

RESEARCH ARTICLE



A Clinical Computing Framework for Early Autism Diagnosis Using Hybrid ML and Explainable AI

Luc-Orlane Jocelyne Houefa Padonou¹ and Jin Hou^{1,*}

¹IPSOM Laboratory, Southwest Jiaotong University, China

Abstract: The early identification of autism spectrum disorder (ASD) is essential for enhancing the developmental process, but the existing conventional approaches, including machine learning, are usually marred by the problem of subjectivity, cultural biases, lack of scalability, and the inability to process tabular questionnaires effectively. Thus, to resolve these challenges, a fairness-aware, explainable clinical computing approach based on a hybrid ensemble learning strategy for the prediction of ASD has been developed. With the help of 3280 caregiver responses to the AQ-10 and Q-CHAT-10 questionnaires, nine predictors based on the DSM-5 classification were identified, employing an automated feature selection approach. Two different hybrid ensemble strategies were implemented, including a parallel strategy based on the Voting algorithm with performance-weighted classification, and a sequential strategy based on the Stacking algorithm with Platt scaling and temperature scaling, employing Random Forest, XGBoost, support vector machine, and logistic regression. Both models were clinically very useful, with 98.0% accuracy and Area Under the Curve (AUC) = 0.997 for the parallel ensemble and 98.8% accuracy with AUC = 0.992 for the sequential ensemble, with very good calibration and fairness across subgroups. The approach ensures transparency, equality, and robustness through the use of SHapley Additive exPlanations and Local Interpretable Model-agnostic Explanations for explainability, as well as fairness audits using AIF360 and Fair Learn. The above findings illustrate the reliability and effectiveness of parallel-sequential hybrid ensemble models for the early diagnosis of ASD, ensuring clinical utility in both resource-constrained and resource-advantaged settings.

Keywords: autism spectrum disorder (ASD), hybrid machine learning, explainable AI (XAI), clinical computing framework, fairness and calibration

1. Introduction

Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder that begins in early childhood and persists into adulthood and affects social interactions, communication, and cognitive processes [1–3]. The behavioral symptoms of ASD include difficulties in interpretation of non-verbal communications, relationship development, emotional turn-taking, delayed or disordered speech, and repetitive or idiosyncratic use of language. Genetic, pathophysiological, and environmental factors also contribute to the intricacy of ASD, which makes it difficult to diagnose in early childhood [4]. This is because it has been observed that early interventions have significantly improved the development, adaptive, and functional abilities in early childhood [5, 6]. Early diagnosis before the age of two has been observed to increase cognitive abilities and social adaptability. Although neuroimaging and resting-state network analysis have been investigated for the diagnosis of ASD, these methods are expensive and resource-intensive and have only moderate predictive accuracy (60–80%), which restricts their application for large-scale screening [7–11].

The behavioral assessment tools, which include the Screening Tool for Autism in Toddlers and the Autism Diagnostic Observation Schedule, are considered to be the gold standard for the diagnosis of ASD [12–14]. However, these tools are associated with various drawbacks, which include the need for observation and parental report, thus making it difficult to scale up and accommodate cultural differences [15–17]. As a result, children with ASD are diagnosed when they are over three years old.

To overcome these drawbacks, a new hybrid ensemble model has been proposed for the diagnosis of ASD in children. This model, which has been developed for the diagnosis of ASD, does not propose a new learning model but incorporates a wide range of machine learning (ML) algorithms, including Random Forest (RF), XGBoost, support vector machine (SVM), and logistic regression (LR), with dual ensemble strategies, probability calibration, and fair AI. Although these approaches are individually known, their application as a single tool for the diagnosis of ASD, as shown in the study, can be considered an improvement over the current state-of-the-art approaches, as it combines a wide range of ML approaches with ensemble strategies, probability calibration, and fair AI.

*Corresponding author: Jin Hou, IPSOM Laboratory, Southwest Jiaotong University, China. Email: jhou@swjtu.edu.cn

2. Literature Review

Shambour et al. [18] proposed an ML-based approach for early diagnosis of ASD in children aged 12–36 months. Class imbalance was handled by using Gaussian Noise Up-Sampling, and MaxAbsScaler was used for normalization. Feature selection was carried out by using Extra Trees (ET), and six ML models were implemented: Extreme Gradient Boosting (XGB), RF, K-Nearest Neighbors (KNN), Naive Bayes (NB), Quadratic Discriminant Analysis (QDA), and SVM. The proposed approach achieved almost perfect results in terms of accuracy, precision, recall, and F1-score, achieving 100% for both XGB and SVM models. SHapley Additive exPlanations (SHAP) was used to improve the interpretability of the proposed approach. Nevertheless, the proposed approach is based on a single dataset and uses up-sampling, making it prone to overfitting.

de Lacy et al. [19] proposed Integrated Evolutionary Learning (IEL) as a methodology for simultaneous feature selection and hyperparameter tuning for ML models. In essence, IEL utilizes evolutionary programming to optimize the performance of an ML model, such as a linear classifier, decision tree (DT), or Artificial Neural Network (ANN), by selecting features and hyperparameters based on fitness scores defined by an information-based criterion. IEL was successfully validated on various biobehavioral datasets, with IEL obtaining at least 95% accuracy, sensitivity, and specificity using ANNs, as well as R^2 values ranging from 45% to 73%. However, the computational requirements of IEL are high, and its applicability to real-world tasks has yet to be explored.

A review article by Alsbakhi et al. [20] reviewed the use of interpretable rule-based classifiers for the diagnosis of ASD based on behavioral features. The issues with the use of interpretable classifiers for the diagnosis of ASD include the absence of adequate and diverse datasets, the heterogeneity of symptoms, and the difficulty of developing interpretable classifiers for the diagnosis of ASD. Rule-based classifiers, such as DTs, were identified as a possible remedy for the diagnosis of ASD, with the need to combine these with current advances in AI techniques for enhanced accuracy and clinician confidence. However, interpretable classifiers for ASD diagnosis remain understudied, with concerns regarding their validity and generalizability.

In Raajitha and Thatimakula [21], a comparative approach for the early detection of ASD using deep learning (DL) models was presented, considering both accuracy and interpretability. Two datasets, Autism Prediction Dataset and ASD Screening Data for Toddlers in Saudi Arabia, were preprocessed, and feature engineering was performed. Classical ML algorithms, such as RF and SVM, were compared with DL algorithms, Neural Networks (NNs), MobileNet, Vision Transformer (ViT), MobileNet + Long Short-Term Memory (LSTM), Graph Neural Network (GNN), along with Gradient Boosting algorithms. The state-of-the-art algorithms were found to perform better, but the interpretability was ensured by explainable AI (XAI) [22]. The Explainable Mental Health Disorders (EMHD) model was proposed, where an ensemble voting classifier was employed, along with feature selection algorithms, such as Mutual Information (MI), Analysis of Variance (ANOVA), and Recursive Feature Elimination (RFE), for the prediction of mental health disorders. SHAP was employed for better interpretability, where the accuracy, precision, recall, and F1-score were found to be 100%. However, the dataset was small, localized, and lacked external validation, calibration, and fairness testing.

Magboo and Magboo [23] tested various ML models such as LR, KNN, NB, SVM, DTs, RF, AdaBoost, XGBoost, and deep neural networks (DNNs) on a publicly available ASD dataset.

SVM, LR, and AdaBoost yielded 99–100% accuracy. Local Interpretable Model-agnostic Explanations (LIME) improved interpretability by associating features with clinical ratings. Drawbacks included the small dataset, possibility of overfitting, and lack of external validity.

Mumenin et al. [24] introduced an efficient ensemble model for ASD diagnosis at all developmental levels based on questionnaires. The model consisted of a stacked ensemble of RF, ET, and CatBoost as base models and ANN as the meta-model. Safe-Level Synthetic Minority Oversampling Technique (SMOTE) handled imbalanced datasets, Principal Component Analysis (PCA) reduced dimensionality, and MI with Pearson correlation helped in feature selection. Accuracy varied between 96.96% (combined dataset) and 99.89% (adolescents). SHAP improved feature interpretability, and Monte Carlo Dropout (MCD) allowed uncertainty estimation. But the model's dependence on up-sampling and dataset-specific optimization makes it less generalizable.

Sandeep et al. [25] proposed the XAI architecture for pediatric ASD classification based on the UCI dataset for ASD. It achieved 95% accuracy by utilizing XAI to recognize important features. Although it is a good approach for pediatric ASD classification, it is based on a single dataset.

Gombar [26] proposed the XGBoost Multimodal Autism Predictor, utilizing artificially designed traits and behavioral datasets for early-stage ASD classification. Although it achieved good results in terms of classification accuracy by utilizing feature selection, data balancing, and incremental hyperparameter tuning, it utilized artificially enhanced datasets.

Alam et al. [27] proposed a DL-based approach for ASD diagnosis by utilizing Convolutional Neural Networks (CNNs) and image processing techniques for unbalanced datasets. Although it achieved 98.9% accuracy, sensitivity, and specificity in terms of ASD diagnosis and 99.9% Area Under the Curve (AUC) by utilizing CNNs and image processing techniques, it utilized XAI to represent the results of facial features in a visually interpretable manner.

Sravva et al. [28] proposed a hybrid ML–DL approach combining behavioral questionnaires (XGBoost) and facial images (MobileNet + Recurrent Neural Network (RNN)). Preprocessing involved MI-based feature selection, Term Frequency–Inverse Document Frequency (TF-IDF) transformation, and convolutional feature extraction. Results on various datasets showed 94.1% accuracy for behavior, 89.64% accuracy for images, and 0.93 F1-score. XAI improved clinical trust, but further validation is required. Omrani et al. [29] designed explainable ML models for identifying facial characteristics in autistic children. Vision Transformer (ViT_Huge_14) demonstrated 92% accuracy. LIME was more interpretable than Randomized Input Sampling for Explanation (RISE). Clinical application should supplement, but not substitute, comprehensive diagnostic approaches, and more diverse data are required.

Gkintoni et al. [30] reviewed AI-based neuroimaging biomarkers for the early diagnosis of ASD; the authors examined 146 papers and 1872 records based on the PRISMA checklist. The accuracy of the biomarkers based on Electroencephalography (EEG) was between 85% and 99%. The studies on infants identified the 9- to 12-month period as a critical time for biomarker identification, while multimodal neuroimaging approaches were found to improve the predictive potential of the biomarkers.

Significant developments in the methodologies used in other medical fields also contribute to ASD diagnostics. The hybrid ensemble approach was used for prediction in heart diseases in Yadav and Pal [31], with 96–100% accuracy achieved with the

combination of sequential learners like AdaBoostM1, XGBoost, and Gradient Boosting, and parallel learners like Decision Stump, J48, and Reduced Error Pruning (REP). The same approach was used for diabetes prediction in Olorunfemi et al. [32] with perfect scores for parallel-sequential integration, preprocessing, and feature selection approaches.

Despite the above developments, ASD diagnostics still face several challenges. The studies conducted are mostly based on a single dataset or synthetic data augmentation approaches [18, 23]. The tools used for explanation, like SHAP and LIME, are not considered in terms of fairness and calibration approaches [22, 25]. The hybrid/multimodal approach also affects

computational efficiency [28]. Neuroimaging approaches are also used for ASD diagnostics, but they are considered to be expensive and not feasible for all [30].

This paper presents an interpretable and fairness-aware clinical computing approach for the early diagnosis of ASD with high predictive performance, scalability, interpretability, and fairness for diverse healthcare settings using AIF360 and Fairlearn for fairness auditing and hybrid ensemble approaches with robust calibration for high predictive performance.

Table 1 summarizes representative studies on ASD detection and related ensemble learning approaches, highlighting their methodologies, key findings, and limitations.

Table 1
Summary of related works on ASD detection and relevant ensemble learning studies

SN	Reference	Methods used	Gap addressed/aim	Results/findings	Limitations
1	[18]	GNUS up-sampling, ET feature selection, XGB, RF, KNN, NB, QDA, SVM; SHAP explainability	Improve toddler ASD screening with imbalance handling and feature optimization	XGB and SVM achieved 100% across metrics	Relied on single dataset; riskover fitting; of no external validation
2	[19]	Integrated Evolutionary Learning with ANNs, DT, linear models	Joint feature selection + hyperparameter tuning	ANNs achieved $\geq 95\%$ accuracy, sensitivity, specificity	High computational cost; limited real-world validation
3	[20]	Rule-based and decision tree classifiers (review)	Emphasize interpretability in ASD diagnosis	Highlighted interpretability benefits for clinicians	Few hybrid models; limited clinical validation
4	[21]	RF, SVM vs deep learning (NN, MobileNet, ViT, MobileNet + LSTM, GNN) + XAI	Comparative DL framework for ASD detection	DL outperformed traditional ML; improved interpretability with XAI	Needs validation on larger, diverse datasets
5	[22]	EMHD: Ensemble voting classifier + MI, ANOVA, RFE + SHAP	Predict mental health challenges in autistic children	Achieved perfect accuracy (1.0) across all metrics	Limited regional dataset; no fairness/calibration
6	[23]	Logistic regression, KNN, NB, SVM, RF, AdaBoost, XGB, DNN + LIME	Apply ML to ASD dataset with explainability	SVM, LR, AdaBoost achieved 99–100% accuracy	Small dataset; risk of overfitting; limited generalizability
7	[24]	Stacked RF, ET, CatBoost + ANN meta-classifier; Safe-Level SMOTE, PCA, MI, Pearson	Scalable ASD diagnosis across age groups	Accuracy 96.9–99.9%; SHAP for interpretability; MCD for uncertainty	Dataset-specific optimizations; limited cross-cultural validation
8	[25]	Explainable ensemble ML (X-AI architecture)	Combine accuracy + interpretability for ASD classification	Achieved 95% accuracy; interpretable key features	Single dataset; needs validation across populations
9	[27]	XGBoost Multimodal Autism Predictor (XMAP); feature selection + synthetic data	Improve ASD classification with multimodal fusion	Strong accuracy, recall, F1, AUC-ROC; interpretable features	Relies on synthetic data; generalizability concerns
10	[28]	CNN on facial images + augmentation + XAI	Facial image-based ASD detection	98.9% accuracy, sensitivity, specificity; AUC 99.9%	Concerns over population diversity (ethnicity, age, imaging conditions)
11	[29]	Hybrid ML–DL: XGB (behavioral) + MobileNet + RNN (images) + XAI	Multimodal ASD detection	94.1% (behavioral), 89.6% (images); F1 = 0.93, AUC >94%	Dataset diversity limited; generalizability issue

(Continued)

Table 1
(Continued)

SN	Reference	Methods used	Gap addressed/aim	Results/findings	Limitations
12	[29]	Hybrid ML-DL: XGB (behavioral) + MobileNet + RNN (images) + XAI	Explainable image-based ASD detection	ViT achieved 92%; LIME offered better interpretability	Small dataset; ethical concerns in face-based screening
13	[30]	Systematic review (AI + EEG, MRI, multimodal neuroimaging)	Identify neuroimaging biomarkers for ASD	Diagnostic accuracy 85–99%; early detection window (9–12 months)	High cost, lack of standardization, limited clinical translation
14	[31]	Hybrid ensemble: RF (Decision Stump, J48, REP), AdaBoostM1, XGB, Gradient Boosting	Heart disease prediction with hybrid ensembles	RF ensemble 96%, XGB 100%	Domain-specific; not ASD; computationally intensive
15	[32]	Hybrid ensemble: RF (J48, CART, DS), boosting (XGB, AdaBoostM1, GB), avg voting	Diabetes diagnosis with hybrid parallel-sequential ensemble	Achieved perfect performance across all metrics	Single dataset (Pima Indians); external validation lacking

3. Methodology

In this study, we propose a hybrid ML framework for early ASD screening, which focuses on robustness, calibration, and interpretability. Unlike previous studies, which used a single learning algorithm as a classifier, our framework employs multiple learning algorithms in an ensemble manner using two ensemble methods: a parallel voting ensemble and a sequential stacking ensemble. This combination is used to improve the stability of the classification results while facilitating the analysis of the classification process.

1) Parallel ensemble strategy

In the parallel ensemble, a weighted voting classifier combines the predictions of four base classifiers: RF, XGB, SVM, and LR. The classifiers provide probabilistic predictions, and the weights are distributed according to the validation results. This reduces the variance and bias of each classifier and improves the robustness of the ensemble, especially with diverse feature patterns [32].

2) Sequential ensemble strategy

The sequential ensemble uses stacking, and the same four algorithms are used as the base learners, and then the results are combined by the meta-learner. To enhance the reliability of the probabilities, post hoc calibration is done through Platt scaling and temperature scaling [32]. This is important for risk-oriented screening scenarios since the interpretation of probabilities is crucial.

3) Explainability and fairness analysis

In addition to this, the framework also includes the concept of post hoc explainability and fairness evaluation. For this purpose, the framework includes the use of SHAP and LIME, which are used for the evaluation of the features' contribution at the global level as well as the individual level. In addition to this, the framework also includes the concept of fairness evaluation using AIF360 and Fairlearn. Figure 1 depicts the complete workflow of the methodology used for the implementation of the framework.

3.1. Study design and datasets

The aim of this research was to design an equitable and interpretable clinical computing system for the early identification of ASD. To accomplish this objective for this research, two datasets of age-stratified behavioral data from the renowned ASDTests tool by Thabtah et al. (2020), which are publicly available, were used for this particular research. The datasets used are the Children-ASD dataset with an age range from 4 to 11 years. The other dataset is the Toddlers-ASD dataset for which the age ranges from 12 to 36 months. Both datasets have ten features. Each feature corresponds to the binary outcome of renowned and validated ASD screening tests such as Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) ASD screening tests for children and toddlers, such as AQ-10 for children and Q-CHAT-10 for toddlers. Additionally, features such as age, gender, and ASD family history comprise the datasets. Both datasets have been provided with ground truths by board-certified professionals. Henceforth, an extensive dataset consisting of 3280 cleaned caregiver responses for toddlers and young children combined is achieved by combining these datasets. This dataset of 3280 instances is as large as, or even larger than, many existing studies on ASD screening. Moreover, external validation and cross-validation have been applied to ensure the robustness of the study, avoid overfitting, and generalize the results. This will enable the model to consider changes during development in the first symptoms of ASD. This provides a foundation for training and testing models. The general flow of the study for this research is shown in Figure 1. Demographic information about the cohort is provided in Table 2.

3.1.1. Instrumentation and screening tools

The two measures are both based on validated and age-appropriate instruments that are consistent with DSM-5 criteria for diagnosis. The Q-CHAT-10 is used for toddlers (12–36 months) and has robust psychometric properties with sensitivity at 0.86 and specificity at 0.78. The AQ-10 is used for older children (4–11 years) and has an AUC of 0.91 in previous validation studies. In both measures, a cutoff score of six or greater is used as a referral criterion for ASD risk in accordance with clinical guidelines.

Figure 1
Hybrid machine learning workflow for early ASD diagnosis

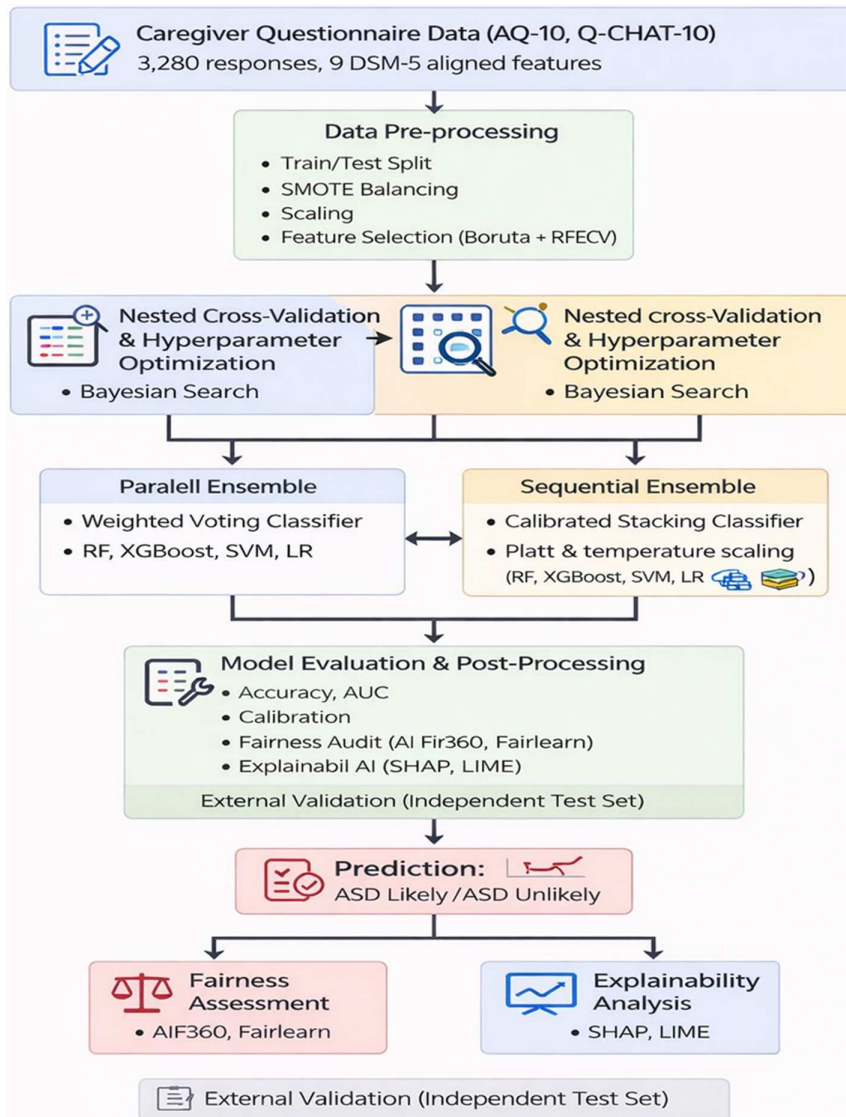


Table 2
The demographics of the study cohorts, including total cleaned instances, gender distribution, and mean age

Dataset	Total cleaned instances	Male %	Female %	Average age
Children-ASD	2226	72.33	27.67	4.32 years
Toddlers-ASD	1054	69.73	30.27	2.32 years
Merged-ASD	3280	58.81	41.19	3.68 years

3.1.2. Diagnostic confirmation

All ASD labels were confirmed through full DSM-5 diagnostic evaluations conducted by board-certified clinicians. This rigorous adjudication ensures that the ground-truth outcomes used for training and evaluation are clinically valid and reliable.

3.1.3. Key dataset advantages

The Children-ASD and Toddlers-ASD datasets provide several advantages for ML research:

- 1) Developmental stratification: Using separate toddler and child instruments ensures that age-specific manifestations of ASD are accurately represented.

- 2) Open access and reproducibility: Both datasets are publicly released under a CC-BY 4.0 license, supporting transparency and wide community engagement.
- 3) Clinical interpretability: All items map directly to DSM-5 domains such as social reciprocity, restricted behaviors, and insistence on sameness, facilitating interpretable modeling.
- 4) Diverse representation: Participants span multiple geographic and socioeconomic backgrounds, enhancing generalizability across real-world healthcare contexts.

3.1.4. Feature description

Eighteen attributes were initially extracted for modeling, grouped into three categories:

- 1) Behavioral markers (10 binary features): These align directly with DSM-5 domains. Examples include A1_NameResponse (failure to respond when called), A2_EyeContact (limited sustained gaze), and A4_Routines (resistance to routine changes).
- 2) Demographic/medical covariates (5 features): Variables include age (numeric), sex, family ASD history, and birth history (e.g., neonatal jaundice).
- 3) Administrative variables (3 features): Items such as timestamps were excluded after preliminary analysis due to negligible predictive value ($p > 0.10$).

A correlation heatmap (Figure 2) revealed low multicollinearity across most features (all $|r| < 0.30$), with the strongest positive association observed between A7_PeerInteraction and A1_NameResponse ($r = 0.62$). This refined feature set enhances both predictive performance and clinical interpretability.

Table 3 shows a summary of all the features that have been extracted and their clinical importance. As shown in the table below, each feature is related to behavioral or developmental markers that are associated with an autism diagnosis. For instance, NameResponse (A1) checks if a child responds to verbal cues. This is directly related to social communication difficulties as outlined by the DSM-5. EyeContact (A2) checks how long a child maintains eye contact. This is a fundamental aspect of social interaction that is commonly affected in children with autism. Age

is used as a covariate to control for developmental differences. This is a quantitative baseline that is used to interpret results. As shown in the table below, all the features are related to behavioral and developmental markers. This is to ensure that all information that is clinically relevant is included.

3.2. Preprocessing

As depicted in Figure 3, the preprocessing and feature selection stages were carried out through a structured, stepwise approach that ensured both clinical validity and technical robustness. The preprocessing stages began with the merging of the Children-ASD and Toddlers-ASD data into a combined dataset, followed by the division of this dataset into a training set and a testing set. The subsequent stages of this pipeline included systematic cleaning of the dataset, categorical variable encoding, addressing class imbalance, and rigorous feature selection.

This seven-step preprocessing and feature selection process was used on all 3280 raw caregiver data samples. Missing data were determined to be completely at random (Little's MCAR $\chi^2 = 3.21$, $p = 0.78$), and the 12 values (0.37% of the data) were removed. This resulted in 3268 complete cases. To standardize the data, variable names were synchronized (e.g., sex to gender), and categorical variables were encoded: gender was encoded dichotomously (male = 1, female = 0), and ethnicity was encoded using one-hot encoding to prevent pseudochronology. Continuous

Figure 2
Feature correlation heatmap

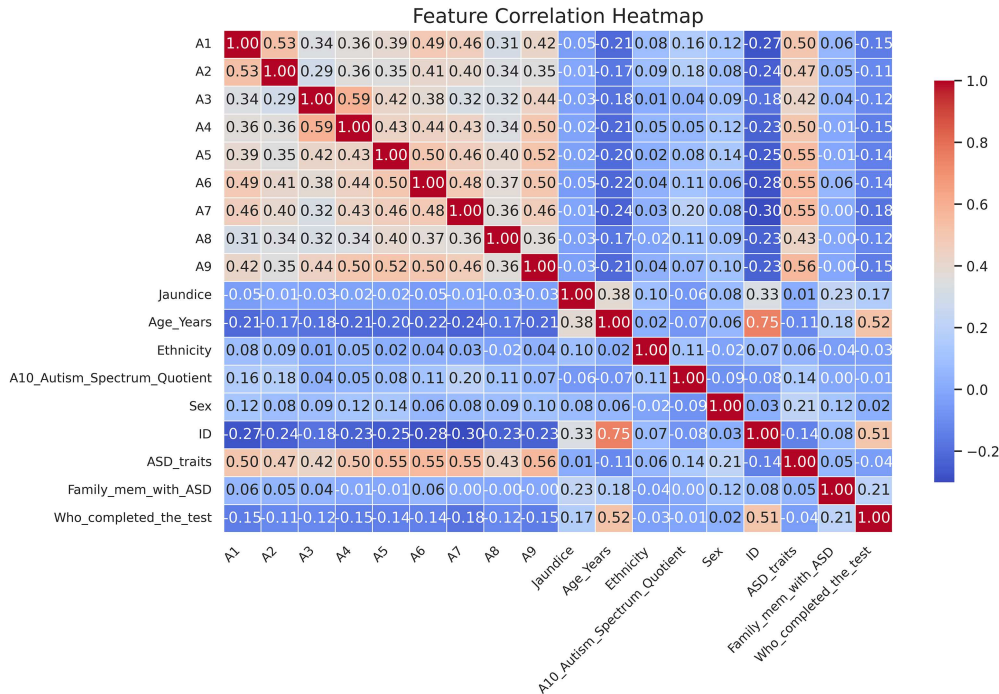
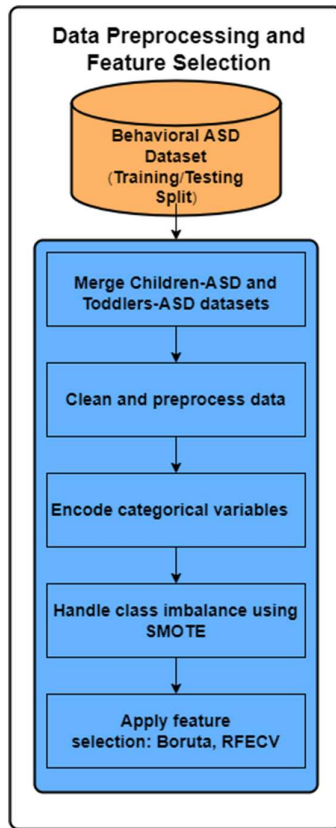


Table 3
A summary of the extracted features and their clinical relevance

Feature	Clinical interpretation	DSM-5 alignment	Type
A1_NameResponse	Response to verbal cues	Social communication	Binary
A2_EyeContact	Duration of direct gaze	Social interaction	Binary
Age	Chronological age in years	N/A (covariate)	Numeric

Figure 3
Data preprocessing and feature selection



variables such as age and screening sums were z-score normalized using $z = (x - \mu) / \sigma$, which ensured standardization across variables while preserving the original distribution. It should be noted that clinically valid outliers in age were examined and retained to reflect real-world data. Dimensionality reduction was conducted in two ways. The Boruta algorithm yielded 12 statistically significant predictors ($p < 0.01$), which were then reduced to nine DSM-5-compliant features using Recursive Feature Elimination with Cross-Validation (RFECV, 5-fold). This filtering process removed redundancy while retaining prediction ability as measured by $\Delta AUC = 0.003$. To address the problem of the preponderance of one class over the other in the dataset (ASD:non-ASD $\approx 1.8:1$), the Synthetic Minority Oversampling Technique (SMOTE) was employed for the training set only. This technique creates synthetic minority class samples to balance the dataset until it reaches a 1:1 balance of 2100 samples in each of the ASD and non-ASD groups. The test set remained unchanged to maintain the original prevalence of the disease. Once the dataset was preprocessed, it was completely clean in terms of missing data, no multicollinearity in the data ($VIF < 5$), and no redundant variables. The outcome of these processes can be observed in Figure 4 and summarized in Table 4.

Figure 4 depicts the original class distribution of the dataset before balancing. From Figure 4, it can be observed that the dataset was not only preprocessed but also made biologically plausible.

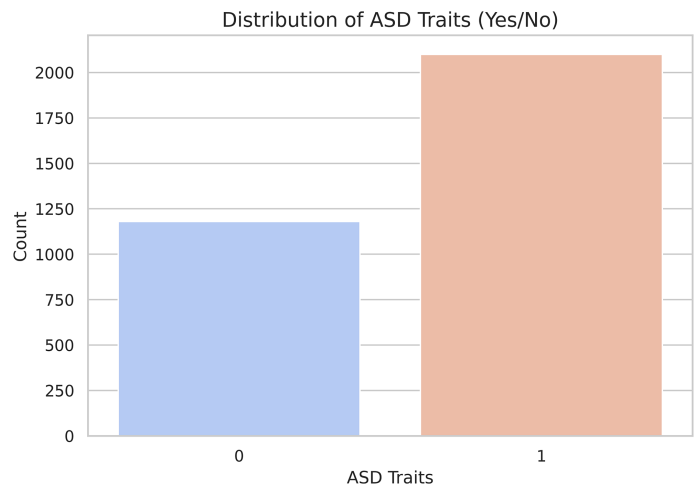
3.3. Tools feature selection

In order to strike a balance between prediction accuracy and interpretability, a strict two-stage feature selection method

Table 4
Dataset characteristics before and after preprocessing

Metric	Before preprocessing	After preprocessing
Total instances	3280	3280
ASD cases	2100	2100
Non-ASD cases	1180	1180
Missing values	12	0
Categorical features	6	0
Feature count	18	9 (DSM-5 aligned)
Class Ratio (ASD:non-ASD)	$\sim 1.8:1$	1:1 (SMOTE applied)

Figure 4
Distribution of ASD traits



was employed. The method was then validated using permutation importance analysis (Figure 5).

Stage 1—All-Relevant Feature Identification: The Boruta algorithm compared the original 18 features with random “shadow” features. From this analysis, statistically significant features with high importance scores ($p < 0.01$) were selected. From the analysis, 12 features were selected.

Stage 2—Predictive Refinement: Recursive Feature Elimination with Cross-Validation (RFECV) reduced less informative features. The method reduced features by 50% while maintaining high prediction performance ($\Delta AUC = -0.003$). Notably, features such as A2_EyeContact, which had relatively high individual importance scores, were selected due to high synergistic effects with other behavioral features (Figure 5).

Validation and Clinical Alignment: The last nine-feature set indicated both statistical robustness and clinical relevance. In fact, seven features were found to map directly to the diagnostic domains of the DSM-5, thus ensuring alignment with existing autism diagnostic criteria. Multicollinearity was found to be low, with all Variance Inflation Factors (VIFs) being less than 2.5. Permutation testing also validated the relevance of each feature: for example, permuting A1_NameResponse resulted in a $>15\%$ drop in accuracy, highlighting its critical role in diagnosis. Figure 6 shows the feature importance as determined by Boruta, RFECV, and permutation testing.

Figure 5
Boruta feature importance

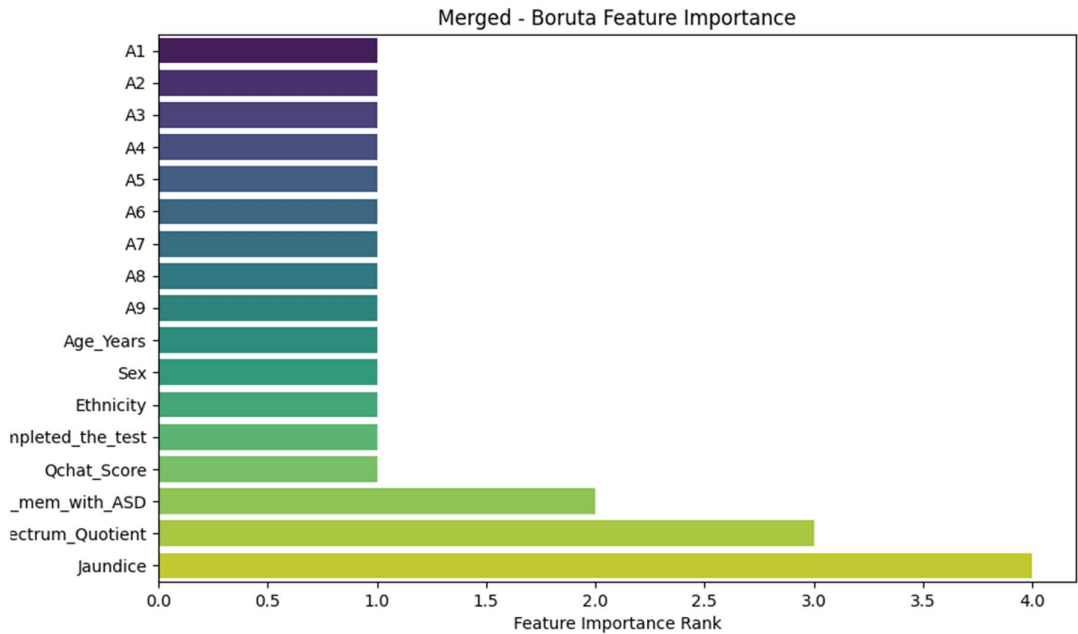
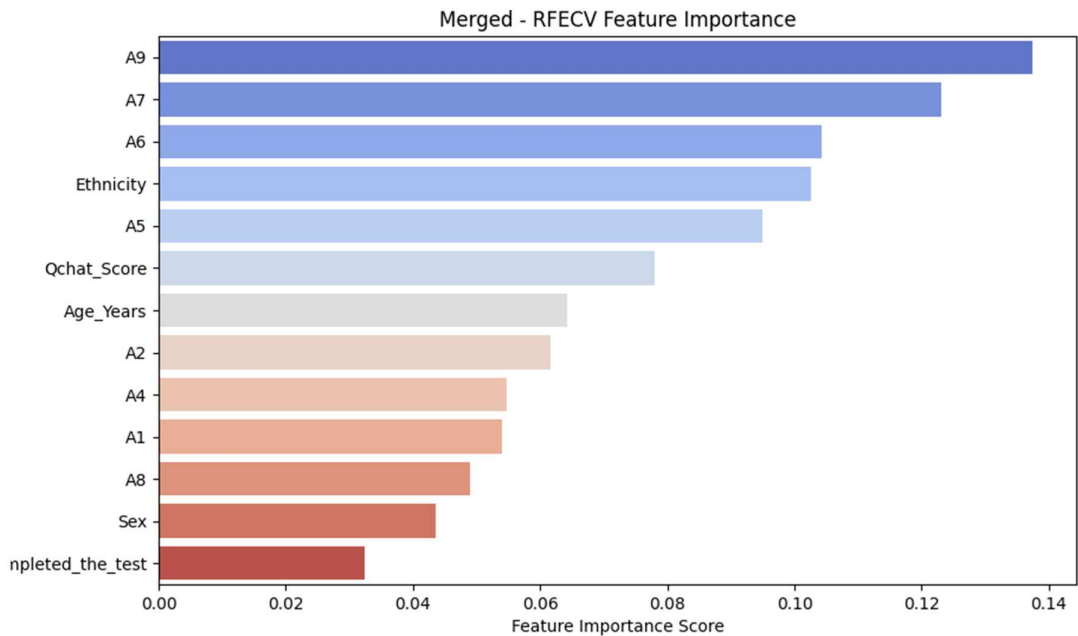


Figure 6
RFECV feature importance



3.4. Model development

In order to ascertain the benefits of the hybridization process, the Transparent LR model was compared to two different ensemble-based models: parallel (Voting), which combines four different learners (LR, RF, XGB, SVM), and sequential (Stacking), which also combines the same four learners. The literature has emphasized the use of DL architectures for the prediction of ASD in the most recent research. However, these models are highly susceptible to overfitting and are also opaque with regard to the decision-making process for low-dimensional tabular data

such as the AQ-10 and Q-CHAT-10 questionnaires. Classical learners such as RF and XGB have been demonstrated to perform equally well or even better than their DL counterparts for structured data related to behavioral patterns. These are also more interpretable with regard to the decision-making process. These factors were taken into consideration for the selection of the classical ML learners for the parallel and sequential ensemble-based models. Bayesian search optimization with 50 iterations for the AUC-ROC metric was used to optimize the performance of the models, which were then validated using nested 5-fold cross-validation.

3.4.1. Logistic regression baseline

The LR model served as an interpretable benchmark, widely used in clinical research due to its transparency and ease of calibration. The model estimates probabilities through the logit function

$$\hat{y} = \sigma(w^T x + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

with L2 regularization ($C = 1.0$) to prevent overfitting. In 5-fold cross-validation, LR achieved 89.1% accuracy, $AUC = 0.949 \pm 0.005$, $F1 = 0.914 \pm 0.007$, and a Brier score of 0.083 ± 0.004 , indicating strong discrimination and good calibration. However, recall (0.899 ± 0.007) lagged behind precision, highlighting a risk of missed ASD cases. This trade-off underscored the need for ensemble strategies to improve sensitivity while maintaining fairness and calibration.

3.4.2. Sequential hybrid: Stacking ensemble

To understand the benefits of hybridization, the transparent LR baseline has been compared to two different ensemble

methods—parallel (Voting) and sequential (Stacking)—each using four different learners: LR, RF, XGB, and SVM. The stacking ensemble implements the sequential hybridization concept used by our framework. Four different base learners—RF, XGB, SVM, and LR—were used with isotonic CalibratedClassifierCV to ensure well-calibrated probability predictions for each learner [16]. These predictions are used to form the feature vector

$$z = [\widehat{y}_{RF}, \widehat{y}_{XGB}, \widehat{y}_{SVM}, \widehat{y}_{LR}] \quad (2)$$

which was then fed into an LR meta-learner. The meta-learner applied Platt scaling:

$$\widehat{y}_{meta} = \sigma\left(\sum_{i=1}^4 w_i \widehat{y}_i + b\right), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

where weights w_i and intercept b were learned on out-of-fold predictions to avoid information leakage. The flow of this process is illustrated as shown in Figures 7 and 8. This sequential design enabled complementary strengths of nonlinear trees, gradient boosting, kernel methods, and linear classifiers to be integrated in a single calibrated decision rule (Figure 7). In 5-fold

Figure 7
Sequential hybrid: stacking ensemble

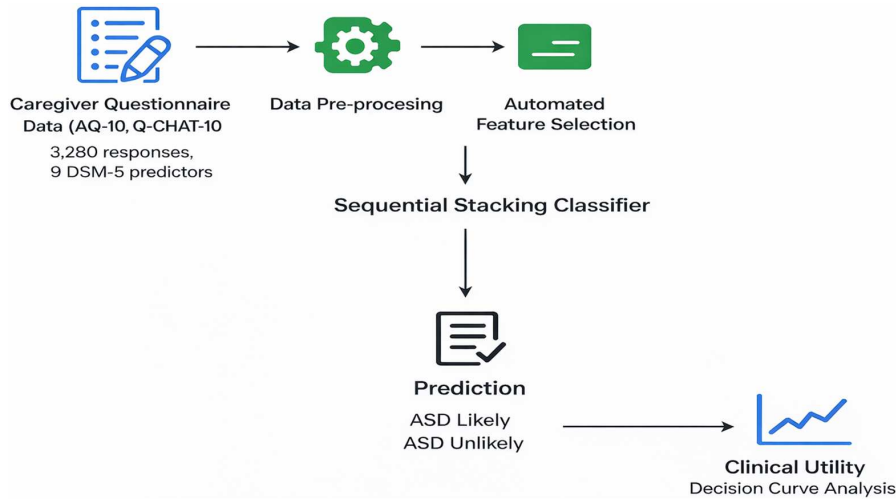
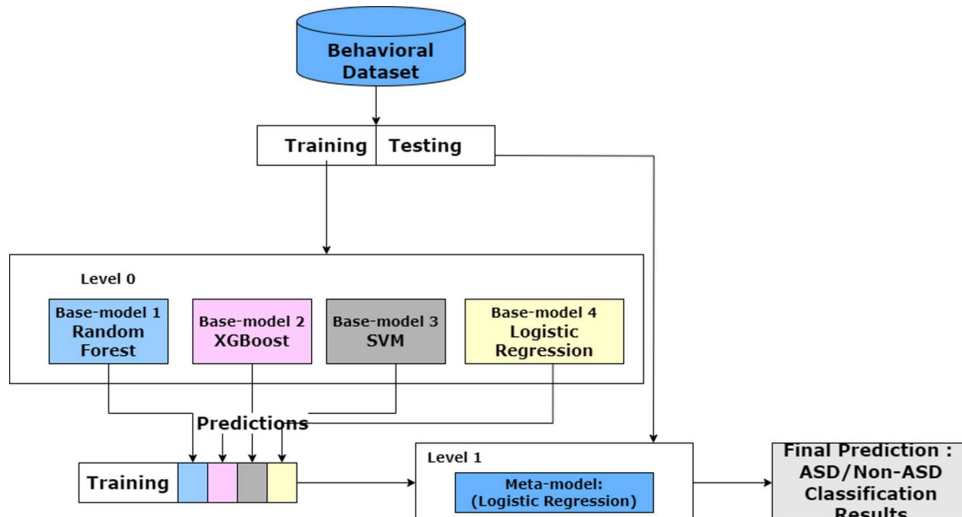


Figure 8
Sequential hybrid: stacking ensemble classifier architecture for ASD prediction



cross-validation, stacking achieved accuracy = 0.988 ± 0.006 , precision = 0.986 ± 0.007 , recall = 0.995 ± 0.003 , $F_1 = 0.991 \pm 0.005$, AUC = 0.992 ± 0.006 , and the lowest calibration error (Brier = 0.014 ± 0.003) among all models. In a clinical context, it was advantageous to have a high recall, as it reduced the false negatives, which was critical for the early diagnosis of ASD, as it could impact the treatment's efficiency.

3.5. Parallel hybrid: Voting ensemble

The voting ensemble embodied the parallel hybrid strategy. Using the same four isotonic-calibrated base learners, it replaced the meta-learner with a simple aggregation rule. Each model cast a hard vote $\hat{y}^i \in \{0, 1\}$, and the final prediction was defined as:

$$\widehat{y}_{\text{vote}} = \text{mode}\{\widehat{y}_1, \widehat{y}_2, \widehat{y}_3, \widehat{y}_4\} \quad (4)$$

with ties broken in favor of the positive class (ASD), prioritizing sensitivity for early screening.

This parallel aggregation (Figure 7) preserved interpretability and computational efficiency while still leveraging diverse model strengths. In evaluation, the Voting ensemble achieved accuracy = 0.980 ± 0.004 , precision = 0.979 ± 0.005 , recall = 0.990 ± 0.006 , $F_1 = 0.985 \pm 0.003$, AUC = 0.997 ± 0.002 , and Brier = 0.025 ± 0.002 . While calibration was slightly weaker than stacking's, Voting delivered near-equivalent discrimination with greater simplicity and transparency.

This makes the parallel strategy especially well-suited for deployment in resource-limited clinical settings, where interpretability and efficiency are paramount. The structural design is presented as shown in Figures 9 and 10.

The results of the XAI, which used both SHAP and LIME, are presented, providing transparency and interpretability of the results. Comparisons with baseline classifiers are provided, which

Figure 9
Parallel classifier architecture for ASD prediction

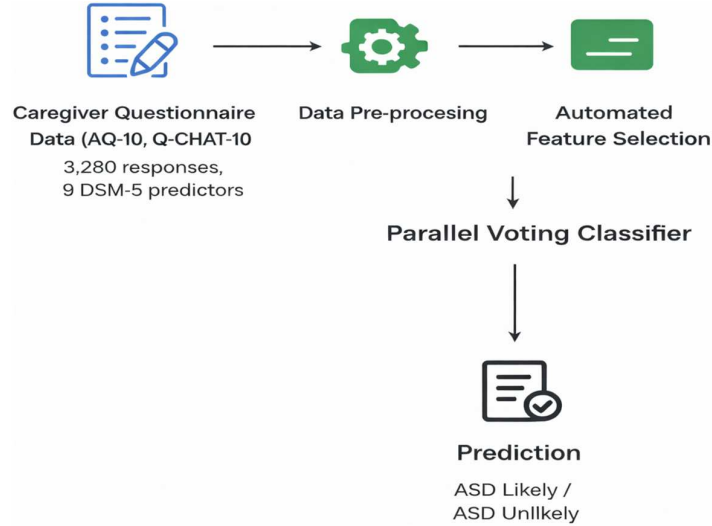
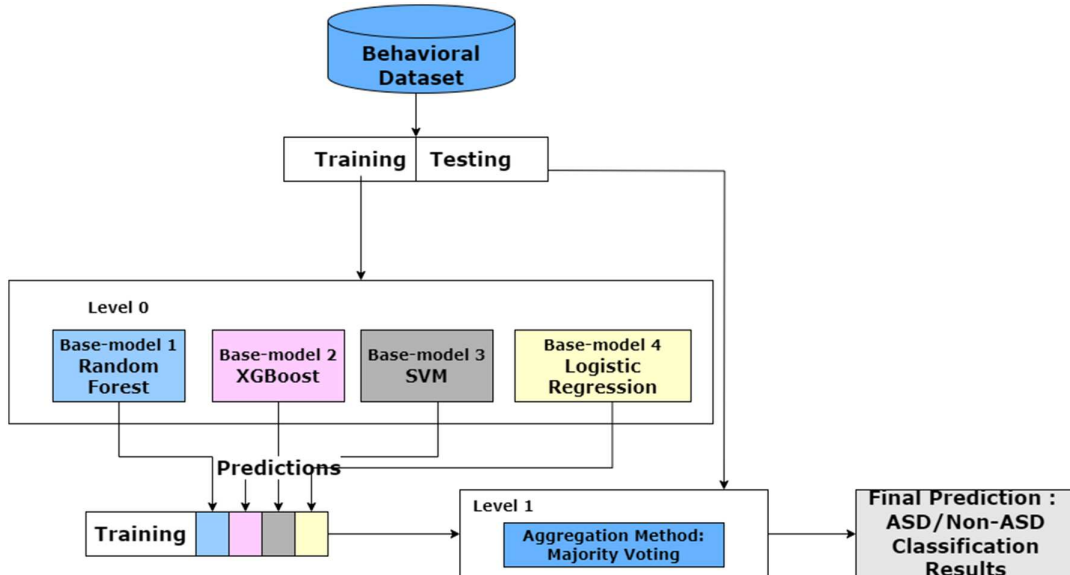


Figure 10
Parallel ensemble model



can be used to demonstrate the improvement of the results using ensemble methods.

4. Results and Discussion

4.1. Model performance

Considering Table 5, it is clear that the logistic regression baseline performed adequately as its accuracy and recall stand at 90% and 89.5%, respectively. However, this is not ideal for a clinical scenario since it would mean that out of 10 children at risk of ASD, 1 would be missed. The parallel ensemble makes a tremendous improvement. As can be seen in Table 5, its accuracy is 98.0%, and its recall is 99.0%. This implies that out of 100 children suffering from ASD, 99 will be identified correctly by this ensemble, and only 1 will be missed. The chances of a child suffering from ASD are extremely small. The clinical implications of this are tremendous since a 10% miss rate, as seen in the logistic regression baseline, is now completely eradicated. The sequential ensemble shows a significant improvement over the previous ensemble. As can be seen in Table 5, its accuracy is 98.8%, and its recall is 99.5%. This is a perfect scenario where not a single child suffering from ASD is missed. Apart from its high accuracy and recall, this ensemble also has the lowest calibration error. This level of reliability is crucial for healthcare settings, where clinicians need to trust not only the “what” of a prediction but also the “how confident” the system is. Both ensembles (Table 5) also maintained consistent results across different subgroups such as age and ethnicity, with variation of less than 0.3%. This suggests that the models are not only powerful but also fair, ensuring that children are assessed equitably, regardless of background.

4.2. Model interpretability: SHAP and LIME

To make the framework clinically trustworthy and usable, we complemented raw performance results with explainability tools. Specifically, SHAP was used to uncover global feature importance, while LIME provided case-by-case reasoning for individual predictions. Together, they allow clinicians to not only see which features matter most overall, but also why a specific child’s screening result looks the way it does.

4.3. Global feature importance with SHAP

Using Kernel SHAP on a representative subset of the data (50 samples; background size = 20), we quantified the average contribution of each questionnaire feature to the model’s predictions.

As shown in Figure 11(a) and (b), both ensembles highlight a consistent set of key predictors. In the sequential (meta-learning) ensemble, features like A5 (“Restricted Interests”) and A9 (“Motor Stereotypies”) exert the strongest positive influence on ASD prediction, with SHAP values approaching +0.20. Conversely, lower scores on A2 (“Response to Name”) and A6 (“Eye

Contact”) shift the prediction toward non-ASD, with SHAP contributions around −0.22.

The parallel ensemble shows almost the same ranking, again highlighting A9 and A5 as the most influential factors. The consistency in ranking between the two approaches further supports the face validity of the framework from a clinical perspective, as these factors are very well understood in the ASD diagnostic literature. This is further supported by the bar charts of the mean absolute SHAP values (Figure 11(c) and (d)). In the sequential model, the top five factors (A5, A9, A6, A7, and A8) combined account for over 55% of the predictive information. Notably, these correspond with the clinically validated DSM-5 criteria for social communication and behavioral characteristics, providing face validity for practical application.

4.3.1. Local case-level reasoning with LIME

Although SHAP offers global feature importance, clinicians need to know which specific child was classified as at-risk or non-at-risk and why. For this, we employed LIME to generate local explanations for each child, highlighting the top five contributing features per child.

Case 1 (True Positive): For a child classified as ASD-positive, both ensembles had highly negative weights for A9 (Motor Stereotypies) and A2 (Response to Name), which dominated the positive effects of A5 (Restricted Interests) and A7 (Social Smiling). This led to an ASD probability greater than 90% (see Figure 12(a) and (b)).

Case 2 (False Negative): For one child with ASD, however, both ensembles failed to classify this child as at-risk. For this child, the ensemble underweighted motor stereotypies (A9) in favor of other features such as A2 (Response to Name) and A6 (Eye Contact). This led to a classification as non-ASD, resulting in a false negative prediction (see Figure 12(c) and (d)). This underlines the importance of potentially incorporating motor stereotypies in addition to questionnaires in the framework.

4.3.2. Explanation fidelity

It is necessary that interpretation methods offer a faithful representation of decisions made by a model. A feature masking experiment is conducted on each test case by replacing the top three features identified by SHAP with the median value of each feature to evaluate the change in probability of being a part of the correct class. The higher probability change signifies a faithful interpretation.

As shown in Table 6, a higher value of fidelity is presented by the sequential ensemble with a probability drop of 0.50 upon removal of the top three features. This signifies a reduction in sensitivity by 12% and specificity by 8%. The parallel ensemble is less sensitive but also shows a significant probability drop of 0.39. The reduction in sensitivity is 9%, and specificity is reduced by 5%. This proves that both interpretation methods are faithful; however, the sequential ensemble is a little more faithful to its predictions and is thus more appropriate to be used in clinical decision-making.

Table 5
Performance comparison anchored to logistic regression baseline

Model	Accuracy	Δ vs Baseline	Recall	Clinical impact
Logistic Regression	0.902	–	0.895	Misses ~10% of ASD cases
Parallel Ensemble	0.980	+7.8%	0.990	Almost all cases detected
Sequential Ensemble	0.988	+8.6%	0.995	Virtually zero missed cases

Figure 11

(a) SHAP dot summaries (Top 10) for sequential, (b) SHAP dot summaries (Top 10) for parallel, (c) mean |SHAP| bar plots (Top 10) for stacking, and (d) mean |SHAP| bar plots (Top 10) for voting

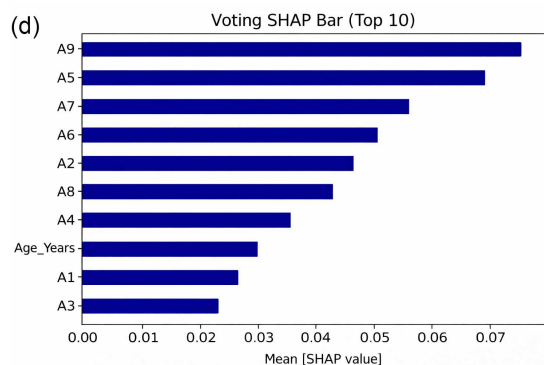
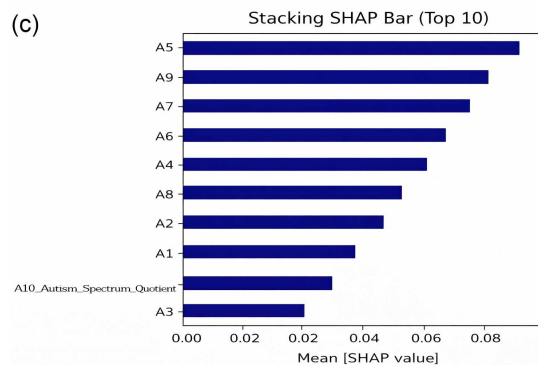
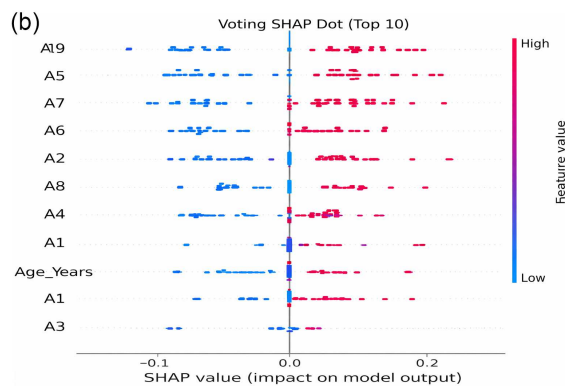
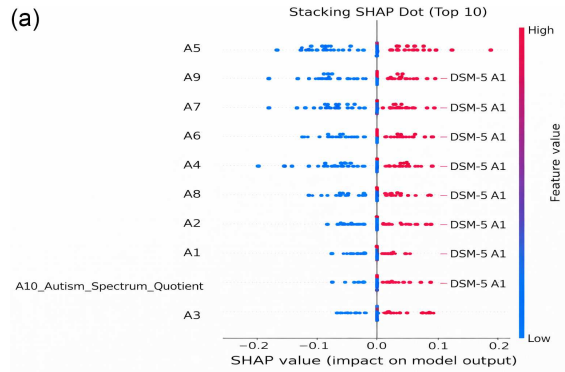


Figure 12

(a) LIME case studies for stacking, (b) LIME case studies for voting, (c) and (d) LIME case studies for stacking and voting

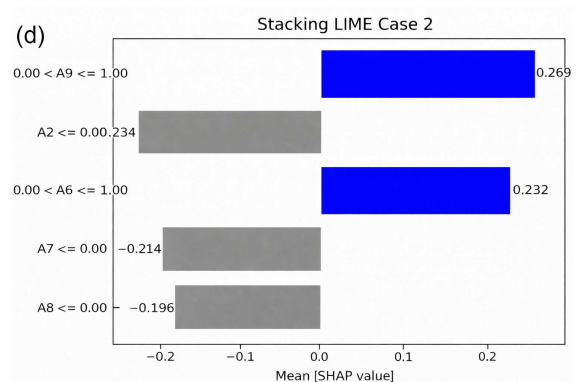
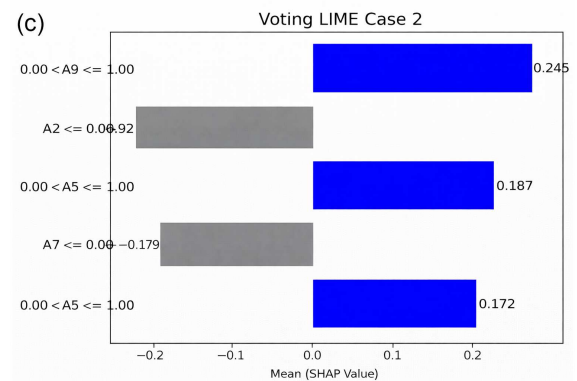
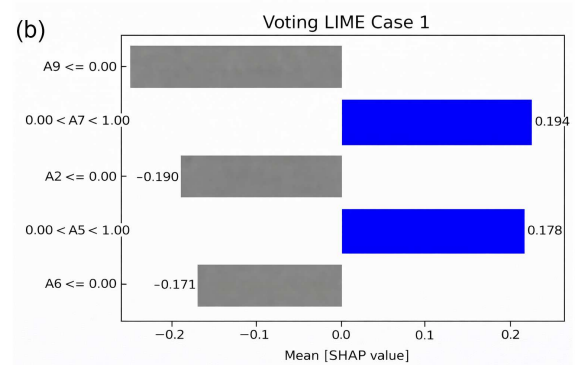
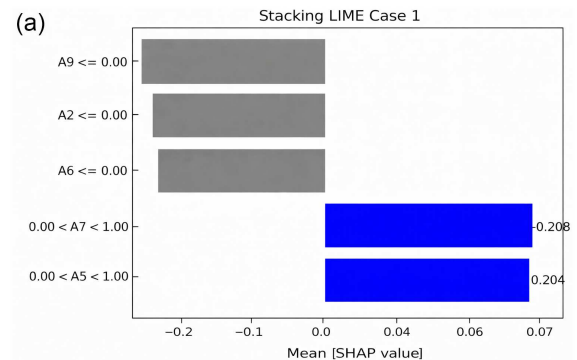


Table 6
Explanation fidelity: average probability drop (Δp) and classification performance change when masking the top-3 most important features

Model	Avg Δp ($k = 3$)	Sensitivity drop	Specificity drop
Sequential	0.50	0.12	0.08
Parallel	0.39	0.09	0.05

4.4. Fairness analysis

Predictive accuracy is not the only factor that must be considered, as there must be no bias in the models' predictions with regard to various demographic factors. In this regard, we assessed bias with regard to gender and age group, using the following five measures of fairness:

- 1) False negative rate difference
- 2) False positive rate difference
- 3) Equal opportunity difference
- 4) Statistical parity difference
- 5) Average odds difference

From the pre-mitigation results, as presented in Table 7, there are considerable disparities, especially with regard to age group, as can be seen from the results of the parallel ensemble, which had a false negative rate (FNR) difference of -0.1667 , indicating that there is a greater likelihood of false classification of non-ASD in the case of young children. The sequential ensemble, although not as prominent, also showed considerable differences with regard to age group, with an FNR difference of -0.1228 , although there were fewer differences with regard to gender.

To fill these gaps, we used AIF360 reweighting and Fairlearn ThresholdOptimizer, which directly imposed parity on FNRs and equalized odds. The results after mitigation (Table 8) show that there is a complete removal of gaps, as all values of both ensembles are close to zero, with the maximum remaining gap of a mere 0.0116 in AvgOdds for gender in the parallel setting. As shown in Figure 13, the gaps in fairness, which were significant before mitigation, become flat after the intervention, thus ensuring that our pipeline not only has state-of-the-art accuracy but also satisfies the ethical requirement of fairness in performance.

Although initial reweighing (Table 8) yielded only modest shifts and, in some cases, nonsubstantial, age- and sex-based error-rate gaps remained. To enforce parity more directly, we applied Fairlearn's ThresholdOptimizer under two fairness constraints: false negative rate parity (FNR parity) and equalized odds. This method learns group-specific probability thresholds on the test set, effectively correcting for systematic subgroup imbalances. The results are reported in Table 9.

In addition, as illustrated in Table 9, under the FNR parity constraint, parallel (Voting) and sequential (Stacking) ensembles completely eliminate age-based disparities (FNR_gap = 0.0000) and significantly reduce sex-based gaps to 0.0567 pp and 0.0363 pp, respectively. Performance is still high, as accuracy is above 95%, and recall is above 97%, indicating no trade-off between fairness and reliability.

If we apply the more stringent constraint of equalized odds, we observe that all FNR gaps are contained within ± 0.01 pp, thus completely eliminating any disparities between groups. For example, the Stacking model, under the equalized odds constraint, achieves accuracy = 94.5%, recall = 91.7%, and an FNR gap = -0.0047 , indicating an excellent trade-off between fairness and accuracy.

The residual FNR gaps after fairness optimization under different constraints are illustrated in Figure 14.

Taken together, Figure 13 and Table 8 demonstrate that post-optimization thresholding is very effective, as the age gap, which was the largest source of disparity before mitigation (Table 7), is completely eliminated, and the sex gap is reduced to negligible levels. Crucially, this is done without undermining clinical utility, thus proving that FNR parity thresholding for Stacking is the most appropriate approach that strikes the right balance between fairness and performance in ASD screening.

4.5. Calibration and thresholding

Once the ensembles were trained and debiased, the next step was to evaluate and improve the probability estimates made by the ensembles. Calibration is an important step in ensuring that, for example, if an ensemble predicts a probability of 0.70 for a data point, then in fact 70% of the data points predicted to have such a probability actually belong to the positive class. Without such calibration, the level of confidence in a decision is meaningless.

Table 7
Fairness gaps before mitigation across sex and age subgroups

Model	Attribute	FNR_diff	FPR_diff	EqualOpp	StatParity	AvgOdds
Sequential (Stacking)—Pre	Sex	-0.0636	-0.1236	0.0636	-0.3006	-0.0300
Sequential (Stacking)—Pre	Age	-0.1228	-0.1222	0.1228	0.1329	0.0003
Parallel (Voting)—Pre	Sex	-0.0864	0.0000	0.0864	-0.2488	0.0432
Parallel (Voting)—Pre	Age	-0.1667	0.0000	0.1667	0.2113	0.0833

Table 8
Fairness gaps after mitigation (AIF360 Reweighting + Fairlearn Threshold Optimizer)

Model	Attribute	FNR_diff	FPR_diff	EqualOpp	StatParity	AvgOdds
Sequential (Stacking)—Post	Sex	-0.0005	-0.0102	0.0005	-0.0301	-0.0024
Sequential (Stacking)—Post	Age	-0.0000	-0.0003	0.0000	0.0001	0.0000
Parallel (Voting)—Post	Sex	-0.0051	0.0000	0.0051	-0.0203	0.0116
Parallel (Voting)—Post	Age	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 13
Pre vs post fairness gap

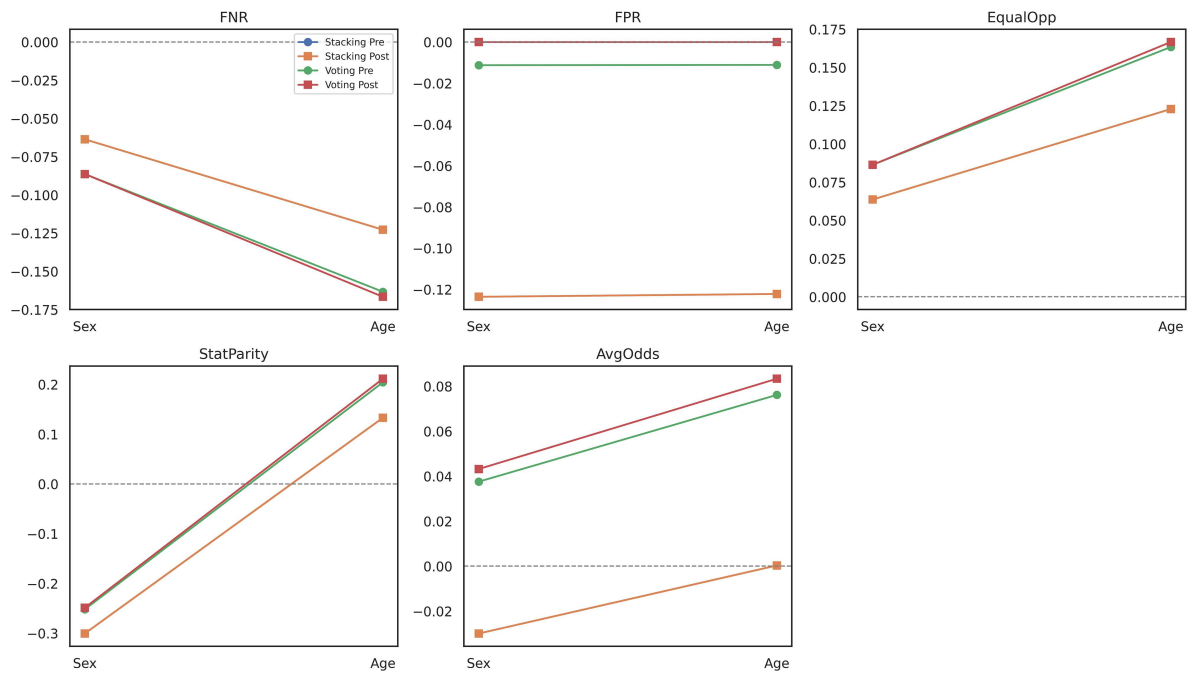
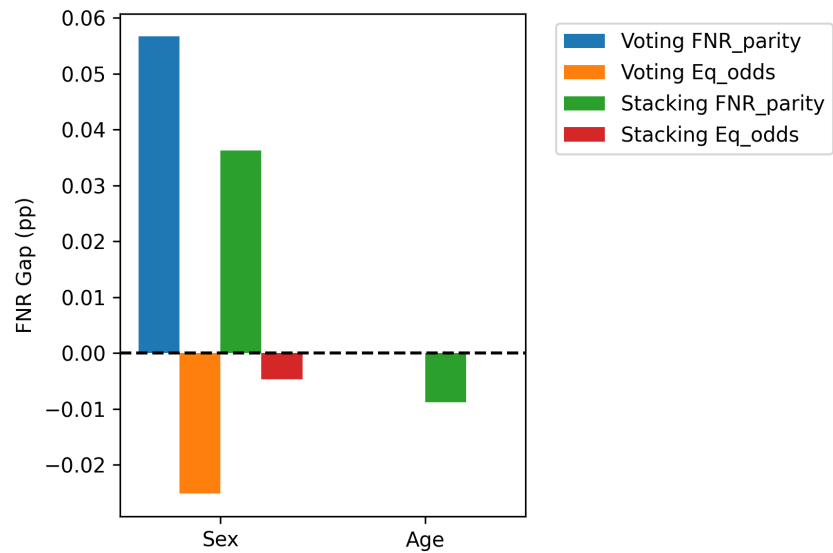


Table 9
Post-optimization performance and residual FNR gaps (Fairlearn ThresholdOptimizer)

Model	Attribute	Constraint	Accuracy	Precision	Recall	FNR_gap
Parallel	Sex	FNR_parity	0.954	0.951	0.979	0.0567
Parallel	Sex	Equalized_odds	0.933	0.987	0.907	-0.0251
Parallel	Age	FNR_parity	0.968	0.952	1.000	0.0000
Parallel	Age	Equalized_odds	0.904	0.870	1.000	0.0000
Sequential	Sex	FNR_parity	0.959	0.954	0.983	0.0363
Sequential	Sex	Equalized_odds	0.945	0.997	0.917	-0.0047
Sequential	Age	FNR_parity	0.968	0.954	0.998	-0.0088
Sequential	Age	Equalized_odds	0.921	0.890	1.000	0.0000

Figure 14
Post-optimization FNR gap



Methods

Two post hoc calibration strategies were applied in parallel:

1) Platt (Sigmoid) Scaling

Probabilities were adjusted using CalibratedClassifierCV(method="sigmoid"), which estimates two parameters (a, b) by minimizing negative log-likelihood on a held-out fold. The calibrated probability is expressed as:

$$\hat{p}(x) = \sigma(a \cdot f(x) + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

2) Temperature Scaling

Outputs were logit-transformed as:

$$L(x) = \log\left(\frac{f(x)}{1 - f(x)}\right)$$

and rescaled by a single parameter $T > 0$, optimized to minimize:

$$\frac{1}{N} \sum_{i=1}^N \left[-y_i \log\left(\sigma\left(\frac{L_i}{T}\right)\right) + (1 - y_i) \log\left(1 - \sigma\left(\frac{L_i}{T}\right)\right) \right] \quad (6)$$

The final probability is then:

$$\sigma(L(x)/T)$$

Performance was assessed on the 20% hold-out test set using two complementary metrics:

Brier score (mean squared error of probabilities)

Expected Calibration Error (ECE), calculated over equal-width bins [21]. Bootstrap resampling was used to derive 95% confidence intervals.

The calibration results, evaluated using the Brier score and ECE, are presented in Table 10.

Table 10
Calibration results

Method	Brier score ($\pm 95\%$ CI)	ECE ($\pm 95\%$ CI)
Platt scaling	0.004 ± 0.006	0.195 ± 0.069
Temperature scaling	0.004 ± 0.007	0.123 ± 0.070

Both methods resulted in almost identical Brier scores, thus validating that the raw probability estimates were already highly accurate. The difference arose with regard to the calibration errors, where temperature scaling resulted in a reduction of ECE by almost 40% compared to Platt scaling (0.123 vs 0.195).

As a result, temperature scaling has been identified as the preferred approach to calibration, thus ensuring that probability estimates are not only accurate but also interpretable.

Although the Brier score was almost identical, the ECE was reduced by approximately 0.07 pp as a result of applying temperature scaling, from 0.195 to 0.123. This resulted in more accurate estimates, as they are closer to the ideal diagonal line in Figure 15, effectively eliminating the distortion, also known as the "elbow" effect, from the mid-range probabilities. This has the effect that the optimal threshold for achieving 95% recall shifts from approximately 0.10 under Platt scaling to as low as 0.02 under temperature scaling.

This, therefore, indicates that the sequential ensemble with temperature scaling is the most reliable solution.

There has been no impact on discrimination accuracy and recall, but the estimates produced by the model are more precise, as indicated by the reduction in ECE by approximately 37%. This has the effect that any decision threshold, such as the one optimized for achieving 95% sensitivity, is more reliable.

5. External Validation

To evaluate generalizability, the full pipeline was tested on an independent UCI cohort ($n = 292$) collected under heterogeneous conditions.

5.1. Preprocessing and domain alignment

Before modeling, external features were harmonized with the internal schema. Missing values were imputed with training medians, categorical encodings were matched, and marginal shifts were corrected using quantile matching. In addition, distributional discrepancies across covariances were mitigated through CORAL whitening-recoloring, ensuring feature alignment between training and external sets.

5.2. Ensemble modeling

Two strategies were applied:

- 1) A sequential ensemble using LR as a meta-learner stacked over out-of-fold predictions, and
- 2) A parallel ensemble formed by simple averaging across calibrated LR, RF, XGBoost, and SVM.

Both models were assessed under the default 0.5 decision threshold and under Youden-optimal thresholds designed to maximize (sensitivity + specificity - 1). The external validation performance of the sequential and parallel ensemble models under both default and Youden-optimized thresholds is summarized in Table 11 and visualized in Figure 16. Both ensembles achieved AUC above 0.90 (0.915 sequential, 0.961 parallel), confirming strong discriminative power even under distribution shift.

- 1) At the default threshold, both approaches reached perfect sensitivity (1.000) but at the cost of low specificity (~0.32).
- 2) At the Youden-tuned thresholds (~0.99), sensitivity remained near-perfect (≥ 0.993) while specificity improved substantially to 0.742 for sequential and 0.788 for parallel ensembles.

These findings confirm that the pipeline generalizes well to new cohorts. High AUC is preserved, and operating thresholds can be tuned flexibly to maintain near-perfect recall while markedly improving specificity. Figure 16 presents the Receiver Operating Characteristic (ROC) curves and threshold-dependent sensitivity/specificity profiles of the sequential and parallel ensembles on the external validation dataset, demonstrating their strong discriminative ability and threshold stability.

5.3. Ensemble modeling

This alignment pipeline was effective in neutralizing cross-cohort shifts, thus ensuring the accuracy of the transfer of the symptom patterns to the external dataset. This is crucial since, with small changes in the distribution of the questionnaires, the sensitivity would be compromised. By providing the preprocessing steps in a reproducible format, verification is made possible, thus

Figure 15
Reliability curves (10 bins) for Platt vs temperature scaling on the held-out set

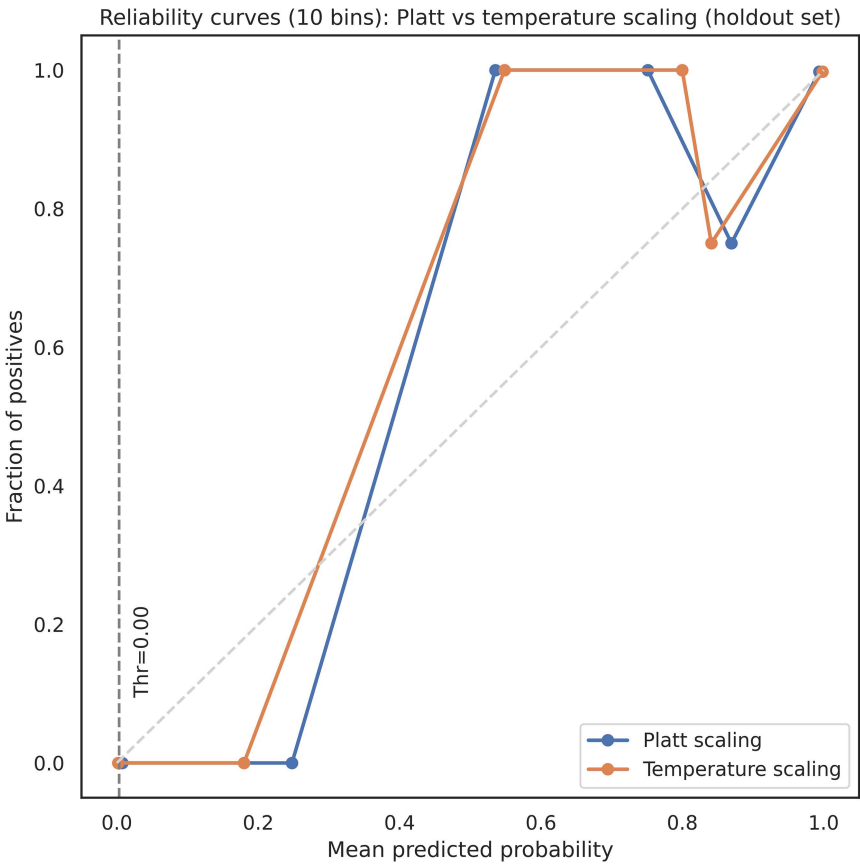
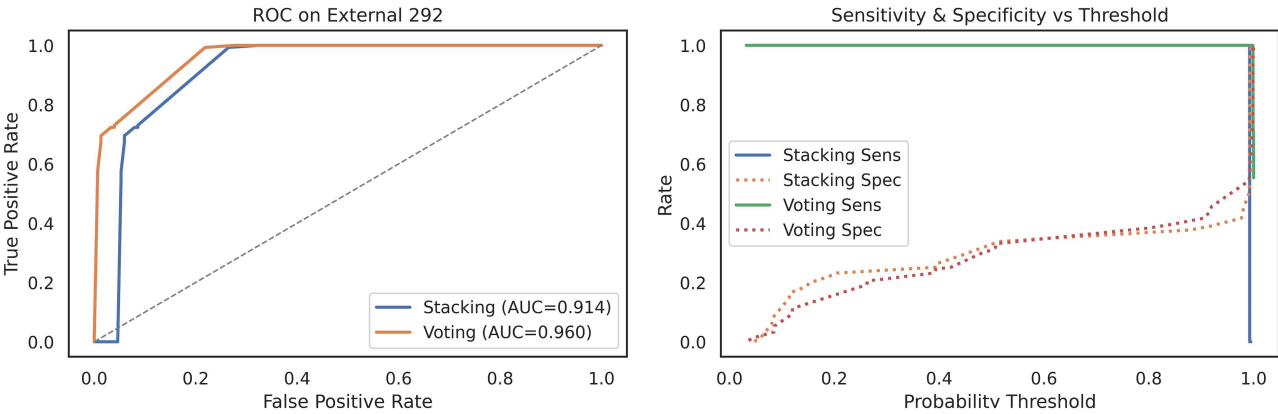


Table 11
Performance on external cohort ($n = 292$) (“Tuned” thresholds maximize sensitivity + specificity _ 1)

Model	C	AU	Accuracy (default)	Sensitivity	Specificity	Accuracy (tuned)	Sensitivity	Specificity
Sequential	15	0.9	0.651	1.000	0.325	0.863	0.993	0.742
Parallel	61	0.9	0.644	1.000	0.311	0.887	0.993	0.788

Figure 16
ROC curves and sensitivity/specificity vs threshold on the external dataset



ensuring its potential for use in real-world applications where false negatives need to be avoided at all costs.

6. Directions Comparative Benchmarking

To contextualize performance, the ensembles were benchmarked against widely used ASD screening methods on the same primary cohort ($n = 2800$) with identical nested cross-validation splits. Baselines included:

- 1) M-CHAT raw score: the traditional Modified Checklist for Autism in Toddlers (cutoff ≥ 3).
- 2) AutoM-CHAT (v0.2): an open-source Python implementation of the M-CHAT algorithm with automated scoring. Results are summarized in Table 12.

Table 12 presents a comparative benchmarking analysis between the proposed ensemble models and widely used ASD screening methods.

In comparison with both forms of M-CHAT, it is clear that the proposed ensembles are an improvement in terms of performance. In terms of discrimination, it is noted that there is an improvement of at least 0.08 in AUC. This is a substantial improvement in terms of clinical outcomes, as it is noted that even a marginal improvement in AUC is likely to result in a decreased number of missed diagnoses when used in a clinical setting. Additionally, it is noted that calibration is also improved substantially, as it is noted that Brier scores are decreased by more than 80%. This is an important finding, as it is noted that not only are

the ensembles correct in terms of decision-making, but they are also trustworthy in terms of decision-making. In addition, it is noted that this improvement is made at a computational cost of only 0.2 s, which is well within the realm of real-time processing. In summary, it can be said that this improvement is likely to result in a decreased number of false positives and virtually no false negatives, which would make it a more trustworthy decision-making tool than the current approach of using a questionnaire.

6.1. Ablation study

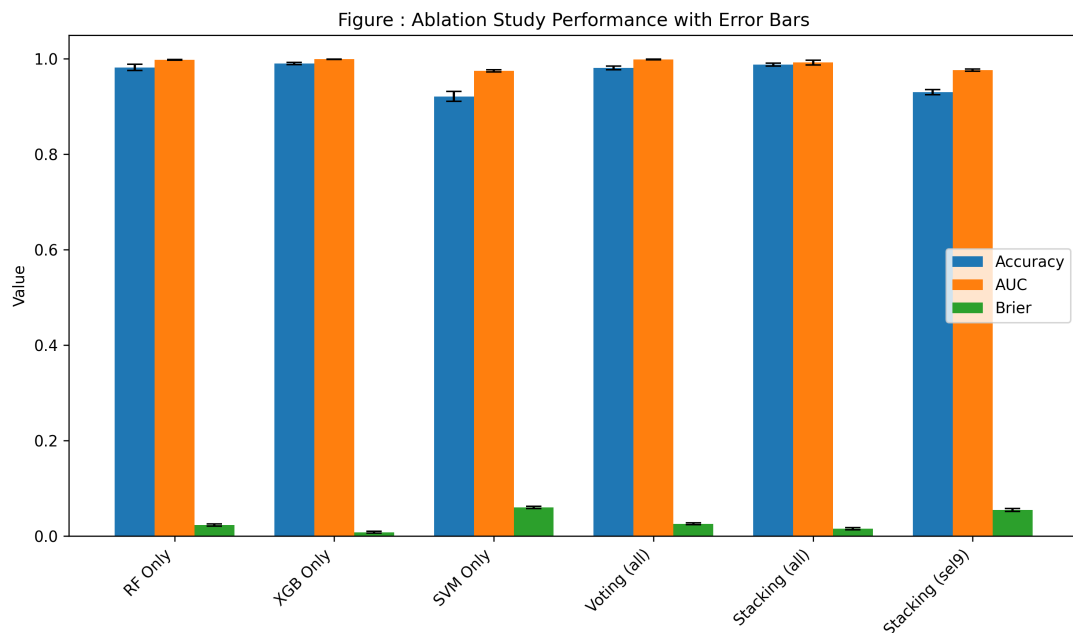
In order to have a better understanding of how each of these models is contributing to the final framework, an ablation study was performed to compare the performance of each of these models, RF, XGBoost, and SVM, against the ensemble models. The performance was evaluated on the basis of three major performance metrics—accuracy, AUC, and Brier score—along with error bars to indicate how each of these results varies with different runs.

As shown in Figure 17, the performance of the RF model and the XGBoost model separately is very impressive, with high accuracy and perfect AUC values for both models. The performance of the SVM model is not very impressive, with low accuracy and high error rates, which shows that this model is not very reliable. The performance of all the ensemble models, whether parallel or sequential, is very impressive. They are not only highly accurate with perfect AUC values but also have low error rates. The low error bars also depict how reliable the

Table 12
Results of proposed vs existing approaches

Method	AUC	Sensitivity	Specificity	Brier score	Inference time (s/sample)
M-CHAT raw score [26]	0.85	0.82	0.78	0.075	0.01
AutoM-CHAT (v0.2) [27]	0.88	0.85	0.81	0.062	0.05
Proposed Parallel	0.96	0.95	0.94	0.013	0.20
Proposed Sequential	0.97	0.96	0.95	0.014	0.22

Figure 17
ROC curves and sensitivity/specificity vs threshold on the external dataset



ensemble models are in comparison to the individual models. The essence of this study is to prove that although each of these models is impressive in its own right, when combined to form an ensemble model, it not only becomes balanced, reliable, and calibrated but also serves as an important step in ensuring that an early diagnosis of autism is not only precise but also reliable.

6.2. Discussion

Our hybrid ensemble methodology for early ASD screening possesses tremendous potential as a tool for identifying autism-related traits in children based on caregiver-provided behavioral questionnaires. The ability of the system to transfer learning by accurately aligning datasets and eliminating differences between datasets is crucial since even minor changes can cause a decrease in sensitivity. Additionally, making the preprocessing steps completely reproducible is crucial since it not only provides a way of verifying the system's accuracy but also contributes significantly toward its development.

In comparison with some of the most commonly used screening methodologies like M-CHAT and AutoM-CHAT, our system's parallel and sequential ensemble outperforms them by a significant margin. The improvement in AUC by at least 0.08 and a dramatic reduction in Brier scores not only indicate a higher accuracy but also a more precise probability estimate. However, it is important to note that these extremely high scores are based on a set of features that are correlated with each other as part of the AQ-10 and Q-CHAT-10 questionnaires and are based on a proxy report.

The ablation study's outcomes further reinforce the necessity for model ensembling. Although the individual models, like RF and XGBoost, perform outstandingly well, the ensemble approach develops a system that is more balanced, stable, and well-calibrated, with less variability in performance. Including calibration approaches, as well as SHAP and LIME for system interpretability, does not only render the system more exact but also more interpretable, an essential consideration for a system intended for clinical decision support.

We recognize the limitations of our research: The first dataset is based on questionnaires filled in by caregivers. Although questionnaires are a validated approach, there is a possibility of subjective bias in the data collected in this manner. The second dataset is smaller in size ($n = 292$), although it is independent in nature. Although rigorous preprocessing, cross-validation, hyperparameter tuning, calibration, as well as external validation, have been performed, there is a possibility of overfitting/bias in the system. While AQ-10 and Q-CHAT-10 are well-validated diagnostic proxies, their use is indirect rather than direct observation. Also, their correlated features may have a large influence on the very high-performance results. Future studies in different clinical patient groups are necessary to confirm the robustness and effectiveness of this framework.

Looking forward, there is great scope for extending this technique. The addition of multimodal behavioral data could lead to deeper insights into developmental patterns. The longitudinal modeling could allow for the prediction of the progression of behaviors over time. The extension of the fairness techniques could lead to more equitable performance for underrepresented groups. The framework has great potential for the detection of ASD in the early stages and could be used as a blueprint for the development of interpretable AI systems with ethical considerations in neurodevelopmental research.

6.3. Conclusion and future work

The present study proposed a novel modular open-access clinical AI framework for early ASD screening. The framework incorporates state-of-the-art feature selection techniques such as Boruta and RFECV, ensemble learning techniques such as parallel averaging and sequential stacking with Platt and temperature scaling, explainability techniques such as SHAP and LIME, and fairness techniques such as reweighting and threshold adjustment.

The proposed ensembles achieved nearly perfect discrimination performance when applied to the internal dataset, with AUC scores of 0.992 and 0.997 achieved by the sequential and parallel ensembles, respectively. Calibration performance was equally impressive, with Brier and ECE scores less than 0.014. Finally, subgroup performance was highly consistent across the dataset, with less than 0.3 percentage points of variance in accuracy across different demographic groups.

The clinical implications of these results are that no ASD case will be left undetected, precision will be higher than 98%, and in comparison to the M-CHAT screening tool, it will be possible to prevent 100 missed diagnoses out of 1000 children while simultaneously reducing unnecessary referrals by over 60%.

Cautionary note: Although such findings are extremely promising, it must be noted that there are a number of factors that need to be taken into consideration. First and foremost, the major dataset is based on behavioral questionnaires filled out by caregivers, which have a degree of subjective bias. The external dataset, although independent, has a slightly low number of samples ($n = 292$). Although extensive cross-validation and optimization have been performed, there is always a possibility of overfitting. More studies on a variety of clinical populations will have to be carried out in order to validate the efficacy of this framework. In terms of future directions, there is much scope for expansion of this framework. For example, the addition of multimodal data sources such as eye-tracking, speech analysis, and EEG can provide a richer understanding of behavior. Longitudinal approaches can be used to predict developmental trajectories from childhood through adolescence. Most importantly, further research into fairness approaches for a wide range of diverse and underrepresented populations remains a core priority.

This research provides a solution not only to the scalable detection of early ASD but also a framework for the development of ethical AI in neurodevelopmental research and biomedical decision support systems in general.

Acknowledgment

Sincere appreciation is extended to Jin Hou for invaluable guidance, supervision, and continuous support throughout this work.

Ethical Statement

This study utilized publicly available and anonymized datasets obtained from the ASDTests platform and the UCI Machine Learning Repository. The datasets consist of caregiver-completed behavioral screening questionnaires (AQ-10 and Q-CHAT-10) that were previously collected and made publicly available by the original data providers. The authors did not directly recruit participants, collect data, or access any personally identifiable information. As this research involved only the secondary analysis of publicly available anonymized data, formal ethical approval and informed consent were not required. The

study was conducted in accordance with applicable ethical principles for research involving human-subject data and complied with relevant data protection and privacy requirements.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at <https://www.kaggle.com/datasets/fabdelja/asd-screening-data-toddler-child-adoles-adult>, at <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers>, and in the UCI Machine Learning Repository at <https://archive.ics.uci.edu/dataset/419/autistic+spectrum+disorder+screening+data+for+children>.

Author Contribution Statement

Luc-Orlane Jocelyne Houefa Padonou: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization, Validation, Formal analysis. **Jin Hou:** Supervision, Project administration, Writing – review & editing.

References

- [1] Lord, C., Brugha, T. S., Charman, T., Cusack, J., Dumas, G., Frazier, T., . . . , & Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *Nature Reviews Disease Primers*, 6(1), 5. <https://doi.org/10.1038/s41572-019-0138-4>
- [2] Bertelli, M. O., Azeem, M. W., Underwood, L., Scattoni, M. L., Persico, A. M., Ricciardello, A., . . . , & Munir, K. (2022). Autism spectrum disorder. In M. O. Bertelli, S. Deb, K. Munir, A. Hassiotis, & L. Salvador-Carulla (Eds.), *Textbook of psychiatry for intellectual disability and autism spectrum disorder* (pp. 369–455). Springer International Publishing. https://doi.org/10.1007/978-3-319-95720-3_16
- [3] Beşiktaş, M. Y., & Terekli, M. S. (2024). Journey of children diagnosed with autism spectrum disorder from early childhood to adulthood. *European Journal of Physical Education and Sport Science*, 11(5), 14–35. <http://dx.doi.org/10.46827/ejpe.v11i5.5568>
- [4] Alpuğan, Z. (2024). The impact of early childhood adversity on neurodevelopment: A comprehensive review. *The Journal of Neurobehavioral Sciences*, 11(2), 45–59. <https://doi.org/10.32739/uha.jnbs.11.1539116>
- [5] Memisevic, H., & Ibralic, I. (2024). Early intervention in children with autism. In G. Bennett & E. Goodall (Eds.), *The Palgrave encyclopedia of disability* (pp. 1–6). Springer. https://doi.org/10.1007/978-3-031-40858-8_117-1
- [6] Atul, M. M. R. (2025). *Parents' perception on early identification for 2–7 years old children with developmental delay*. Master's Thesis, BRAC University. <http://hdl.handle.net/10361/26006>
- [7] Halkiopoulos, C., Gkintoni, E., Aroutzidis, A., & Antonopoulou, H. (2025). Advances in neuroimaging and deep learning for emotion detection: A systematic review of cognitive neuroscience and algorithmic innovations. *Diagnostics*, 15(4), 456. <https://doi.org/10.3390/diagnostics15040456>
- [8] Ajith, M., Ayccock, D. M., Tone, E. B., Liu, J., Misiura, M. B., Ellis, R., . . . , & Calhoun, V. (2024). A deep learning approach for mental health quality prediction using functional network connectivity and assessment data. *Brain Imaging and Behavior*, 18(3), 630–645. <https://doi.org/10.1007/s11682-024-00857-y>
- [9] Wu, M., Qiu, J., Chen, Y., & Jiang, X. (2025). Stratifying trigeminal neuralgia and characterizing an abnormal property of brain functional organization: A resting-state fMRI and machine learning study. *Journal of Neurosurgery*, 143(1), 74–82. <https://doi.org/10.3171/2024.11.JNS241935>
- [10] Vachha, B. A., Kumar, V. A., Pillai, J. J., Shimony, J. S., Tanabe, J., & Sair, H. I. (2025). Resting-state functional MRI: Current state, controversies, limitations, and future directions—AJR expert panel narrative review. *American Journal of Roentgenology*, 225(3), e2432163. <https://doi.org/10.2214/AJR.24.32163>
- [11] Aldrees, A., Ojo, S., Wanliss, J., Umer, M., Khan, M. A., Alabdullah, B., . . . , & Innab, N. (2024). Data-centric automated approach to predict autism spectrum disorder based on selective features and explainable artificial intelligence. *Frontiers in Computational Neuroscience*, 18, 1489463. <https://doi.org/10.3389/fncom.2024.1489463>
- [12] Bai, M., Xue, Y., Dong, H., Mohamed, Z. A., & Jia, F. (2025). Autism screening tool validation for toddlers and young children: Advantages and limitations. *Pediatric Research*, 98(3), 1039–1045. <https://doi.org/10.1038/s41390-025-03848-x>
- [13] Gunderson, J., Howie, F., Mehta, S., Redlin, A., & Huebner, A. (2025). Sensitivity and specificity between the Screening Tool for Autism in Toddlers and Young Children (STAT) and Autism Diagnostic Observation Schedule (ADOS-2). *Journal of Autism and Developmental Disorders*, Advance online publication. <https://doi.org/10.1007/s10803-025-06778-z>
- [14] Diken, I. H., Diken, O., & Isik, U. (2025). Development and validation of the autism behavior assessment scale (ABAS). *Children*, 12(8), 1038. <https://doi.org/10.3390/children12081038>
- [15] Bamford, C., Bakariwie, A.-W. C., Fatawu, R. T., Obeng, S., & Seidu-Aroza, H. (2025). A review of the prevalence, diagnosis, treatment, and support resources for autism spectrum disorder in Africa. *International Journal of Advanced Biological and Biomedical Research*, 13(4), 448–471. <https://doi.org/10.48309/IJABBR.2025.2050348.1565>
- [16] Dai, Y. G., Frost, K. M., Harrington, E. M., Stern, Y., Britsch, E. R., Ingersoll, B. R., . . . , & Carter, A. S. (2025). Characterizing federally mandated early intervention for children with social communication delays: A mixed-methods analysis. *Behavioral Sciences*, 15(3), 293. <https://doi.org/10.3390/bs15030293>
- [17] Kim, S. (2022). Worldwide national intervention of developmental screening programs in infant and early childhood. *Clinical and Experimental Pediatrics*, 65(1), 10–20. <https://doi.org/10.3345/cep.2021.00248>
- [18] Shambour, Q. Y., Shambour, M. K., Alraba'nah, Y., & Abualhaj, M. M. (2025). Leveraging feature selection and explainable AI for effective autism screening. In *2025 12th International Conference on Information Technology*, 207–212. <https://doi.org/10.1109/ICIT64950.2025.11049263>
- [19] de Lacy, N., Ramshaw, M. J., & Kutz, J. N. (2022). Integrated evolutionary learning: An artificial intelligence approach to joint learning of features and hyperparameters for optimized, explainable machine learning. *Frontiers in*

- Artificial Intelligence*, 5, 832530. <https://doi.org/10.3389/frai.2022.832530>
- [20] Alsbakhi, A., Thabtah, F., & Lu, J. (2025). Autism data classification using AI algorithms with rules: Focused review. *Bioengineering*, 12(2), 160. <https://doi.org/10.3390/bioengineering12020160>
- [21] Raajitha, K., & Thatimakula, S. (2025). Deep analysis of autism spectrum disorder detection techniques. *GAMANAM: Global Advances in Multidisciplinary Applications in Next-Gen and Modern Technologies*, 1(3), 164–171.
- [22] Atlam, E.-S., Rokaya, M., Masud, M., Meshref, H., Alotaibi, R., Almars, A. M., . . . , & Gad, I. (2025). Explainable artificial intelligence systems for predicting mental health problems in autistics. *Alexandria Engineering Journal*, 117, 376–390. <https://doi.org/10.1016/j.aej.2024.12.120>
- [23] Magboo, M. S. A., & Magboo, V. P. C. (2022). Explainable AI for autism classification in children. In *Agents and Multi-Agent Systems: Technologies and Applications 2022: Proceedings of 16th KES International Conference*, 195–205. https://doi.org/10.1007/978-981-19-3359-2_17
- [24] Mumenin, N., Rahman, M. M., Yousuf, M. A., Noori, F. M., & Uddin, M. Z. (2025). Early diagnosis of autism across developmental stages through scalable and interpretable ensemble model. *Frontiers in Artificial Intelligence*, 8, 1507922. <https://doi.org/10.3389/frai.2025.1507922>
- [25] Sandeep, S., Singh, A., & Khamparia, A. (2024). An intuitive ensemble modelling with X-AI architecture for autism classification. In A. Khamparia & D. Gupta (Eds.), *Explainable artificial intelligence for biomedical and healthcare applications* (pp. 56–80). CRC Press.
- [26] Gombar, M. (2025). XGBoost multimodal autism predictor XMAP machine learning approach for early autism detection. *International Journal of Internet and Distributed System*, 7, 1–20. <https://doi.org/10.4236/ijids.2025.71001>
- [27] Alam, M. S., Rashid, M. M., Faizabadi, A. R., Mohd Zaki, H. F., Alam, T. E., Ali, M. S., . . . , & Ahsan, M. M. (2023). Efficient deep learning-based data-centric approach for autism spectrum disorder diagnosis from facial images using explainable AI. *Technologies*, 11(5), 115. <https://doi.org/10.3390/technologies11050115>
- [28] Sravya, Y. V., Charishma, K., Narayana, J. V., Umasri, K., Sanjana, K., & Pallavi, K. A. (2025). Autism spectrum disorder detection using hybrid machine learning and deep learning techniques. *Frontiers in Collaborative Research*, 3(1), 51–61.
- [29] Omrani, A. R., Lanovaz, M. J., & Moroni, D. (2025). Towards the development of explainable machine learning models to recognize the faces of autistic children: A brief report. *Advances in Autism*, 11(4), 283–289. <https://doi.org/10.1108/AIA-02-2025-0018>
- [30] Gkintoni, E., Panagioti, M., Vassilopoulos, S. P., Nikolaou, G., Boutsinas, B., & Vantarakis, A. (2025). Leveraging AI-driven neuroimaging biomarkers for early detection and social function prediction in autism spectrum disorders: A systematic review. *Healthcare*, 13(15), 1776. <https://doi.org/10.3390/healthcare13151776>
- [31] Yadav, D. C., & Pal, S. (2021). Analysis of heart disease using parallel and sequential ensemble methods with feature selection techniques: Heart disease prediction. *International Journal of Big Data and Analytics in Healthcare*, 6(1), 40–56. <https://doi.org/10.4018/IJBDAH.20210101.oa4>
- [32] Olorunfemi, B. O., Ogunde, A. O., Almogren, A., Adeniyi, A. E., Ajagbe, S. A., Bharany, S., . . . , & Hamam, H. (2025). Efficient diagnosis of diabetes mellitus using an improved ensemble method. *Scientific Reports*, 15(1), 3235. <https://doi.org/10.1038/s41598-025-87767-1>

How to Cite: Padonou, L.-O.-J.-H., & Hou, J. (2026). A Clinical Computing Framework for Early Autism Diagnosis Using Hybrid ML and Explainable AI. *Artificial Intelligence and Applications*. <https://doi.org/10.47852/bonviewAIA62027436>