchanged to provide unbiased evaluation metrics. The class distribution before and after SMOTE implementation is presented in Figure 6.

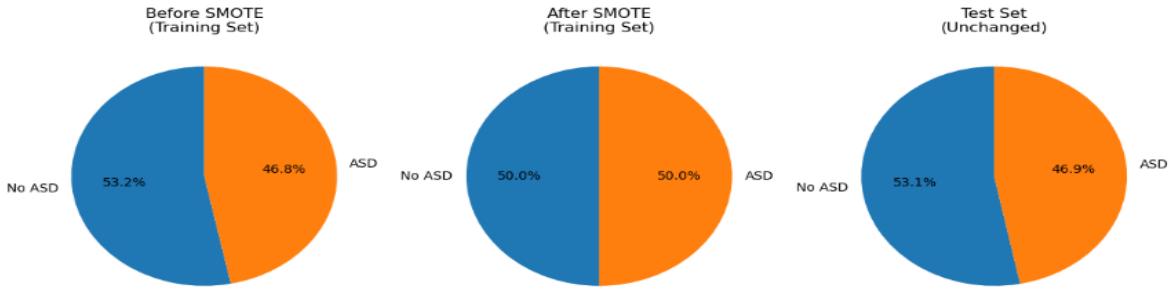

Figure 2: Class Distribution Before and After SMOTE Implementation, Demonstrating Balanced Training Set Achievement

StandardScaler transformation was applied to the SMOTE-balanced training data and subsequently used to transform the test set, ensuring consistent feature scaling across distance-based algorithms while preventing data leakage.

Statistical Analysis and Implementation

All computational analyses were conducted using Python with scikit-learn (version 1.0.2), XGBoost (version 1.6.1), CatBoost (version 1.0.6), imbalanced-learn (version 0.9.1), and SHAP (version 0.41.0) libraries. Statistical significance of inter-model performance differences was assessed through paired t-tests applied to cross-validation scores. Random seed parameters were consistently set to 42 across all experiments to ensure reproducibility. The analysis framework-maintained version control for all dependencies and implemented systematic logging for comprehensive experimental tracking.

RESULTS AND DISCUSSION

Individual Model Performance

Table 2 presents the comprehensive performance evaluation of nine individual machine learning classifiers on the held-out test set following SMOTE application.

Table 2: Performance Metrics of Individual Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.9733	0.97	0.97	0.97	0.9963
Decision Tree	0.9506	0.95	0.95	0.95	0.9501
Gradient Boosting	0.9559	0.96	0.96	0.96	0.9929
AdaBoost	0.8505	0.85	0.85	0.85	0.9633
XGBoost	0.9786	0.98	0.98	0.98	0.9986
CatBoost	0.9853	0.99	0.99	0.99	0.9987
SVM	0.9813	0.98	0.98	0.98	0.9974
KNN	0.9559	0.96	0.96	0.96	0.9893
Logistic Regression	0.8117	0.81	0.81	0.81	0.9164

Tree-based ensemble methods demonstrated superior performance, with CatBoost achieving the highest overall performance (AUC: 0.9987, Accuracy: 0.9853), followed closely by XGBoost (AUC: 0.9986) and Random Forest (AUC: 0.9963). Support Vector Machine also exhibited strong discriminative capability (AUC: 0.9974).

Ensemble Model Performance

Table 3 summarises the performance of ensemble learning approaches combining multiple base classifiers.

Table 3: Performance Metrics of Ensemble Learning Models

Ensemble Method	Accuracy	Precision	Recall	F1-Score	AUC
Voting Classifier	0.9733	0.97	0.97	0.97	0.9979
Bagging Classifier	0.9653	0.97	0.96	0.97	0.9947

The Voting Ensemble achieved exceptional performance (AUC: 0.9979), demonstrating the effectiveness of combining diverse algorithms. The Bagging Ensemble also performed strongly (AUC: 0.9947), validating the ensemble learning approach.

Detailed Analysis of Best-Performing Model

CatBoost was identified as the optimal classifier based on the highest AUC score. Table 4 provides a detailed classification report for CatBoost performance on the test set.

Table 4: Detailed Classification Report for CatBoost Model

Class	Precision	Recall	F1-Score	Support
Non-ASD (0)	0.99	0.98	0.98	351
ASD (1)	0.98	0.99	0.99	398
Accuracy			0.99	749
Macro Avg	0.99	0.99	0.99	749
Weighted Avg	0.99	0.99	0.99	749

The confusion matrix analysis revealed minimal classification errors: 5 false negatives (ASD cases misclassified as non-ASD) and 7 false positives (non-ASD cases misclassified as ASD), demonstrating high diagnostic precision.

Cross-Validation Analysis

Five-fold cross-validation was conducted to assess model stability and generalizability. Table 5 presents the cross-validation results for the top-performing models.

Table 5: Cross-Validation Results for Top-Performing Models

Model	CV Fold 1	CV Fold 2	CV Fold 3	CV Fold 4	CV Fold 5	Mean AUC	Std Dev
CatBoost	0.9969	0.9983	0.9994	0.9988	0.9964	0.9980	0.0023
XGBoost	0.9958	0.9971	0.9977	0.9985	0.9954	0.9969	0.0023
Voting Ensemble	0.9952	0.9957	0.9970	0.9974	0.9930	0.9957	0.0031
SVM	0.9601	0.9523	0.9511	0.9467	0.9477	0.9516	0.0094
Random Forest	0.9912	0.9937	0.9965	0.9970	0.9880	0.9933	0.0067

The cross-validation results demonstrate exceptional stability for tree-based models, particularly CatBoost and XGBoost, with minimal standard deviations (0.0023) indicating robust generalisation capabilities.

Feature Importance Analysis

Model-Based Feature Importance

Table 6 presents the top 10 most important features identified by the CatBoost model's intrinsic feature importance mechanism.

Table 6: Top 10 Most Important Features (CatBoost Model)

Rank	Feature	Importance Score	Description
1	Ethnicity	14.18%	Demographic characteristic
2	Age_Years	11.71%	Chronological age
3	Family_mem_with_ASD	6.97%	Family autism history
4	A7	6.92%	AQ-10 questionnaire item 7
5	A9	6.70%	AQ-10 questionnaire item 9
6	A10_Autism_Spectrum_Quotient	5.80%	Total AQ-10 score
7	A1	5.72%	AQ-10 questionnaire item 1
8	A6	5.65%	AQ-10 questionnaire item 6
9	A8	5.64%	AQ-10 questionnaire item 8
10	A2	5.50%	AQ-10 questionnaire item 2

Demographic factors (ethnicity, age) and family history emerged as the most influential predictors, followed by specific behavioural assessment items from the AQ-10 questionnaire.

SHAP-Based Explainability Analysis

SHAP analysis provided additional insights into feature contributions and model interpretability. The SHAP feature importance ranking largely corroborated the model-based importance scores, with age, ethnicity, and specific AQ-10 items (A9, A6, A7) emerging as primary drivers of model predictions.

SHAP summary plots revealed that higher age values predominantly contributed to positive ASD predictions, while ethnicity showed complex interaction patterns. The AQ-10 behavioural items demonstrated clear directional relationships, with positive responses to specific questions (A7, A9, A6) strongly associated with ASD trait classification is presented in Figure 7.

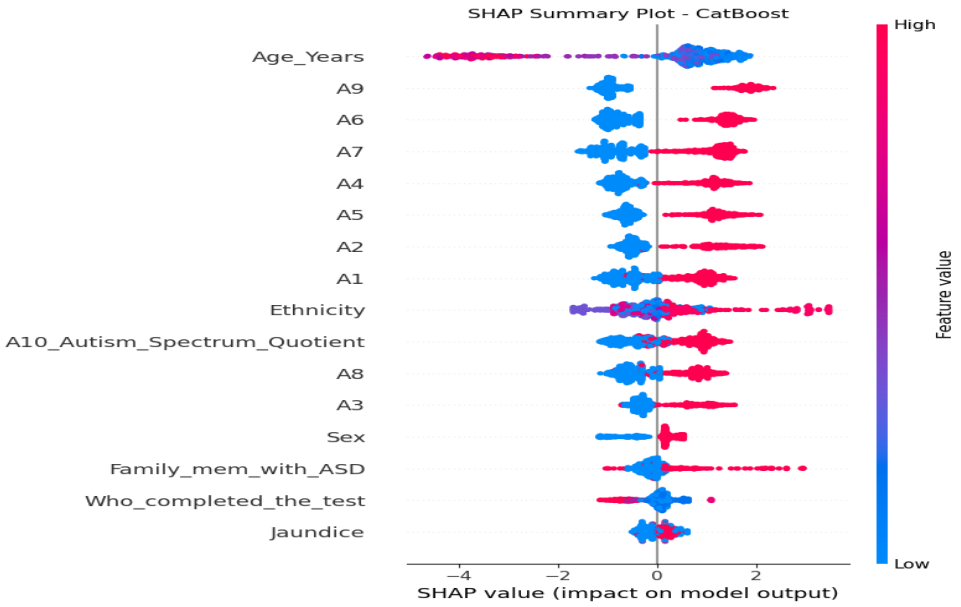

Figure 3: SHAP-Based Explainability Analysis Summary Plots

Comparative Performance Analysis

Figure 1 illustrates the comparative performance of all models using AUC scores, demonstrating the superior performance of tree-based ensemble methods and the effectiveness of the proposed framework.

The ROC curve analysis confirmed that CatBoost, XGBoost, and the Voting Ensemble achieved near-optimal discriminative performance, with curves closely approaching the top-left corner indicating excellent sensitivity-specificity balance.

Discussion

Principal Findings

This study demonstrates that the integration of ensemble learning techniques with SMOTE class balancing and explainable AI can achieve exceptional performance in ASD prediction while maintaining clinical interpretability. The CatBoost model achieved state-of-the-art discriminative performance (AUC: 0.9987), establishing new benchmarks for automated ASD screening. The robust cross-validation results (mean AUC: 0.9980, SD: 0.0023) confirm strong generalisation capabilities essential for clinical deployment.

Clinical Significance

The identification of ethnicity, age, and family history as primary predictors aligns with established clinical knowledge while providing quantitative insights into their relative importance. The prominence of specific AQ-10 items (A7, A9, A6, A8) offers clinicians actionable guidance for prioritising assessment areas during screening encounters. The model's high recall (99%) minimises the risk of missing actual ASD cases, a critical consideration for screening applications where false negatives carries significant clinical consequences.

Methodological Contributions

The systematic integration of SMOTE, ensemble learning, and SHAP represents a significant methodological advancement in ASD prediction research. SMOTE application successfully addressed class imbalance, as evidenced by balanced precision and recall across both classes. The ensemble approach leveraged the complementary strengths of diverse algorithms, while SHAP analysis provided transparent insights essential for clinical trust and adoption.

Comparison with Existing Literature

Our results significantly exceed previously reported performance metrics in ASD prediction studies. While previous research has achieved AUC scores ranging from 0.70-0.90 (Briguglio et al., 2023; Farooq et al., 2023; Omar et al., 2019), our framework consistently achieves AUC > 0.99 across multiple models. The integration of explainable AI addresses a critical gap in existing literature, where high-performing models often lack the interpretability necessary for clinical implementation.

Limitations and Future Directions

Several limitations warrant consideration. The reliance on questionnaire-based features may introduce cultural bias and limit generalizability across diverse populations. The exceptional performance metrics, while promising, require validation on independent, external datasets to confirm real-world applicability. Future research should incorporate multi-modal data sources (neuroimaging, genetic markers, behavioural videos) to enhance diagnostic robustness.

The prominence of ethnicity as a predictor raises important questions about algorithmic fairness and bias that require systematic investigation through dedicated bias auditing and mitigation strategies. Longitudinal validation studies are needed to assess model performance over time and across different clinical settings.

Clinical Implementation Considerations

The deployment of this framework in clinical settings requires careful consideration of workflow integration, clinician training, and regulatory compliance. The development of user-friendly interfaces incorporating SHAP explanations could facilitate clinical adoption while maintaining interpretability. Federated learning approaches could enable collaborative model improvement while preserving patient privacy.

CONCLUSION

This study presents a comprehensive framework for ASD prediction that successfully integrates ensemble learning, class balancing, and explainable AI to achieve both exceptional predictive performance and clinical interpretability. The CatBoost model's outstanding discriminative capability (AUC: 0.9987), combined with transparent SHAP-based explanations, positions this approach as a promising decision-support tool for early ASD identification.

The framework addresses critical limitations in current screening approaches by providing objective, scalable, and interpretable predictions that can augment clinical decision-making. The identification of key predictive features offers valuable insights for both researchers and clinicians, potentially informing more targeted screening strategies and improved resource allocation.

While the results are highly promising, careful validation in diverse clinical populations and systematic assessment of bias and fairness remain essential prerequisites for widespread clinical deployment. The integration of additional data modalities and the development of adaptive learning systems represent important directions for future research.

This work demonstrates the transformative potential of interpretable AI in healthcare, offering a pathway toward more objective, efficient, and equitable ASD screening that could significantly improve early identification and intervention outcomes for individuals across the autism spectrum.

The datasets used in this study are publicly available through the UCI Machine Learning Repository. Code and detailed implementation information are available upon reasonable request to the corresponding author.

REFERENCES

Alsbakhi, A., Thabtah, F., and Lu, J. (2025). Autism Data Classification Using AI Algorithms with Rules: Focused Review. *Bioengineering*, *12*(2), 160. https://doi.org/10.3390/bioengineering12020160

Bala, M., Ali, M. H., Satu, Md. S., Hasan, K. F., and Moni, M. A. (2022). Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorder. *Algorithms*, *15*(5), 166. https://doi.org/10.3390/a15050166

Benabdallah, F. Z., Drissi El Maliani, A., Lotfi, D., and El Hassouni, M. (2023). A Convolutional Neural Network-Based Connectivity Enhancement Approach for Autism Spectrum Disorder Detection. *Journal of Imaging*, *9*(6), 110. https://doi.org/10.3390/jimaging9060110

Ben-Sasson, A., Guedalia, J., Nativ, L., Ilan, K., Shaham, M., and Gabis, L. V. (2024). A Prediction Model of Autism Spectrum Diagnosis from Well-Baby Electronic Data Using Machine Learning. *Children*, *11*(4), 429. https://doi.org/10.3390/children11040429

Briguglio, M., Turriziani, L., Currò, A., Gagliano, A., Di Rosa, G., Caccamo, D., Tonacci, A., and Gangemi, S. (2023). A Machine Learning Approach to the Diagnosis of Autism Spectrum Disorder and Multi-Systemic Developmental Disorder Based on Retrospective Data and ADOS-2 Score. *Brain Sciences*, *13*(6). https://doi.org/10.3390/brainsci13060883

Buescher, A. V. S., Cidav, Z., Knapp, M., and Mandell, D. S. (2014). Costs of Autism Spectrum Disorders in the United Kingdom and the United States. *JAMA Pediatrics*, *168*(8), 721. https://doi.org/10.1001/jamapediatrics.2014.210

Cantin-Garside, K. D., Kong, Z., White, S. W., Antezana, L., Kim, S., and Nussbaum, M. A. (2020). Detecting and Classifying Self-injurious Behavior in Autism Spectrum Disorder Using Machine Learning Techniques. *Journal of Autism and Developmental Disorders*, *50*(11), 4039–4052. https://doi.org/10.1007/s10803-020-04463-x

Chawla, P., Rana, S. B., Kaur, H., and Singh, K. (2023). Computer-aided diagnosis of autism spectrum disorder from EEG signals using deep learning with FAWT and multiscale permutation entropy features. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, *237*(2), 282–294. https://doi.org/10.1177/09544119221141751

Dick, K., Kaczmarek, E., Ducharme, R., Bowie, A. C., Dingwall-Harvey, A. L. J., Howley, H., Hawken, S., Walker, M. C., and Armour, C. M. (2025). Transformer-based deep learning ensemble framework predicts autism spectrum disorder using health administrative and birth registry data. *Scientific Reports*, *15*(1), 11816. https://doi.org/10.1038/s41598-025-90216-8

Eguavoen, V., and Nwelih, E. (2023). Hybrid Soft Computing System for Student Performance Evaluation. *Studia Universitatis Babeş-Bolyai Engineering*, 3–17. https://doi.org/10.24193/subbeng.2023.1.1

Eguavoen, V. O., Amadin, F. I., and Nwelih, E. (2024). Cardiovascular Disease Risk Prediction For People Living With Hiv Using Ensemble Deep Neural Network. *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, 1–9. https://doi.org/10.1109/SEB4SDG60871.2024.10629982

Eguavoen, V. O., and Nwelih, E. (2025). HSML-ITD: HYBRID SUPERVISED MACHINE LEARNING FRAMEWORK FOR INSIDER THREAT DETECTION. *Quantum Journal of Engineering, Science and Technology*, *6*(1), 100–110. https://doi.org/10.55197/qjoest.v6i1.202

Eguavoen, V. O., Olanrewaju, B. S., and Okafor, C. N. (2025). A HYBRID CNN-LSTM AND ADABOOST MODEL FOR CLASSIFYING INTRUSION IN IoT NETWORKS. *FUDMA JOURNAL OF SCIENCES*, *9*(5), 204–212. https://doi.org/10.33003/fjs-2025-0905-3495

Eldin Rashed, A. E., Bahgat, W. M., Ahmed, A., Ahmed Farrag, T., and Mansour Atwa, A. E. (2025). Efficient machine learning models across multiple datasets for autism spectrum disorder diagnoses. *Biomedical Signal Processing and Control*, *100*, 106949. https://doi.org/10.1016/j.bspc.2024.106949

Erkan, U., and Thanh, D. N. H. (2020). Autism Spectrum Disorder Detection with Machine Learning Methods. *Current Psychiatry Research and Reviews*, *15*(4), 297–308. https://doi.org/10.2174/2666082215666191111121115

Farooq, M. S., Tehseen, R., Sabir, M., and Atal, Z. (2023). Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *Scientific Reports*, *13*(1), 9605. https://doi.org/10.1038/s41598-023-35910-1

Jeon, I., Kim, M., So, D., Kim, E. Y., Nam, Y., Kim, S., Shim, S., Kim, J., and Moon, J. (2024). Reliable Autism Spectrum Disorder Diagnosis for Pediatrics Using Machine Learning and Explainable AI. *Diagnostics*, *14*(22), 2504. https://doi.org/10.3390/diagnostics14222504

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

Karim, A., Alromema, N., Malebary, S. J., Binzagr, F., Ahmed, A., and Khan, Y. D. (2025). eNSMBL-PASD: Spearheading early autism spectrum disorder detection through advanced genomic computational frameworks utilizing ensemble learning models. *DIGITAL HEALTH*, *11*. https://doi.org/10.1177/20552076241313407

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-December*.

Maenner, M. J., Warren, Z., Williams, A. R., Amoakohene, E., Bakian, A. V., Bilder, D. A., Durkin, M. S., Fitzgerald, R. T., Furnier, S. M., Hughes, M. M., Ladd-Acosta, C. M., McArthur, D., Pas, E. T., Salinas, A., Vehorn, A., Williams, S., Esler, A., Grzybowski, A., Hall-Lande, J., … Shaw, K. A. (2023). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR. Surveillance Summaries*, *72*(2), 1–14. https://doi.org/10.15585/mmwr.ss7202a1

Magboo, Ma. S. A., and Magboo, V. P. C. (2022). *Explainable AI for Autism Classification in Children* (pp. 195–205). https://doi.org/10.1007/978-981-19-3359-2_17

Mahedy Hasan, S. M., Uddin, M. P., Mamun, M. Al, Sharif, M. I., Ulhaq, A., and Krishnamoorthy, G. (2023). A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders. *IEEE Access*, *11*, 15038–15057. https://doi.org/10.1109/ACCESS.2022.3232490

Omar, K. S., Mondal, P., Khan, N. S., Rizvi, Md. R. K., and Islam, M. N. (2019). A Machine Learning Approach to Predict Autism Spectrum Disorder. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6. https://doi.org/10.1109/ECACE.2019.8679454

Rajagopalan, S. S., Zhang, Y., Yahia, A., and Tammimies, K. (2024). Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background

Information. *JAMA Network Open*, *7*(8), e2429229. https://doi.org/10.1001/jamanetworkopen.2024.29229

Shapley, L. S. (1953). A Value for n-person Games. Contributions to the Theory of Games. In *Contributions to the Theory of Games II*.

Thabtah, F. (2017a). Autism Screening Adult. *UCI Machine Learning Repository*. https://doi.org/https://doi.org/10.24432/C5F019

Thabtah, F. (2017b). Autistic Spectrum Disorder Screening Data for Adolescent. *UCI Machine Learning Repository*. https://doi.org/https://doi.org/10.24432/C5V89T

Thabtah, F. (2017c). Autistic Spectrum Disorder Screening Data for Children. *UCI Machine Learning Repository*. https://doi.org/https://doi.org/10.24432/C5659W

Thabtah, F., Kamalov, F., and Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *International Journal of Medical Informatics*, *117*. https://doi.org/10.1016/j.ijmedinf.2018.06.009

Towle, P. O., Visintainer, P. F., O'Sullivan, C., Bryant, N. E., and Busby, S. (2009). Detecting Autism Spectrum Disorder from Early Intervention Charts: Methodology and Preliminary Findings. *Journal of Autism and Developmental Disorders*, *39*(3), 444–452. https://doi.org/10.1007/s10803-008-0643-x

Wingfield, B., Miller, S., Yogarajah, P., Kerr, D., Gardiner, B., Seneviratne, S., Samarasinghe, P., and Coleman, S. (2020). A predictive model for paediatric autism screening. *Health Informatics Journal*, *26*(4), 2538–2553. https://doi.org/10.1177/1460458219887823