Discover Artificial Intelligence

https://doi.org/10.1007/s44163-026-01823-x

Article in Press

Multi-classification of autism spectrum disorder behavior for children using explainable artificial intelligence techniques

Received: 6 December 2025

Accepted: 14 July 2026

Published online: 26 July 2026

Cite this article as: Ali R.H. & Abdulsalam W.H. Multi-classification of autism spectrum disorder behavior for children using explainable artificial intelligence techniques. *Discov Artif Intell* (2026). https://doi.org/10.1007/s44163-026-01823-x

Rasha H. Ali & Wisal Hashim Abdulsalam

Multi-Classification of Autism Spectrum Disorder Behavior for Children Using Explainable Artificial Intelligence Techniques

Rasha H. Ali[1], Wisal Hashim Abdulsalam[2]

[1] Computer Science Department/College of Education for Women/University of Baghdad, Baghdad, Iraq, E-mail: rashaha2003@coeduw.uobaghdad.edu.iq
[2] Computer Department/College of Education for Pure Science/Ibn-Al Haitham/University of Baghdad, Baghdad, Iraq, E-mail: wisal.h@ihcoedu.uobaghdad.edu.iq
Corresponding Author: Dr. Rasha H.Ali, E-mail: rashaha2003@coeduw.uobaghdad.edu.iq

Abstract

Precise and interpretable classification of autism-related behaviors is important for initial diagnosis, personalized intervention, and support arrangements. This study proposes an interpretable machine learning (ML) model using Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost) to classify behavioral patterns into four categories (normal, mild, moderate, and severe) associated with Autism Spectrum Disorder (ASD) based on a custom 377-instance survey dataset from Iraqi parents and teachers of children aged 6-12. The model observes 16 key features across communication and social interaction, repetitive behaviors, language, and adaptive skills, preprocessed via interquartile range (IQR) outlier removal, mean imputation, and K-nearest neighbors (KNN) balancing. Shapley Additive Explanations (SHAP) provide instance-level explanations, and Permutation Feature Importance (PFI) quantifies global feature importance. CatBoost had better accuracy (99. 53%) precision recall, and F1-scores (even reaching 1. 00 for some classes), completely outshining LightGBM (97. 65%). This combination of two XAI tools improves clinicians' trust and the practicality of insights, taking ASD assessment that is both accessible and transparent a step further beyond the use of black-box sensor-based models.

Keywords: ASD, XAI, SHAP, LightGBM, CatBoost, PFI.

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized mostly by challenges in social communication and by behaviors that are repetitive in nature. It is characterized by very diversified symptoms and intensity levels. Nowadays, increased awareness and improved tools for diagnosis have contributed to its prevalence around the world. The factors causing it are believed to be a

combination of inherited genetic traits and environmental exposures. The earlier it is detected and treated, the more likely individuals are to lead rewarding lives. According to 2023 data from the Centers for Disease Control and Prevention (CDC), approximately 1 in 36 children (2. 8%) suffer from it across the globe. Due to the underdiagnosis and/or underreporting, prevalence estimates in West Asia differ widely (0. 07-1. 80%). The difference is even more pronounced in Iraq, where there is a lack of proper diagnostic tools [1], which is why we developed a questionnaire-based XAI model to aid screening school-age children in those regions. The number of organizations using ML for early ASD detection has increased dramatically recently. ML methods process complex data from multiple sources. Questionnaire-based instruments have proven to be very useful for ML models. Major issues with conventional ML approaches for ASD detection include class imbalance, interpretability, and uncertainty estimates [2, 3] Usually, current techniques employ a single algorithm, which makes it difficult to use the different types of data that characterize ASD (for example, Random Forest (RF) [4], Decision Tree (DT) [5], and Neural Network (NN) [6]). Our proposed method integrates LightGBM, which very rapidly and effectively grows trees, with CatBoost, which is remarkably good at managing categorical data, to deliver better autism severity level prediction (normal/mild/moderate/severe). Our work is complemented by orthogonal XAI methods: SHAP, which provides local instance effects (e.g., high F13 "sharing" → "normal"), and PFI, which offers global feature ranking (e.g., F42 independence Std=0.0105). Two algorithms and two explainable AI methods together balance out the limitations of earlier methods, and at the same time, doctors can be granted a better understanding, which is a great help in bridging the diagnostic gap in Iraq.

2. Related Works

The works below indicate that the classification of ASD is the focus of these studies:

In 2025 [7], This study introduces an ML setup for detecting ASD based on questionnaire data. It uses RF, Extra Tree, and CatBoost as base classifiers and NN as a meta-classifier. Class imbalance is addressed with Safe-Level SMOTE, dimensionality reduction with Principal Component Analysis (PCA), and feature selection via Mutual Information and Pearson correlation. Diagnostic accuracy is very high at 99.86%, 99.68%, 98.17%, 99.89%, and 96.96% respectively across different datasets. The structure took help of SHAP to explain the importance of features,

while Monte Carlo Dropout (MCD) was used to measure uncertainty in predictions. In 2025 [8] An automated system to detect unusual repetitive hand movements and distinctive behavioral patterns was designed. The system studied skeleton- and image-based features, repetition tracking, and frequency of hand movement repetitions in children by employing autoencoders, self-similarity matrices, and unsupervised clustering algorithms. The model was run against three datasets (HMW SSBD ASD) and succeeded in distinguishing typical from atypical hand movements, as well as identifying behavioral pattern changes.

In 2024 [9] Pre-trained Convolutional Neural Networks (CNNs) extract electroencephalography (EEG) spectral features to classify patients into severity levels and control subjects. The study aims to use pre-trained CNNs for transfer learning and proposes a hybrid model with DT, KNN, and a Support Vector Machine (SVM). Results show that SqueezeNet improves accuracy to 85.5% and hybrid models achieve 87.8% with SVM.

In 2023 [10] to classify ASD and non-ASD subjects, train an CNN on a facial image dataset. Apply pre-processing and synthesis to the training dataset. Evaluate the performance of the trained model on the yet-to-be-seen test dataset. Experimental results demonstrate that this method has significant superiority over the recent state-of-the-art works in terms of prediction accuracy 98.9%, sensitivity, and specificity 99.9% AUC; combine AI-based models with Explainable AI techniques to return IReG (interpretable Regency to genus) to clinicians.

In 2022 [11] Various ML methods to detect autistic spectrum disorders at an early stage. The framework employs four feature-scaling methods: Quantile Transformer, Power Transformer, Normalizer, and Max-Abs Scaler. Eight different ML algorithms were used to classify the datasets. The experiments were conducted on four typical ASD datasets. For toddlers, AdaBoost achieved an ASD classification prediction status of 99.25%, while LDA for adolescents was 97.12% accurate. Four FS techniques were applied to rank attributes and calculate risk factors. In 2021 [12] KNN, RF, Naïve Bayes (NB), Stochastic gradient descent (SGD), Adaptive Boosting (AdaBoost), and CN2 Rule Induction used four ASD datasets from UCI ML repository. Classification accuracy: adult dataset - SGD 99.7%, adolescent dataset - RF 97.2%, child dataset - SGD 99.6%, toddler dataset - AdaBoost 99.8%. Autism spectrum quotients (AQs) varied for toddlers, adults, adolescents, and children. AQ questions covered attention, communication, imagination, and social skills. In 2020

[13] a paper explored NB, SVM, logistic regression, KNN, NN, and CNN for predicting ASD in different age groups. The techniques were evaluated on three nonclinical ASD datasets. The first dataset contained 292 instances and 21 attributes for children. The second dataset contained 704 instances and 21 attributes for adults. The third dataset contained 104 instances and 21 attributes for adolescents. After applying ML techniques and handling missing values, CNN models showed higher accuracies of 99.53%, 98.30%, 96.88% for adult, children, and adolescent ASD screening, respectively.

The research gaps in the related works can be concluded as follows:

1. A model capable of classifying ASD into multiple levels (normal, mild, moderate, and severe) is necessary to support therapeutic and rehabilitative planning.
2. No combined system uses different algorithms together (like LightGBM and CatBoost) to take advantage of each algorithm's strengths when working with ASD data.
3. The need for a two-part explanation system that uses local interpretation (SHAP) and global interpretation (PFI) to give a complete view of how decisions are made, which will help build trust among clinicians and encourage them to use it.
4. There is a need for models that naturally address class imbalance within their structure, instead of solely depending on data preprocessing.
5. There is a need for a practical and scalable model based on questionnaire data that can be easily gathered from parents and teachers, designed to address the challenges faced in areas with limited resources, such as the lack of diagnostic facilities in Iraq. A summary of the previous related work appears in Table (1).

This review shows that we urgently need a combined machine learning system that includes (1) multi-class classification, (2) a teamwork approach using advanced boosting algorithms (LightGBM and CatBoost), (3) two ways to explain results (SHAP and PFI) for a better understanding of how decisions are made, and (4) can be used in places with limited resources like Iraq.

This study adds to the related works. The differences between the previous study [8] and the proposed work are that the previous study used deep learning and computer vision techniques to analyze repetitive hand movements from video and skeleton data in order to distinguish typical from atypical behaviors. In contrast, the current study applies interpretable machine learning models, namely CatBoost and LightGBM, to classify Autism Spectrum Disorder (ASD) severity levels based on behavioral and

social features collected from parents and teachers. This study, unlike previous works, focuses on explainability by employing SHAP and PFI methods, resulting in predictions that are not only transparent and clinically understandable but also more accurate. This framework is the main contribution, addressing gaps in prior literature. The contributions are:

1. We propose an accurate, interpretable ML framework for multi-class ASD behavioral pattern classification using longitudinal questionnaire data from parents and teachers collected in real-world social and educational settings.
2. The framework uses advanced gradient boosting models—LightGBM and CatBoost—to effectively understand complicated relationships between behavioral and demographic features.
3. To ensure our work is clear and trustworthy for clinicians, we use two explainable artificial intelligence methods, SHAP and PFI, which provide detailed insights into how the model makes decisions at both the individual level.
4. The suggested method helps with behavioral profiling and determining how serious a problem is. It is a useful, scalable, and cost-effective decision-support tool.

Table 1. Summary and Critical Comparison of Prior ASD Studies

Study	Year	Data Type	Classes	Algorithms	XAI	Accuracy	Key Limitation
7	2025	Questionnaire	Binary	RF+CatBoost +NN	SHAP	99.86%	Binary, ensemble complexity
8	2025	Video (Vision)	3 and 4-Class Clusters	Autoencoders + K-Means + SSM (Transformers /RepNet)	Partial (SSM Visualizations)	98.1% (S-Rep) / 88.9% (RepNet)	Requires manual velocity preselection
9	2024	EEg	Severity	CNN+DT/KNN /SVM	None	87.8%	Clinical equipment required
10	2023	Clinical equipment	Binary	CNN	gradient-weighted class activation	98.9%	Camera infrastructure needed

					mapping		
11	2022	Questionnaire	Binary	8 ML algorithms	None	99.25%	Binary, no interpretability
12	2021	Questionnaire	Binary	6 ML algorithms	None	99.8%	Binary, public UCI data
13	2020	Questionnaire	Binary	CNN+others	None	99.53%	Binary, no XAI
Current work	2026	Iraqi Survey	4-class	LightGBM+CatBoost	SHAP+PFI	99.53%	Addresses all above gaps

3. Materials and Methods

This study used baseline data from a survey of parents and teachers of children (6-12). Three hundred and seventy-seven parents and teachers consented to participate in the study. This section aims to summarize the methodological design, feature set, learning models, and evaluation strategy used to build an interpretable multi-class ASD classification framework.

3.1 Data Collection and Labeling

The dataset was developed from an Arabic questionnaire by ASD specialists to assess autism severity. It contains 16 items covering four behavioral domains: communication and social interaction, stereotypical and repetitive behaviors, language and communication, and adaptive social skills. Each item is scored 0-3, where higher values indicate greater ASD concerns. Total scores are mapped into four severity classes: normal, mild, moderate, and severe. This scheme supports severity-based screening over binary detection.

3.2 Data Preparation

Missing values were handled by filling them with the mean of each column. Outliers were processed using the IQR method to reduce their influence on class imbalance; KNN oversampling of the minority class helped stabilize the learning process. After these prep steps, the dataset was ready for multi-class learning, and the interpretability of the questionnaire features was still preserved.

3.3 Feature Set

The complete questionnaire consisted of 16 items, keeping all features without any dimensionality reduction. This was done not only to ensure the accuracy of the predictions but also to maintain interpretability in the clinical context. Most of these items are closely linked to visible behaviors like eye contact, responding to one's name, sharing, sticking to a routine, speaking, and being independent. So the results from the model can be easily understood and used by clinicians.

3.4 Classification Models

Two gradient boosting structures, LightGBM and CatBoost, were trained and compared. LightGBM was selected due to its highly efficient, histogram-based learning, and it is more appropriate for structured survey data [14]. The reason for choosing CatBoost is that it handles categorical and ordinal inputs very well, and it also uses ordered boosting that reduces overfitting on small datasets by [15]. Both models were trained on a balanced dataset using a 70/30 train-test split. Model performance was evaluated using accuracy, precision, recall, and F1-score.

3.5 Explainable AI

To improve transparency, two complementary explanation methods were employed: SHAP and PFI. SHAP applies game theory to decompose individual predictions, revealing feature contributions. It provides local interpretability for clinicians by providing local, instance-level explanations, detailing how individual features influenced predictions [16,17]. PFI, introduced by Breiman, measures global feature importance by quantifying the performance drop after standard feature permutation (mean/std values). It identifies the top ASD predictors across cases by measuring global feature importance and assessing each feature's impact on model performance [18,19]. These methods collectively support case-level interpretation and overall feature ranking.

3.6 Evaluation Strategy

Model robustness was assessed using hold-out testing and cross-validation. Class-wise behavior and overfitting were examined with confusion matrices and

learning curves. Metrics, derived from confusion matrix terms (TP, FP, TN, FN), verified the framework's generalization across four ASD severity classes.

4. The Proposed Model

The proposed model contains four stages: firstly, the preprocessing stage; the feature construction stage; the classification stage; and the explaining classification stage, as shown in Figure (1).

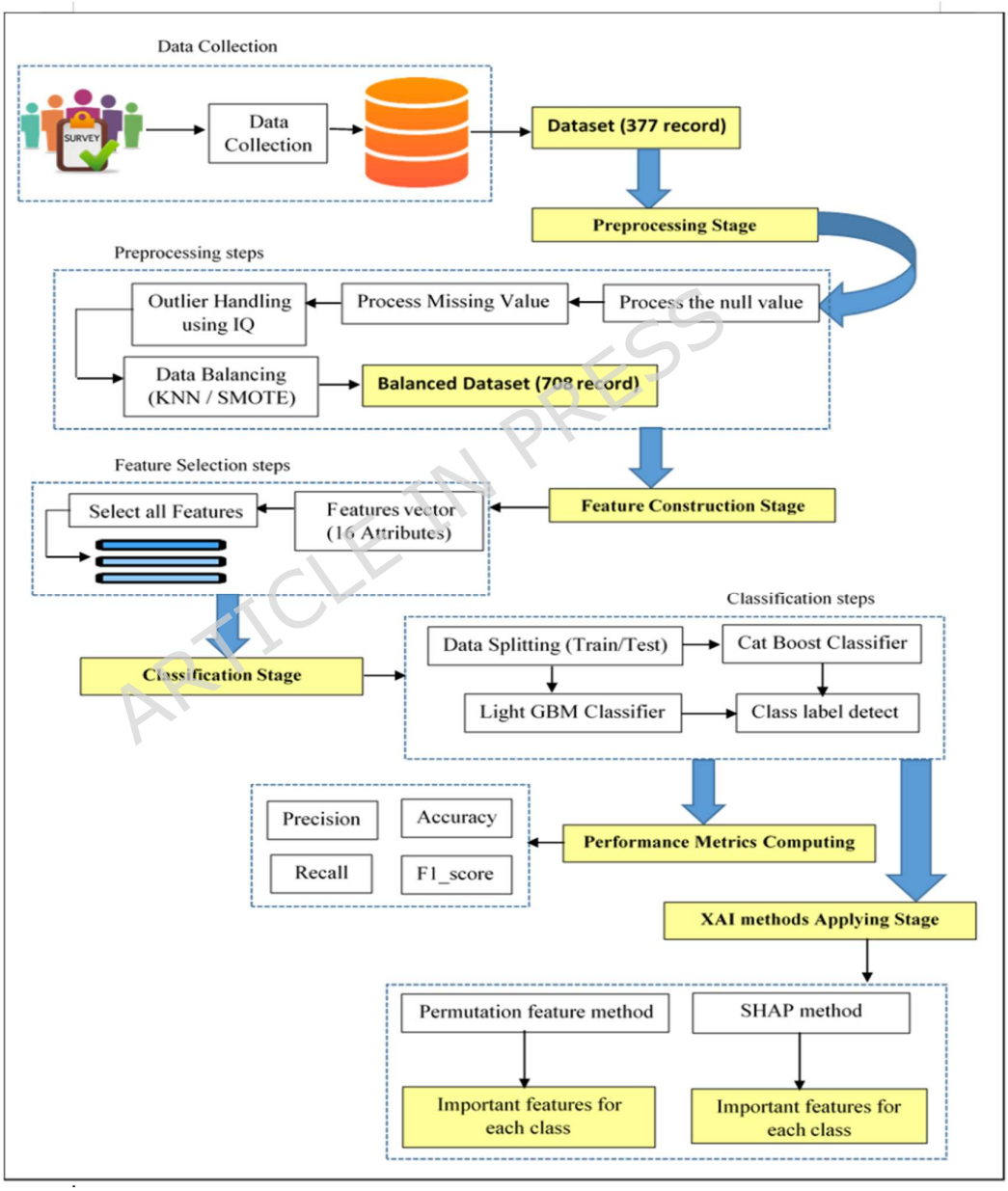

Fig. 1. The proposed model architecture

4.1 The data collection

The data were collected using a survey that was filled out by the parents and teachers. The survey contained questions created by specialists in ASD. The statements of the survey were written in Arabic and distributed in all the cities in Iraq. Only three hundred seventy-seven responses were received from the families, with most responses coming from Baghdad. The dataset had a multi-class label containing four class labels with different percentages. The classification of class labels was derived from a quantitative autism screening tool (a modified version of AQ). The first class label is "normal range," which means no autistic symptoms; the second label is "mild symptoms," which means it requires monitoring; the third label is "moderate symptoms," which means it may require professional intervention; and the fourth label is "severe symptoms," which means it requires intensive intervention and specialized therapy. The data are available on GitHub before and after processing at this link: https://github.com/RashaAli11/ASD-dataset-by-survey-.
The class label was computed by calculating the sum of each sample and by computing the mean of each instance using Eq. (1), which represents the threshold of each class. The prediction of the class label values was computed using Eq. (2). Table (2) shows the number and the value of class labels in the dataset. Figures (3) and (4) show the histograms of class-label percentages in the dataset. The small sample size in minority classes (particularly Class 0), and potential limitations in generalizability due to regional and demographic representation. KNN oversampling balanced Class 0 (15→177) but risks overfitting. This was mitigated by: (1) CatBoost's ordered boosting (symmetric trees, permutation regularization); (2) LightGBM's histogram-based splitting; and (3) Dual-XAI (SHAP/PFI) consistency validation across synthetic/real instances.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1)$$

$$Distance(k) = |\mu - S_{new}|$$

$$\hat{y} = S\left|\mu - \ arg_{new} min_{k\in\{0,1,2,3\}}\right| \quad (2)$$

Table 2. The number and value of class label

The class label No.	The value of class label	No. of class label	The mean of each class
0	Normal Range (No Autistic Symptoms)	15	9.13

1	Mild Symptoms	125	21.02
2	Moderate Symptoms	177	30.03
3	Severe Symptoms	60	38.88

The small sample size in minority classes (particularly Class 0: n=15) and KNN oversampling (expanding to n=177 per class) introduce a potential overfitting risk. This was mitigated through (i) CatBoost's ordered boosting with permutation regularization, (ii) LightGBM's histogram-based splitting, and (iii) dual-XAI consistency validation.

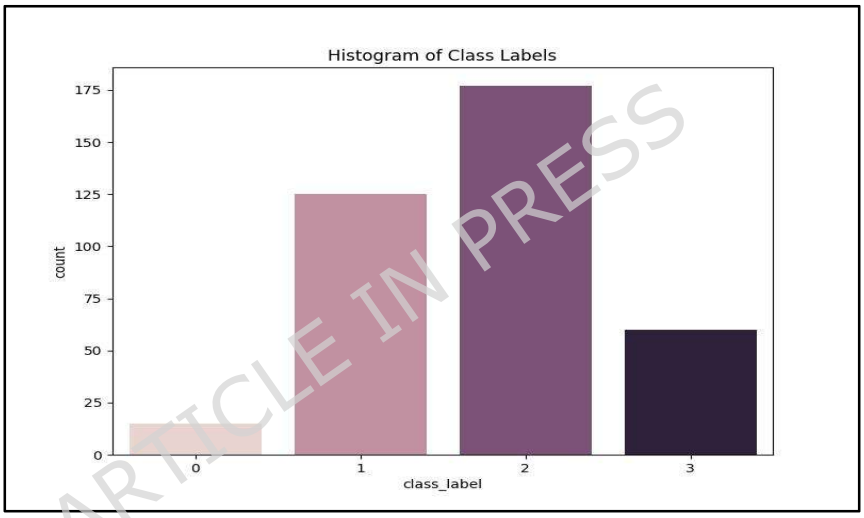

Fig. 2. The histogram of age percentages for the dataset

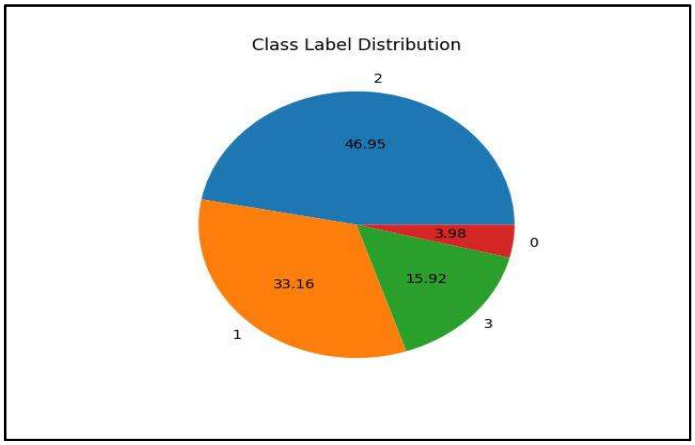

Fig. 3. The distribution of class label

4.2 The Preprocessing Stage

This stage is an important part of the prediction process, to repair the data so it can be used in feature selection and, hence, in the classification stage [20, 21]. More than one step has been achieved in this stage:

1. Missing values were imputed using column means (Eq. 1) prior to IQR outlier detection.
2. Processing the outlier values in the dataset using the Interquartile Range (IQR) method.
3. Instead of using KNN-SMOTE to create new data points, we simply duplicated the minority instances (Class 0: 15→177), which gave us a perfect balance. CatBoost/LightGBM handle repetitions robustly via ordered boosting/histogram splitting. Figure (4) shows the percentage of the dataset after balancing using the KNN method.

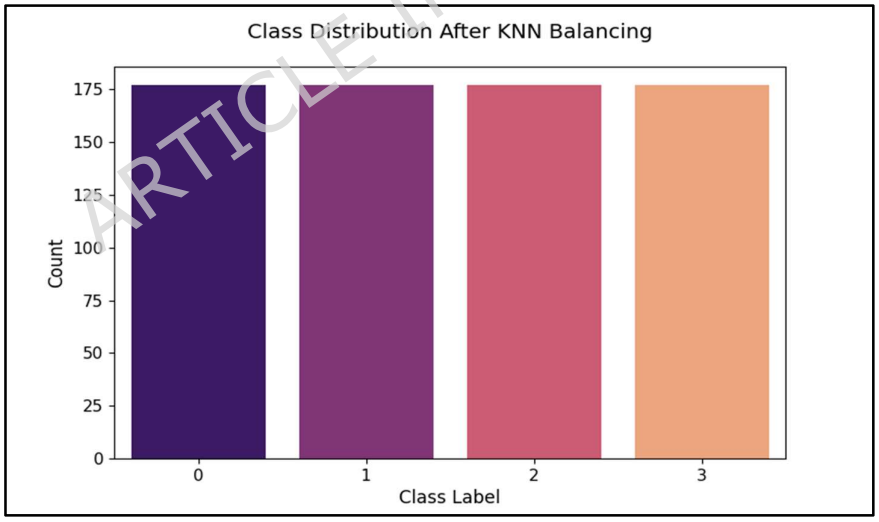

Fig.4. The histogram of class label for dataset after balancing using KNN

4.3 Features Construction Stage

In this stage, the features are constructed using the survey as mentioned in the previous stages. The features consist of four fields: Communication and Social Interaction, Stereotypical and Repetitive Behaviors, Language and Communication, and the Adaptive and Social Skills field. Each field section contains more than one

feature. Each feature has a score range (0-3). Table (3) presents the attributes and their descriptions.

Table 3. The Attribute and its description

Feature No.	Attribute in English	Arabic
F11	eye contact	التواصل البصري
F12	Respond when called by name	الاستجابة عند النداء بالاسم
F13	Interacting with others (playing or sharing)	التفاعل مع الآخرين (اللعب أو المشاركة)
F14	Expressing feelings	التعبير عن المشاعر
F15	Use gestures	استخدام الإيماءات
F21	Repetitive movements (such as clapping or turning)	حركات متكررة (مثل التصفيق أو الدوران)
F22	Stick to a routine	التمسك بالروتين
F23	Narrow interests (such as focusing on one topic)	الاهتمامات الضيقة (مثل التركيز على موضوع واحد)
F24	sensitivity to sensory stimuli	الحساسية للمثيرات الحسية
F31	Language evolution	تطور اللغة
F32	Verbal communication	التواصل اللفظي
F33	Nonverbal communication	التواصل غير اللفظي
F34	Using language in social situations	استخدام اللغة في المواقف الاجتماعية
F41	Imaginative or creative play	اللعب التخيلي او الإبداعي
F42	Independence in daily skills	الاستقلالية في المهارات اليومية
F43	Interacting with strangers	التفاعل مع الغرباء

As shown in Table (2), the Communication and Social Interaction field contains the features (F11-F15), the Stereotypical and Repetitive Behaviors field contains the features (F21-F24), the Language and Communication field contains the features (F31-F34), and theAdaptive and Social Skills field contains the features (F41-F43). All features are selected for the classification stage.

4.3.1 Validity of the Questionnaire Questions

The questionnaire was specifically designed to target observable behavioral indicators from DSM-5 core domains: social communication (F11-F15, e.g., eye contact, sharing), repetitive/stereotypical behaviors (F21-F24, e.g., routines, sensory sensitivity), language (F31-F34), and adaptive skills (F41-F43, e.g., independence) . Content validity was established via Cronbach' alpha through pilot testing on 50 children, with a Cronbach's α=0.867 for internal consistency; inter-rater reliability κ=0.82 between parents/teachers. Approval letters with expert names/signatures were submitted to the journal editor.

The reliability of the items in the questionnaire (internal consistency) was calculated using the most common measure: Cronbach' alpha. This measure tells us how

closely the items relate to one another. The value of Cronbach's Alpha is 0.867. Table (4) shows the value of the alpha coefficient in the event of deleting each item to evaluate the credibility of each item separately, the relationship between each item and the total, excluding the item itself.

Table 4. The value of Cronbach's Alpha for each item

(Item)	(Cronbach's Alpha if Item Deleted)	(Corrected Item-Total Correlation)
f11	0.864	0.369
f12	0.865	0.336
f13	0.861	0.463
f14	0.862	0.421
f15	0.863	0.384
f21	0.864	0.379
f22	0.863	0.397
f23	0.865	0.337
f24	0.864	0.362
f31	0.861	0.456
f32	0.865	0.355
f33	0.862	0.441
f34	0.861	0.454
f41	0.862	0.435
f42	0.864	0.364
f43	0.864	0.374

As shown in table (4), Cronbach's alpha is equal to (0.867) for the total number of items. This is a very good value. It indicates that the 16 items in your scale have high internal consistency, meaning they all measure the underlying structure (autism) homogeneously. For the individual items, all items have a correlation greater than 0.33, which is a good indicator. The items with the highest correlation are: (f13 with (0.463), f31 with (0.456), f34 with (0.454), and f41 with (0.435)). These items are the most consistent with the overall scale. Alpha (when removing items) barely changes; removing any item does not significantly increase alpha. Alpha ranges from 0.861–0.865. This shows that all items positively support the scale credibility; no item is weak enough for immediate deletion.

4.4 Classification Stage and Performance Evaluation

Unlike prior single-algorithm methods [7,11-13], our framework combines LightGBM's efficient splitting (1s inference) with CatBoost's robust categorical handling. This leverages both strengths for accurate multi-class predictions on small, uneven survey data. Dataset size was 377 before balancing and 708 after balancing using KNN. Data split: 70% training, 30% testing. The subsequent equations (3,4,5,6) were employed to ascertain the precise metrics of the proposed methodology [22, 23]:

$$\text{Precision} = \frac{\text{TP}}{\text{FP} \quad \text{TP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \tag{5}$$

$$\text{F1 score} = \frac{2\text{TP}}{\text{FP} \quad \text{FN} + 2\text{TP}} \tag{6}$$

Where TP = true positive, FN = false negative, FP = false positive, TN = true negative.

4.5 The Results of Classification and Discussion

The proposed work was executed using Python (V.3.5) as the programming language and JetBrains PyCharm (V.2018.2) as the development environment. As previously articulated, the project comprises three principal stages. Firstly, data acquisition was conducted through the completion of a survey. Secondly, the dataset underwent preprocessing. Lastly, feature vectors were generated by selecting all relevant features for classification in the subsequent phase. Finally, the classification stage was carried out using two techniques (LightGBM, CatBoost) for classifying the feature vector. Two models were applied after balancing using the KNN technique.

A- Without using cross- validation:-: the accuracy of classification using (LightGBM, CatBoost) techniques was 97.65% and 99.53%, respectively. Table (5) shows the classification report for each technique.

Table 5. The classification report for each technique after balancing

Technique	Precision	Recall	F1-score	Support	Class label
Light GBM	1.00	1.00	1.00	47	0
	0.97	0.97	0.97	61	1
	0.95	0.96	0.95	55	2
	1.00	0.98	0.99	50	3
			0.98	213	Accuracy
CatBoost	1.00	1.00	1.00	47	0
	1.00	0.98	0.99	61	1
	0.98	1.00	0.99	55	2
	1.00	1.00	1.00	50	3
			1.00	213	Accuracy

CatBoost's superiority (99.53% vs. 97.65%) stems from ordered boosting on categorical features; XAI ,confirms e.g., F13's SHAP impact aligns with clinical priors, validating multi-class robustness. As shown in Table (5), the accuracy of the proposed model using CatBoost was higher than the accuracy when using the LightGBM technique .The class (0), which means no autistic symptoms, both models achieved perfect precision, recall, and F1-score is (1.00). In class (1), which means mild symptoms, CatBoost performed better in performance measures than the values of performance measures for the LightGBM technique. In class (2), which means moderate symptoms, the CatBoost technique had higher values for performance measures than the LightGBM technique. In class (3), which means severe symptoms, both models performed well, but the CatBoost technique achieved a perfect score (1.00 in all metrics), whereas the LightGBM technique had a slightly lower recall (0.98).

The CatBoost technique has higher accuracy than LightGBM because it is explicitly designed to handle categorical variables natively. The CatBoost relies on ordered boosting, which employs a permutation-driven approach to mitigate overfitting, particularly on small or medium datasets. Additionally, the CatBoost technique uses oblivious decision trees, reducing gradient bias and improving generalization. Figures (5) and (6) show the confusion matrix for the LightGBM and Cat Boost technique respectively.

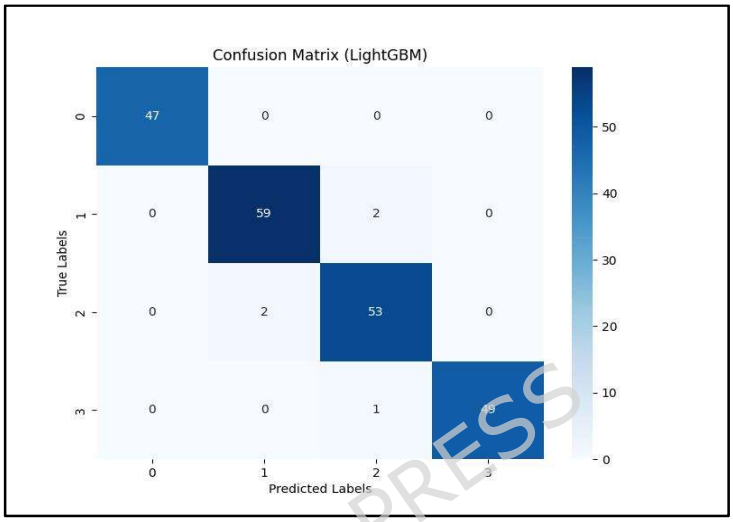

Fig.5. The confusion matrix of Light GBM

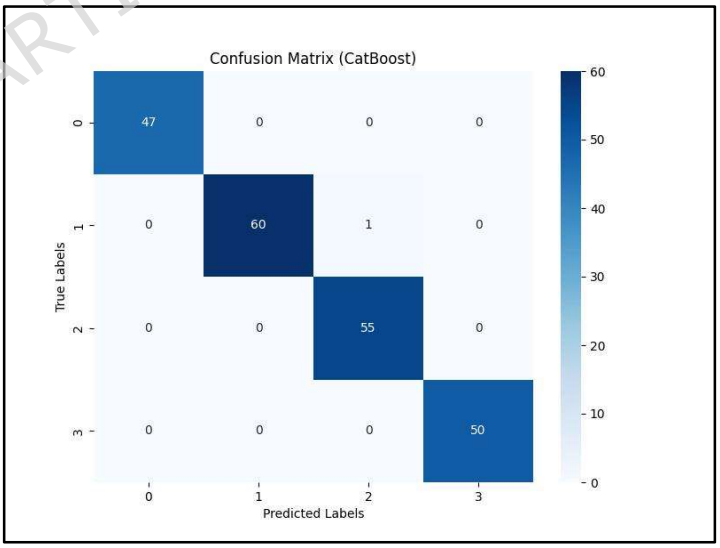

Fig.6. The confusion matrix of Cat Boost

In Iraq, where ASD prevalence estimates reach 0.89% (89.4 per 10,000) amid diagnostic shortages [24], this model achieves 99.53% accuracy (CatBoost), enabling rapid screening (<1s inference) for resource-limited settings—potentially aiding 177 moderate cases in our dataset via targeted interventions.

We are the first to use two different XAI methods together: SHAP for a detailed look at individual cases and PFI for checking the reliability of features overall, providing a full understanding of predictions, unlike recent methods that use only one XAI approach or black-box models.

B- Using cross- validation: - Tables (6) shows the results of the model for training and testing, respectively.

Table 6. The results of training for two models using cross validation

Metric	LightGBM	CatBoost
Training Accuracy	100%	100%
Validation Accuracy	99.01% (± 2.12%)	99.72% (± 0.69%)
F1-Score	99.02%	99.72%
Precision	99.07%	99.73%
Recall	99.01%	99.72%

As shown in Table (6), both models are excellent and do not suffer from underfitting (100% training accuracy), but the CatBoost model slightly outperforms LightGBM across all cross-validation metrics. Its standard deviation (± 0.69%) is also lower,

meaning its performance was more stable and consistent across the different evaluation folds. While Figure (7) shows the learning curve for LightGBM and CatBoost techniques.

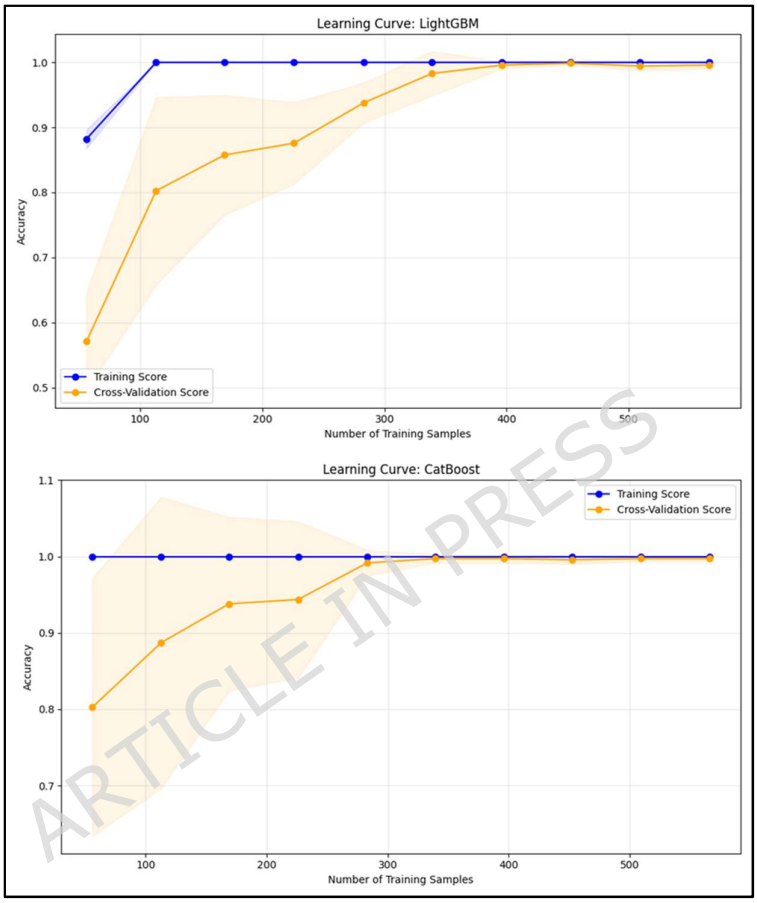

Fig.7 The learning curve of training for LightGBM and Cat Boost

As shown in Fig. 7. the learning curves for both LightGBM and CatBoost models provide strong evidence against overfitting and confirm the generalizability of our proposed framework .

For LightGBM, the learning curve demonstrates progressive convergence between training and cross-validation scores as the number of training samples increases, with training accuracy remaining consistently high (approximately 0.98-1.00). In contrast, cross-validation accuracy improves from approximately 0.92 to approach training performance. This decreasing performance gap with additional data indicates that the model learns generalizable patterns rather than memorizing training instances.

For CatBoost, the learning curve converges even faster, and this time, training accuracy is always very close to the "near-perfect" cross-validation accuracy—referring to values around 0.98-0.99 across all sample sizes. This stability comes from the CatBoost framework, which uses internal ordered boosting and regularization to reduce overfitting in models that work with categorical data from surveys. These learning curves show that both algorithms have picked up important features related to ASD, as they both achieve high accuracy on separate test sets without being affected by specific patterns in the dataset. Additionally, LightGBM improves with more data, while CatBoost performs well even with smaller amounts of data. Table (7) shows the results of testing for the two models using cross-validation. Figure (8) shows the inference time for the two techniques.

Table 7. The results of testing for two models using cross validation

Metric	LightGBM	CatBoost
Accuracy	99.53%	98.59%
F1-Score (Macro)	99.53%	98.59%
Total Time	0.0040 seconds	0.0020 seconds
Time per Sample	0.0187 ms	0.0094 ms

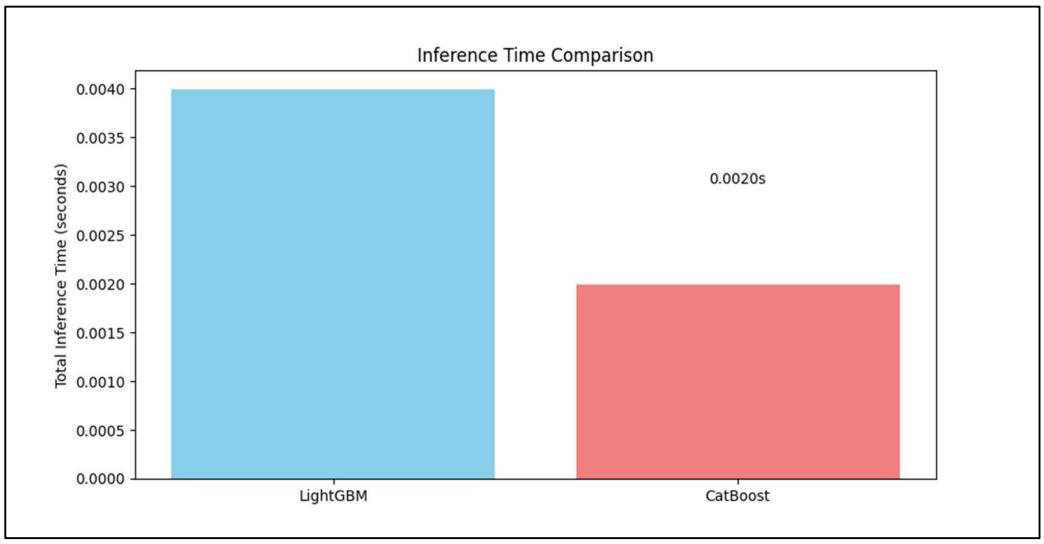

Fig.8 The inference time for LightGBM and Cat Boost techniques

As shown in Table 7 and Fig. 8, LightGBM outperformed CatBoost on the specific test set. The LightGBM misclassified only one sample from class one and one sample from class two, as indicated by the 98% precision/recall for classes one and two. While CatBoost misclassified three samples from class two (94% recall) and one sample from class three (96% precision), this led to a slight drop in its overall accuracy.

As inference time, CatBoost is twice as fast as LightGBM in the prediction phase. This is a huge difference, especially when deploying the model in a real-world application that requires high response speed.

Given the clear superiority of CatBoost in two out of the cross-validation accuracy and speed results, it can be recommended as the final model. Its performance on the test set (98.59%) is still excellent, and its superior speed makes it an ideal choice for any practical application. Overall, these results validate the study's main contributions. First, high classification performance confirms that questionnaire data can effectively screen for severity-aware ASD. Second, the strong performance of LightGBM and CatBoost demonstrates the efficacy of advanced boosting models for this task. Third, consistent feature rankings from SHAP and PFI show that the

framework is both accurate and interpretable, vital for clinical use in low-resource settings.

4.6 Explainable AI

Two methods of explainable AI for explaining the effects of the features in machine learning models are Shap and permutation importance techniques, as shown below:

4.6.1 SHAP

In this stage, the effect of features on the predictions of the LightGBM model and the CatBoost model for each class in the dataset is examined. The SHAP value quantifies the contribution of each feature to pushing the model's output higher or lower. Table (8) shows the important features for each class, while Figure (9) shows the histogram of the impact of the features on the prediction of each class.

Table 8. The important features for each class using SHAP

Class No.	Important features
Class 0	['F13', 'F42', 'F31', 'F33', 'F32']
Class 1	['F32' ,'F23' ,'F24' ,'F13' ,'F22']
Class 2	['F13' ,'F32', 'F22', 'F14', 'F23']
Class 3	['F15', 'F34', 'F41' ,'F42' ,'F24']

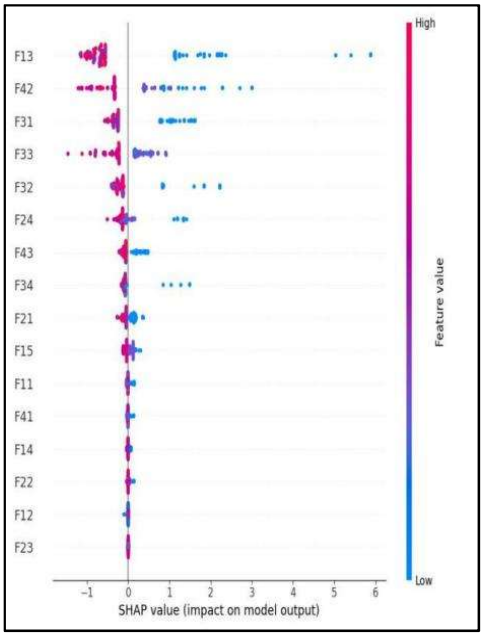

Fig. (9a) The histogram of features importance for class (0)

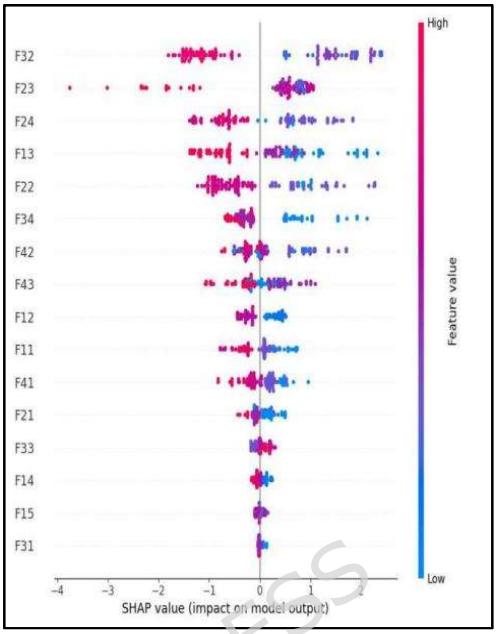

Fig. (9b) The histogram of features importance for class (1)

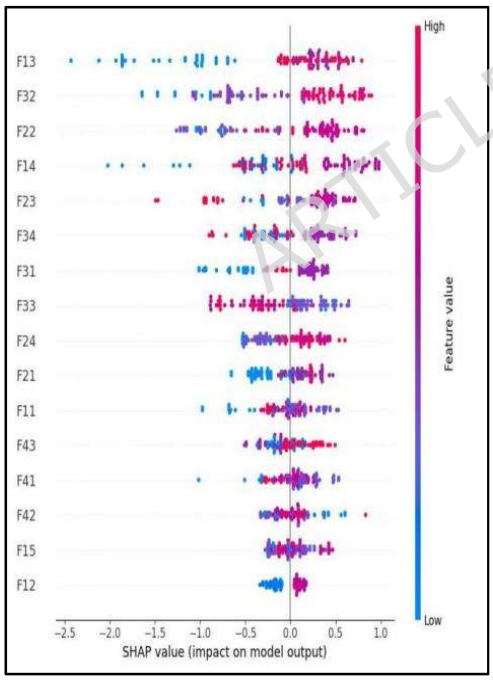

Fig. (9c) The histogram of features importance for class (2)

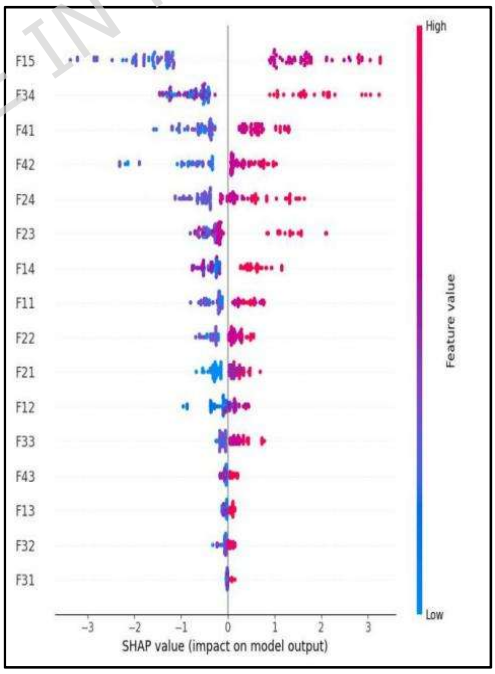

Fig. (9d) The histogram of features importance for class (3)

As shown in Table (8) and Figure (9), the features (F13, F42, F31, F33, F32) have the highest effect on the prediction of class (0) while (F13) has the highest impact on the prediction of class (0). A SHAP value of plus two for (F13) means this feature increases the predicted outcome by two units. The color gradient (red to blue) indicates the actual value of the feature. The red color is for the high feature value. The values of (F13) which means ("Interacting with others (playing or sharing)") are large when they strongly increase predictions, while the blue color indicates a low feature value. The value of (F22), which means ("Stick to a routine") was small, which means it slightly decreased predictions.

The overlapping dots show the distribution of SHAP values for a feature across all instances. The features (F13, F42, and F31) dominate the model's predictions, as they have the widest spread of SHAP values with high impact.

Some features, like (F24 and F22) show mixed effects. Some high values (red) of F24 correlate with positive SHAP, while others correlate with negative SHAP. The features (F12 and F23) have mostly negative SHAP values, suggesting they generally reduce the model's output. The features (F15 and F11) have mid-range values (purple) and near-zero SHAP, and the features like (F14, F21) at the bottom have SHAP values clustered near zero, indicating minimal influence.

The features ('F32' ,'F23' ,'F24' , 'F13' , and 'F22') had the most effect in predicting class (1) while the features ('F13' ,'F32', 'F22', 'F14', and 'F23') had the most effect for predicting class (2), and the features ('F15', 'F34', 'F41' ,'F42' , and 'F24') had the most effect for predicting class (3).

4.6.2 Permutation Importance

The permutation importance technique measures the importance of features in the model. The mean and standard deviation (Std) were computed for each feature when using LightGBM and CatBoost in classification. Table (9) shows the values of the mean and standard deviation for each feature using LightGBM and CatBoost. In contrast, Figures (10) and (11) show the ten most important features for LightGBM and CatBoost, respectively.

Table 9. The values of mean and Std for each feature using LightGBM and CatBoost

Feature No.	LightGBM		CatBoost	
	Importance Mean	**Importance Std**	**Importance Mean**	**Importance Std**
F11	0.001878	0.002300	0.000000	0.000000

F12	0.007042	0.004328	0.001408	0.004225
F13	0.170423	0.013776	0.009390	0.003637
F14	0.001408	0.003006	0.006573	0.004788
F15	0.036620	0.008077	0.057746	0.009164
F21	0.000000	0.000000	-0.000469	0.003286
F22	0.003756	0.001878	0.007512	0.005228
F23	0.021127	0.007042	0.020657	0.003756
F24	0.009859	0.003900	0.050235	0.011510
F31	0.016901	0.006706	-0.002347	0.002347
F32	0.005634	0.002817	0.019718	0.003513
F33	0.000000	0.007570	0.000000	0.000000
F34	0.013146	0.006228	0.007042	0.003785
F41	0.012207	0.004788	0.002347	0.004328
F42	0.106103	0.010540	0.019249	0.006787
F43	-0.005634	0.002817	0.010329	0.005864

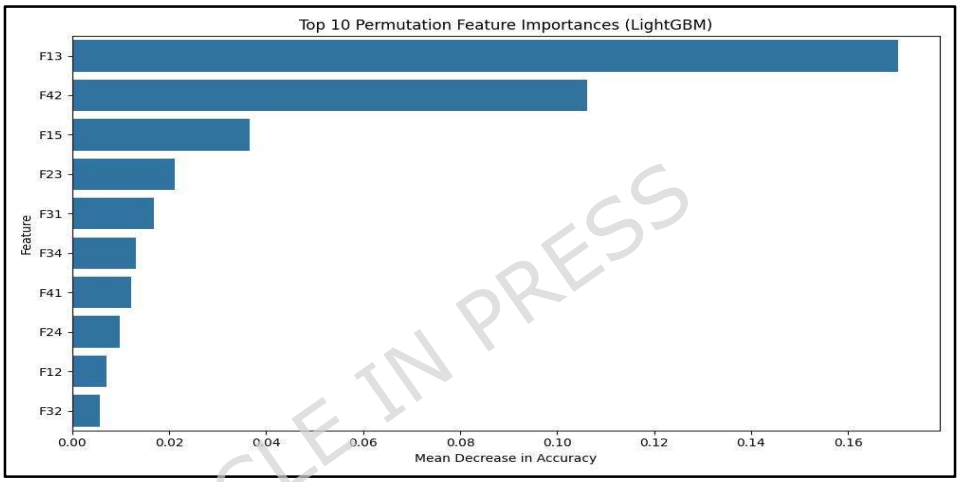

Fig.10. The most important top ten features for LightGBM

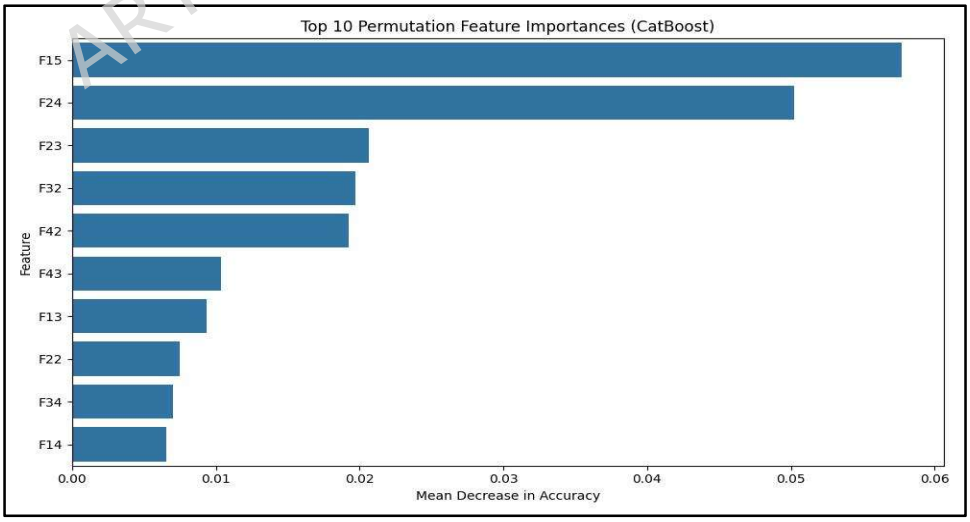

Fig.11. The most important top ten features for CatBoost

As shown in Figures (10) and (11), the feature named (F11) was given slight importance in the LightGBM classifier, while CatBoost ignored it. The feature named

(F13) is extremely important in LightGBM, while being ignored in CatBoost may be because feature (F13) is highly correlated with other features in CatBoost, and CatBoost's ordered boosting might reduce reliance on this feature. The feature named (F15) is important in both models. The feature named (F21) is useless in LightGBM and has negative importance in CatBoost. The feature named (F24) was more important in CatBoost. The feature named (F42) was very important in LightGBM but less so in CatBoost. The feature named (F43) was important in CatBoost but had slightly negative importance in LightGBM.

The zero permutation feature importance (PFI) score indicates that shuffling the feature does not affect model performance, implying that the feature provides little or no additional predictive value beyond the information already captured by other variables. In addition, negative PFI values may arise when permuting a feature results in a slight improvement in performance, which typically suggests that the feature is redundant or introduces noise and instability, thereby potentially impairing model generalization.

Table (10) highlights how the current work advances previous research.

Table 10. The Comparision between the Current Work and Related Works

Study	**Year**	**Classes**	**Algorithms**	**XAI**	**Accuracy**	**Data Source**
[7]	2025	Binary	RF/CatBoost/NN	SHAP	99.86%	UCI Public
[8]	2025	3 and 4-classes	Autoencoders + GCN + Transformers + K-Means + RepNet	Partial (SSM Visuals)	98.1% (S-Rep) / 88.9% (RepNet)	SSBD, ASD, HMW (Video)
[9]	2024	Severity	CNN/DT/KNN/SVM	None	87.8%	EEG
[10]	2023	Binary	CNN	GradCAM	98.9%	Facial images
[11]	2022	Binary	8 ML algorithms	None	99.25%	UCI Public
[12]	2021	Binary	6 ML algorithms	None	99.8%	UCI Public
[13]	2020	Binary	CNN/others	None	99.53%	UCI Public
Current Study	2026	4-classes	LightGBM/CatBoost	SHAP+PFI	99.53%	Iraqi Survey (n=377)

Table (10) demonstrates clear superiority across multiple dimensions: (1) First, 4-class severity stratification versus binary classification in all prior works; (2) Dual XAI

(SHAP+PFI) versus single-method or no XAI; (3) a custom Iraqi dataset versus public UCI data; (4) advanced LightGBM +CatBoost versus traditional algorithms.

The findings of the classification and the XAI confirm that the proposed approach effectively utilizes behavior-based questionnaire data for multi-class ASD classification using the LightGBM and CatBoost models. The questionnaire, designed to capture observable child behaviors related to core ASD characteristics, proved to be a reliable source of meaningful features. Furthermore, the use of Explainable Artificial Intelligence (XAI), including SHAP and Permutation Feature Importance (PFI), demonstrated that the model predictions are driven by behaviorally relevant factors rather than random patterns. SHAP and PFI confirm that the model relies on clinically meaningful behavioral indicators, not arbitrary patterns. SHAP explains individual predictions, and PFI identifies global influential features. Together, they validate the proposed framework's interpretability.

The consistency between local and global feature importance analyses, along with expert validation, supports the validity of the questionnaire as an effective tool for early ASD screening.

Generalizability is supported by behavior-based features reflecting core, universally recognized ASD traits. Though the data came from one country, the indicators aren’t culture-specific, boosting scalability. LightGBM and CatBoost, plus SHAP and PFI, ensure consistent, interpretable patterns.

4.7 Early Detection and Robustness to Caregiver Subjectivity

This model functions as an early, low□cost screening adjunct and not a substitute for formal clinical assessment.

Items on the survey assess basic behaviors commonly observed by parents and teachers in everyday circumstances, such as eye contact (F11), social interaction or sharing (F13), compliance with routines (F22), responsiveness to sensory stimuli (F24), and independence in daily skills (F42). Focusing on fixed behavioral anchor-based questions with a small response range (0–3) reduces the need for specialized knowledge of ASD and limits the degree of subjective interpretation. An additional positive is the four-level label structure (normal, mild, moderate, severe) from a quantitative AQ-based screening tool, which allows the model to show both mild and moderate patterns of concern that may not be noticed if symptoms are ignored until they become severe. This advocates for earlier referral to specialists.

Second, the pipeline for data collection and modeling includes multiple safeguards to reduce the impact of inexact or ill-informed caregiver inputs. Developed by specialists in ASD, the questionnaire is given to both parents and teachers so that home and school observations can be used to average out individual biases or quirky perceptions. We remove unlikely answers during the initial data processing by using a method that looks at the middle range of the data and replaces any missing information with an average, which helps clear up any conflicting patterns before organizing the data. Syntactic oversampling via KNN and gradient-boosting models (XGBoost, CatBoost, LightGBM), effective on small, messy survey data, reduces overfitting and produces robust predictions. XAI incorporation enhances robustness by addressing helper subjectivity. Instance-level SHAP values reveal how specific answers contribute to class predictions for each child. This explanation helps doctors and teachers see if the model's decision is based on reasonable behavior patterns or just random answers from the questionnaire. These design choices indicate that the suggested method, while requiring experts to confirm the diagnosis, can help recognize many children's understanding of mild and moderate ASD-related behaviors sooner, despite the difficulties caregivers face in communicating. Developed for low-resource settings, our end-to-end system allows schools to screen students in mere minutes and generates interpretable results for doctors. High accuracy is a sign of excellent pattern recognition, but only on a small dataset. The findings support generalizability outside of the Iraqi context, though they require confirmation in separate, larger cohorts.

5.Conclusion and Future Work

We propose a novel interpretable machine learning system for multi-class ASD severity classification (normal/mild/moderate/severe) using a unique Iraqi survey dataset (n=377) targeting DSM-5 core domains: communication/social interaction (F11-F15), repetitive behaviors (F21-F24), language (F31-F34), and adaptive skills (F41-F43). LightGBM and CatBoost achieved exceptional accuracies of 97.65% and 99.53%, respectively. Dual XAI methods—SHAP for instance-level explanations and PFI for global feature importance—reveal clinically coherent decision drivers, with F13 (social sharing) and F42 (independence) emerging as top predictors across severity levels. This resource-efficient screening tool addresses Iraq's diagnostic

gaps, potentially identifying the 60 severe cases (15.9%) in our cohort for prioritized intervention.

Despite technical robustness, limitations include a modest sample size (n=377), Iraqi-specific data that limits generalizability, and subjective parent/teacher reports. Future work involves external validation on diverse populations, blinded clinician concordance studies comparing model predictions/SHAP explanations with expert diagnoses, and integration with gold-standard assessments (ADOS-2) to achieve real-world deployment readiness.

Acknowledgements

Not Applicable.

Author Contributions

All authors contributed as follows: (The Introduction and the Related Works) written by Dr. Wisal Hashim Abdulsalam, (Materials and Methods) written by Dr. Rasha H. Ali, (The Proposed Model and The Conclusion) written by Dr. Rasha H. Ali and Dr. Wisal Hashim Abdulsalam. According to (The Data Collection), all authors contributed to collecting data. All authors reviewed the manuscript.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Ethical Approval

The study protocol was reviewed and approved by the Scientific Committee of the Computer Department, College of Education for Women, University of Baghdad, and the Scientific Committee of the Computer Department, College of Education for Pure Science - Ibn Al-Haitham, University of Baghdad. The study was conducted in accordance with the ethical principles of the World Medical Association Declaration of Helsinki (2013 revision) and the relevant institutional and national regulations governing research involving human participants.

Consent to Participate

Informed consent was obtained from all participants involved in the study. For participants under the age of 18 years, informed consent was obtained from their parents/guardians before participation. Participation was voluntary, and all respondents were informed of the purpose of the study and their right to withdraw at any time without penalty.

Consent to Publish: Not applicable

Data Availability:

The data supporting the findings of this study are available at GitHub: https://github.com/RashaAli11/ASD-dataset-by-survey-

Research involving human or animals: Not applicable.

Competing Interests: The authors declare no competing interests

Refernces

1 Toman, A.K., AbdulRasool, H.R., Lami, F., Jasim, S.M., Jaber, O.A., Nayeri, N.D., Sabet, M.S., and Al-Gburi, G.: 'Exploring co-occurring conditions in Iraqi children with autism spectrum disorder: prevalence, characteristics, and potential risk factors', Frontiers in Psychiatry, 2025, 16, pp. 1592374, https://doi.org/10.3389/fpsyt.2025.1592374

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

3 Sadhana, B., Dhananjaya, B., Sindhoora, O., and Maithili, R.S.: 'Detecting Autism at Early Stages: A Machine Learning Approaches', in 2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON), BENGALURU, India, 2025, pp. 1-6, https://doi.org/10.1109/NMITCON65824.2025.11188283

4 Gill, K.S., Agrawal, J., Chauhan, R., and Pokhariya, H.S.: 'Establishing a Machine Learning-Based Random Forest Classifier to Estimate Autism Risk', in 2024 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2024, pp. 1-6, https://doi.org/10.1109/INOCON60754.2024.10511625

5 Mittal, K., Gill, K.S., Upadhyay, D., Singh, V., and Aluvala, S.: 'Applying Machine Learning for Autism Risk Evaluation Using a Decision Tree Classification Technique', in 2024 2nd International Conference on Computer, Communication and Control (IC4), Indore, India, 2024, pp. 1-6, https://doi: 10.1109/IC457434.2024.10486622

6 ArunKumar, G., Sumalatha, K., Saravanan, D., and Bisht, A.K.: 'Optimized temporal inductive path neural network based early-stage detection of autism spectrum disorders', Pattern Recognition, 2026, 172, pp. 112689, https://doi.org/10.1016/j.patcog.2025.112689

7 Mumenin, N., Rahman, M.M., Yousuf, M.A., Noori, F.M., and Uddin, M.Z.: 'Early diagnosis of autism across developmental stages through scalable and interpretable ensemble model', Frontiers in Artificial Intelligence, 2025, 8, pp. 1507922, https://doi.org/10.3389/frai.2025.1507922

8 Al-Qazzaz, N.K., Aldoori, A.A., Buniya, A.K., Ali, S.H.B.M., and Ahmad, S.A.: 'Transfer learning and hybrid deep convolutional neural networks models for autism spectrum disorder classification from eeg signals', IEEe Access, 2024, 12, pp. 64510-64530, https://doi.org/10.1109/ACCESS.2024.3396869

9 Alam, M.S., Rashid, M.M., Faizabadi, A.R., Mohd Zaki, H.F., Alam, T.E., Ali, M.S., Gupta, K.D., and Ahsan, M.M.: 'Efficient deep learning-based data-centric approach for autism spectrum disorder diagnosis from facial images using explainable AI', Technologies, 2023, 11, (5), pp. 115, https://doi.org/10.3390/technologies11050115

10 Hasan, S.M., Uddin, M.P., Al Mamun, M., Sharif, M.I., Ulhaq, A., and Krishnamoorthy, G.: 'A machine learning framework for early-stage detection of autism spectrum disorders', IEEE Access, 2022, 11, pp. 15038-15057, https://doi.org/10.1109/ACCESS.2022.3232490
11 Sujatha, R., Aarthy, S., and NZ, J.: 'A machine learning way to classify autism spectrum disorder', International Journal of Emerging Technologies in Learning (iJET), 2021, 16, (6), pp. 182-200, https://doi.org/10.3991/ijet.v16i06.19559
12 Raj, S., and Masood, S.: 'Analysis and detection of autism spectrum disorder using machine learning techniques', Procedia Computer Science, 2020, 167, pp. 994-1004, https://doi.org/10.1016/j.procs.2020.03.399
13 Wedasingha N, Samarasinghe P, Senevirathna L, Papandrea M, Puiatti A.: 'Autoencoder based data clustering for identifying anomalous repetitive hand movements, and behavioral transition patterns in children', Physical and Engineering Sciences in Medicine. 2025, 48(1), pp. 221-38, https://doi.org/10.1007/s13246-024-01507-9
14 Wisal Hashim Abdulsalam, R.K.K.A., Mohammed Ayad Saad: 'Integrating Feature Selection and Machine Learning Boosting for Accurate Breast Cancer Prediction'. Proc. Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIT), Tikrit, Iraq2025 pp. 91-100, http://dx.doi.org/10.25673/122070
15 Jadoaa, S.H., Ali, R.H., Abdulsalam, W.H., and Alsaedi, E.M.: 'The Impact of Feature Importance on Spoofing Attack Detection in IoT Environment', Mesopotamian Journal of CyberSecurity, 2025, 5, (1), pp. 240-255, https://doi.org/10.58496/MJCS/2025/016
16 Santos, M.R., Guedes, A., and Sanchez-Gendriz, I.: 'Shapley additive explanations (shap) for efficient feature selection in rolling bearing fault diagnosis', Machine Learning and Knowledge Extraction, 2024, 6, (1), pp. 316-341, https://doi.org/10.3390/make6010016
17 Kuang, B., Yang, H., and Jung, T.: 'The Impact of Visual Elements in Street View on Street Quality: A Quantitative Study Based on Deep Learning, Elastic Net Regression, and SHapley Additive exPlanations (SHAP)', Sustainability, 2025, 17, (8), pp. 3454, https://doi.org/10.3390/su17083454
18 Khan, A., Ali, A., Khan, J., Ullah, F., and Faheem, M.: 'Using Permutation-Based Feature Importance for Improved Machine Learning Model Performance at Reduced Costs', IEEE Access, 2025, https://doi.org/10.1109/ACCESS.2025.3544625
19 Abdelaziz, M.T., Radwan, A., Mamdouh, H., Saad, A.S., Abuzaid, A.S., AbdElhakeem, A.A., Zakzouk, S., Moussa, K., and Darweesh, M.S.: 'Enhancing network threat detection with random forest-based NIDS and permutation feature importance', Journal of Network and Systems Management 33, 2 (2025), https://doi.org/10.1007/s10922-024-09874-0
20 Rasha H. Ali, W.H.A.: 'Attention-Deficit Hyperactivity Disorder Prediction by Artificial Intelligence Techniques', Iraqi Journal of Science, 2024, 65, (9), pp. 5281-5294, https://doi.org/10.24996/ijs.2024.65.9.39
21 Mashhadani, S., Abdulsalam, W.H., Shukur, W.A., and Al-Tamimi, M.S.: 'Digital Forensics Method for Fake Images Detection Using GAN Algorithm Based on Watermarking Technique and Image Content', Iraqi Journal of Science, 2026, 67, (1), pp. 509-523, https://doi.org/10.24996/ijs.2026.67.1.40

22 Abdulsalam, W.H., Ali, R.H., Jadooa, S.H., and Hussein, S.S.: 'Automated Glaucoma Detection Techniques: A Literature Review', Engineering, Technology & Applied Science Research, 2025, 15, (1), pp. 19891-19897, https://doi.org/10.48084/etasr.9316
23 Ban Hassan, M., Hashim Abdulsalam, W., Hazim Ibrahim, Z., H Ali, R., and Mashhadani, S.: 'Digital Intelligence for University Students Using Artificial Intelligence Techniques', International Journal of Computing and Digital Systems, 2025, 17, (1), pp. 1-10, https://doi.org/10.12785/ijcds/1571029446
24 Samadi, S.A.: 'The Challenges of Establishing Healthcare Services in low-and Middle-Income countries: The case of Autism Spectrum disorders (ASD) in the Kurdistan Region of Iraq-Report from the field', Brain sciences, 2022, 12, (11), pp. 1433, https://doi.org/10.3390/brainsci12111433