Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 20, 2025

FROM DATA TO DIAGNOSIS: KNOWLEDGE-DRIVEN, EXPLAINABLE AI FOR RELIABLE EARLY AUTISM DETECTION

Qusai Yousef Shambour*	Department of Software Engineering, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan	q.shambour@ammanu.edu.jo
Mahran Al-Zyoud	Department of Networks and Cybersecurity, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan	m.zyoud@ammanu.edu.jo
Abdelrahman H. Hussein	Department of Networks and Cybersecurity, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan	a.husein@ammanu.edu.jo

* Corresponding author

ABSTRACT

Aim/Purpose	The primary aim of this study is to address the persistent challenge of delayed autism spectrum disorder (ASD) diagnosis in toddlers. Early detection enables timely interventions that can improve developmental outcomes; however, conventional approaches rely on lengthy and resource-intensive behavioral assessments. We therefore introduce an interpretable AI screening framework designed to accelerate ASD triage while providing clinically understandable rationales to support decision-making.
Background	Traditional ASD diagnosis depends on expert behavioral evaluation and parent reports which, despite their value, are time-consuming and capacity-limited, delaying access to early intervention. With ASD prevalence rising, scalable and effective approaches are urgently needed. This study proposes a robust AI framework for early ASD detection that integrates targeted preprocessing, feature selection, principled model optimization, and post-hoc explanations, aiming to

Accepting Editor Geeta Sandeep Nadella | Received: August 4, 2025 | Revised: September 28, November 7, 2025 | Accepted: November 8, 2025.
Cite as: Shambour, Q. Y., Al-Zyoud, M., Hussein, A. H. (2025). From data to diagnosis: Knowledge-driven, explainable AI for reliable early autism detection. *Interdisciplinary Journal of Information, Knowledge, and Management, 20,* Article 32. https://doi.org/10.28945/5652

	improve diagnostic utility and clarity for end users in clinical and community settings.
Methodology	We develop a unified, reproducible pipeline that combines data preprocessing, class balancing, feature selection, and Bayesian hyperparameter tuning. The pipeline also incorporates SHapley Additive exPlanations (SHAP) to provide model explanations. Six diverse machine learning models – Extreme Gradient Boosting (XGB), Histogram-based Gradient Boosting (HGB), Random Forest (RF), Naïve Bayes (NB), Mixture Discriminant Analysis (MDA), and Multi-layer Perceptron (MLP) – are evaluated to assess framework robustness rather than to crown a single best classifier. A cross-cultural dataset of toddlers aged 12–36 months (n=1,560) is constructed by merging two public sources containing Q-CHAT-10 items with demographic attributes. Preprocessing removes non-informative variables and encodes categorical features; Gaussian noise-based upsampling (GNUS) mitigates post-merge imbalance; RobustScaler stabilizes training. Gradient Boosting Feature Selection (GBFS) ranks and reduces features to enhance parsimony and interpretability. Performance is reported via accuracy, precision, recall, F1, and Matthews Correlation Coefficient (MCC). Model behavior is elucidated with SHAP to reveal feature contributions and decision pathways.
Contribution	This work presents an interpretable AI framework for early ASD detection that couples performance with clinician-oriented explanation in a single pipeline. Rather than optimizing for accuracy alone, we emphasize synergy among preprocessing, balancing, feature selection, and explanation – the multimodel evaluation evidence adaptability across algorithmic families. GBFS and SHAP are used to ensure concise, explainable predictions. Notably, the framework achieved very strong internal validation results (high F1 and MCC across folds) with XGB, while SHAP-derived patterns aligned with clinical heuristics. Results are promising but preliminary, pending external, multi-site validation.
Findings	GNUS and robust normalization improved generalization on the cross-cultural dataset. With GBFS-selected features, XGB achieved near-ceiling internal scores across key metrics, a trend observed – though to a lesser extent – in other models after comparable optimization. SHAP consistently highlighted behaviors such as gaze-following and social/emotional responsiveness among the most influential predictors, in line with clinical practice. Collectively, the findings indicate that interpretable ML can complement conventional screening, while warranting prospective and external validation to assess generalizability and potential dataset shift.
Recommendations for Practitioners	Clinicians and community programs may consider adopting interpretable ML as a screening aid to prioritize referrals and shorten time-to-assessment. Attention to features repeatedly identified as influential can guide focused early interventions and resource allocation.
Recommendations for Researchers	Future studies should test the framework on larger and more diverse cohorts to evaluate generalizability. Exploring ensembles and deeper architectures, as well as alternative preprocessing, resampling, and feature selection strategies, may further enhance performance, particularly for cases that are borderline.
Impact on Society	Earlier, more reliable screening can improve access to services during critical neurodevelopmental windows. Integrating interpretable AI into practice may

also strengthen clinician confidence in ML-assisted tools, supporting responsible, human-centered deployment and broader public health benefits.

Future Research — Next steps include conducting real-world pilots across various clinical/community settings, integrating with complementary diagnostic tools to build multimodal platforms, and systematically evaluating balancing/optimization choices. These directions will help translate the framework into practical impact and inspire analogous applications in pediatric neurodevelopmental assessment.

Keywords — data-to-diagnosis, early diagnosis, autism spectrum disorder, machine learning, explainable artificial intelligence, interpretable AI, knowledge-driven screening

INTRODUCTION

Autism spectrum disorder (ASD) represents a heterogeneous neurodevelopmental condition characterized by enduring challenges in social communication and interaction, alongside restricted, repetitive behavioral patterns. With a global prevalence of approximately 1 in 54 children, ASD imposes significant lifelong consequences not only on diagnosed individuals but also on families, healthcare infrastructures, and educational systems (Lauritsen, 2013). The heterogeneity of symptom presentation, ranging from subtle social reciprocity deficits to pronounced sensory sensitivities, complicates diagnostic processes and necessitates personalized intervention strategies (Fernell et al., 2013). Timely and accurate identification is critical, as the "data-to-diagnosis" journey presents a window of opportunity during which targeted interventions can significantly alter developmental trajectories and enhance adaptive functioning (Lord & Luyster, 2006). Despite this urgency, diagnostic delays persist globally due to the reliance on subjective clinical assessments, resource disparities, and the nuanced manifestation of early behavioral markers that often overlap with those of other developmental conditions (McCarty & Frye, 2020).

Conventional diagnostic frameworks utilize standardized instruments, such as the Autism Diagnostic Observation Schedule (ADOS) and the Childhood Autism Rating Scale (CARS), which monitor social communication impairments and repetitive behaviours using clinician observation and input from caregivers (Thabtah & Peebles, 2019). These tools are regarded as the benchmark for diagnosing ASD; however, their administration is resource-intensive, requiring highly trained specialists and often taking months or even years to complete. As a result, this contributes to long waiting lists and many children being left undiagnosed until they reach school age, as opportunities for early intervention are severely limited (Fekar Gharamaleki et al., 2022). Furthermore, the subjective nature of these assessments can introduce limitations, such as inter-clinician variability and potential sampling biases, highlighting the necessity of better, objective, and scalable measures that augment, rather than replace, clinical expertise (Omar et al., 2019).

In light of these challenges, artificial intelligence (AI) and machine learning (ML) have emerged as promising, transformative solutions in healthcare research and ASD diagnostics, offering the ability to translate complex data into actionable clinical knowledge (Abu Owida et al., 2025; Abu-Shareha et al., 2025; Prasad & Bhatia, 2025). By analyzing diverse data sources (e.g., behavioral questionnaires and electronic health records, neuroimaging), ML models can identify complex predictive patterns, potentially shortening the path from initial screening to diagnosis (Farooq et al., 2023; Lu et al., 2009; Luong et al., 2025; Shambour et al., 2021, 2022; Tran et al., 2023). However, a critical barrier to the clinical adoption of these models has been their typical "black-box" nature. High-performing models often provide accurate predictions but fail to offer clinicians transparent, understandable rationales for their decisions. This lack of transparency undermines trust in their utility in real-world clinical workflows (Abdullah et al., 2021).

This gap has, to some extent, led to the emergence of Explainable AI (XAI), a field devoted to making AI decisions interpretable for human users. Techniques such as SHAP aim to bridge the interpretability gap by quantifying the contribution of each input feature to a model's prediction, thereby aligning computational evidence with clinical reasoning (Lundberg & Lee, 2017). Nevertheless, achieving robust and trustworthy AI requires more than just applying XAI post-hoc; it demands a holistic, integrated pipeline that addresses fundamental data challenges, including class imbalance, feature redundancy, and cultural variability, that can undermine model performance and generalizability (Ferrari et al., 2020). Unfortunately, without such a pipeline, even the most accurate algorithms risk continuing diagnostic biases and remaining unfit for clinical deployment.

This study directly addresses these challenges by introducing a comprehensive AI framework designed for the entire journey from data to diagnosis in early ASD detection. Our work advances beyond existing research in three key ways:

1) *Integrated pipeline:* We unify advanced data preprocessing (Gaussian Noise UpSampling for class balancing, RobustScaler normalization), embedded feature selection (Gradient Boosting Feature Selection), Bayesian hyperparameter optimization, and SHAP-based explainability into a single, reproducible pipeline.
2) *Clinically actionable outputs:* Rather than solely pursuing accuracy, we prioritize the creation of clinically actionable insights. Our framework is designed to produce not just a prediction, but an auditable, clinician-ready rationale that aligns with established clinical understanding, thereby facilitating adoption.
3) *Robust validation:* We demonstrate the framework's efficacy and consistency using a cross-cultural dataset and a diverse suite of six ML models – XGB, HGB, RF, NB, MDA, and MLP – proving its robustness across different algorithmic paradigms.

Preliminary results are compelling, with our optimized pipeline achieving near-perfect, cross-validated performance with the XGB classifier. More importantly, SHAP analysis confirms that the model's decision-making is driven by well-established behavioral markers such as gaze-following and social responsiveness, validating its clinical relevance. This synergy of high accuracy and clear explanations represents a significant step toward trustworthy, knowledge-driven AI tools that can be integrated into clinical workflows to support earlier and more equitable ASD diagnosis.

The remainder of this paper is organized as follows. The next section reviews related work on ML and XAI for ASD detection. Then, the design and implementation of our proposed framework are detailed. The experimental setup and results are then discussed, followed by an examination of the findings, their implications, and the limitations. Finally, the paper concludes and suggests directions for future research.

RELATED WORKS

The application of ML in the context of ASD has undergone a profound and rapid evolution, transitioning from a nascent field focused predominantly on diagnostic efficiency to a sophisticated, multidisciplinary endeavor. This evolution is characterized by a progressive shift in priorities, moving from the foundational goal of achieving high predictive accuracy to a more holistic and clinically-oriented pursuit of transparent, multimodal, and ethically robust decision-support systems. The trajectory of this research can be understood by tracing its development across three distinct yet overlapping phases. The first phase was marked by an intensive drive to automate screening and enhance classification accuracy, establishing the viability of ML for this clinical application. The second phase witnessed a significant expansion into more complex methodologies, including advanced feature engineering and the integration of multimodal data streams, which enriched the diagnostic potential of these systems. The current and most recent phase represents a critical turn towards operationalizing

these technologies for real-world clinical use, defined by a strong emphasis on XAI and the establishment of frameworks for responsible and ethical implementation. This comprehensive journey reflects the growing maturity of the field, as it strives to create tools that are not only technically powerful but also clinically meaningful, trustworthy, and aligned with the nuanced demands of ASD assessment and care.

Early research in this domain was fundamentally driven by the goal of automating the screening process and improving the accuracy of diagnostic classifications. During this foundational period, researchers primarily explored the capabilities of traditional ML models to distinguish between individuals with and without ASD based on structured clinical and behavioral data. For example, some studies developed mobile applications using decision tree-based models with promising accuracy rates, but limited interpretability for model results (Omar et al., 2019). This work highlighted the practical utility of ML but lacked a complete, transparent data-to-diagnosis pipeline. Other foundational studies explored traditional classifiers like Naïve Bayes and logistic regression, demonstrating ML's potential to identify relationships between clinical and demographic factors (Usta et al., 2019). However, these studies also continued to highlight overarching issues of variability of datasets and limited external validation of their approaches, which indicated that improved accuracy on its own was not enough for real-world clinical uptake (Usta et al., 2019).

The subsequent phase of research introduced and implemented more sophisticated and computationally intensive techniques to further boost predictive power and model robustness. Methods such as AdaBoost and, most notably, RF, became the new standard, with numerous studies reporting their superior ability to handle the complex and often noisy data characteristic of ASD research. These ensemble models proved particularly effective when applied to age-stratified datasets, where they could learn the subtle, age-dependent manifestations of ASD traits, with some models achieving remarkable accuracy rates exceeding 99% (Sujatha et al., 2021; Uddin et al., 2023). Data-centric strategies, such as systematic preprocessing with the synthetic minority over-sampling technique (SMOTE) for class balancing and advanced feature selection using principal component analysis (PCA), also yielded near-perfect classification results (Aldrees et al., 2024; Uddin et al., 2023). Despite these methodological advances, the primary focus remained on quantitative metrics. Some frameworks began to bridge this gap by linking diagnosis with post-diagnosis educational planning, yet they still relied on aggregate feature rankings instead of providing case-specific, interpretable rationales (Hajjej et al., 2024).

Recognizing the key limitations of purely performance-driven models, often referred to as "black-box" systems, the field has recently undertaken a critical step towards integrating XAI methodologies. This move is motivated by the need to foster clinical trust, enhance the utility of the model, and provide actionable insights that resonate with clinical practice. The introduction of SHAP represented a significant milestone in the research process, enabling the measurement of a feature's individual contribution to a specific prediction (Mumenin et al., 2025). This has now escalated into a number of XAI tools designed during this era for specific data types. For instance, Local Interpretable Model-Agnostic Explanations (LIME) has also been developed to explain what input features have affected predictions, e.g., what regions of a facial image influenced a prediction, for deep learning models (Atlam et al., 2025). In neuroimaging, Gradient-weighted Class Activation Mapping (Grad-CAM) developed salience maps that indicate the regions that most influenced a CNN's decision and can be directly compared to neurologically relevant findings (Varghese et al., 2024). Recent reviews have recommended triangulating these methods as a way to improve transparency and clinical auditability in ASD decision support (Agrawal & Agrawal, 2025).

Concurrently with the growing interest in XAI, there has been a significant and ongoing emphasis on transitioning to a multimodal data stream approach to create a richer, more nuanced, and ultimately more accurate understanding of ASD. The underlying rationale behind this is that ASD is a complicated neurodevelopmental disorder that can be expressed in equally complex ways, and therefore, relying on data collected from a single modality could provide a reductive view of the disorder. By integrating information from various sources, researchers believe they will be better able to encapsulate

the multifaceted nature of the disorder. Among the most prominent and powerful complementary data streams developed in recent years are eye-tracking and electroencephalography (EEG) (Fonseca et al., 2025; Jaradat et al., 2025; Sun et al., 2024; Wei et al., 2023). Models based on eye-tracking data have achieved high accuracy in toddlers (Jaradat et al., 2025; Wei et al., 2023), and EEG-based pipelines have successfully identified neurophysiological markers (Fonseca et al., 2025). Data fusion studies that combine EEG and eye-tracking have shown that integrating these modalities not only improves classification accuracy but also provides a deeper mechanistic understanding of the interplay between neural dynamics and attentional processes in ASD (Sun et al., 2024). Given the relative success of these multimodal approaches, the inclusion of multiple approaches will become increasingly essential for developing diagnostic tools that are both clinically informative and accurate, pending validation in larger multicenter cohorts (Sun et al., 2024).

Alongside the technical challenges of multimodal integration, the growing sophistication of ML in ASD research has brought ethical considerations to the forefront. It is now widely recognized that performance metrics alone are insufficient for responsible deployment. Developers must proactively assess and mitigate algorithmic bias to ensure models are fair across different demographic subgroups, including sex, age, and socioeconomic strata. Furthermore, data privacy and governance have become critical, necessitating strict consent protocols, data minimization practices, and secure data-sharing frameworks suited to pediatric and multi-site data. These ethical imperatives – highlighted in recent reviews discussing bias risks and the need for robust regulatory safeguards – are central to ensuring that AI tools are deployed in a manner that is safe, equitable, and aligned with the core principles of clinical practice, thereby fostering trust and facilitating the successful translation of these technologies from research to real-world use (Rêgo & Araújo-Filho, 2024).

DESIGN AND IMPLEMENTATION

This section outlines a knowledge-driven, transparent, and XAI framework for early ASD detection in toddlers, emphasizing data robustness, interpretability, and actionable clinical insight. The proposed pipeline was designed to guide the full journey "from data to diagnosis," ensuring that each stage – from preprocessing and feature selection to model training and interpretability – contributes to meaningful, trustworthy diagnostic support. With a culturally diverse dataset, the framework tackles class imbalance with GNUS and employs RobustScaler for feature normalization. GBFS identifies the most important diagnostic markers. Subsequently, a set of six ML models was subjected to Bayesian hyperparameter optimization to enhance performance and mitigate overfitting.

A central innovation of this framework is the integration of XAI, specifically SHAP, which provides clear, transparent insights into model decision-making, thereby transforming raw data into knowledge-ready, clinically actionable explanations. Figure 1 illustrates the architecture of the proposed framework. The following section presents the experimental setup and evaluation of this framework, including dataset description, preprocessing, feature selection, model training, hyperparameter tuning, and performance metrics.

EXPERIMENTAL DESIGN AND EVALUATION

This section presents a comprehensive overview of the experimental procedures employed to assess the performance, robustness, and interpretability of the proposed transparent and XAI framework for ASD detection in toddlers. The evaluation includes a detailed description of the merged cross-cultural datasets, the complete preprocessing and feature selection pipeline, the configuration and optimization of ML models, and the integration of XAI techniques. Additionally, we outline the experimental setup, including class balancing and validation strategies, and define the performance metrics and statistical analyses used to ensure methodological rigor, reproducibility, and interpretability throughout the data-to-diagnosis process. All experiments were implemented in Python 3.10 (64-bit) on a Windows 11 Pro environment equipped with an Intel(R) Core(TM) i9-14900HX CPU @ 2.20

GHz and 32 GB RAM, using Scikit-learn 1.3.1 for model implementation, scikit-optimize (skopt) 0.9.0 for hyperparameter tuning, and SHAP 0.42 for explainability, ensuring full reproducibility.

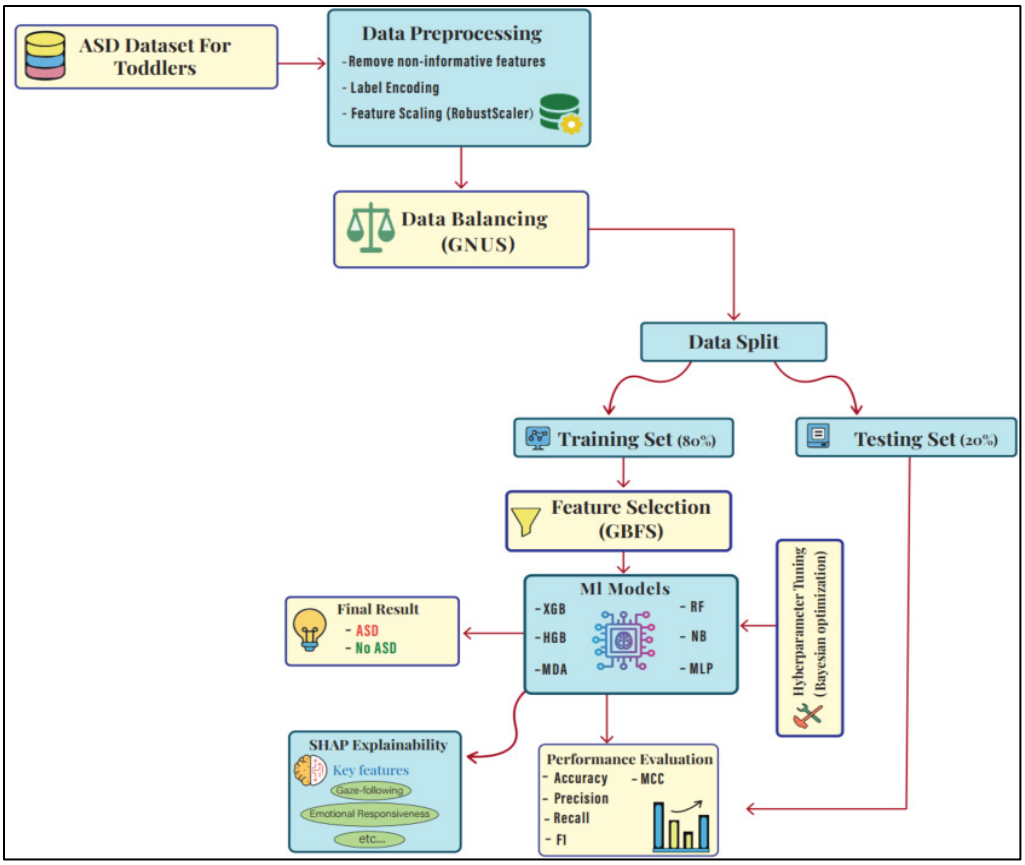

Figure 1. Design of the proposed framework

Datasets Description

Two publicly available datasets were utilized in the study to create an ASD toddler dataset for cross-cultural purposes, ensuring demographic and cultural diversity. The first dataset is from Thabtah et al. (2018), which comprises 1,054 instances with 17 categorical features derived from the Q-CHAT-10 screening tool, including ten behavioral variables and several demographic variables. This dataset contains 735 ASD-positive instances and 319 ASD-negative instances. The second dataset, sourced by Alkahtani et al. (2023), was collected using an Arabic adaptation of the Q-CHAT-10 instrument, comprising 341 cases of ASD-positive individuals and 165 cases of ASD-negative individuals.

After merging, the combined dataset included a total of 1,560 instances, with 1,076 ASD-positive and 484 ASD-negative cases. This strategy was motivated by the need to enhance the knowledge-value, scientific validity, and generalizability of the framework, ensuring that the resulting pipeline is relevant for diverse real-world clinical scenarios.

Dataset Preprocessing

Preprocessing began with feature selection to eliminate attributes that were not predictive. We remove "Case no" and "Who is completing the test" because there are no meaningful comparisons to be made through them as predictors. Exploratory analyses confirmed the absence of missing values or duplicates, obviating the need for imputation or deduplication.

To address class imbalance, we adopted Gaussian Noise Upsampling (GNUS) rather than SMOTE or ADASYN. SMOTE's linear interpolation can introduce synthetic samples that unrealistically smooth clinical distributions, while ADASYN may oversample boundary/noisy regions. GNUS first duplicates minority-class instances and then injects zero-mean Gaussian perturbations into selected numeric/ordinal encodings of Q-CHAT and demographic features, preserving clinically plausible patterns while enhancing diversity and reducing the risk of overfitting (Beinecke & Heider, 2021). Following GNUS upsampling, the final working dataset comprised 2,138 instances with perfectly balanced classes: 1,069 ASD-positive and 1,069 ASD-negative cases. This balanced dataset provided a robust and knowledge-rich foundation for training explainable models.

Feature transformation involved label encoding of categorical variables to retain ordinal relationships and RobustScaler normalization. The latter standardizes the data, centering on the median and scaling by interquartile ranges, thereby minimizing the influence of outliers. This reproducible data-to-diagnosis preprocessing protocol establishes methodological rigor, maximizes the knowledge content of the input, and enables transparent downstream interpretation.

Feature Selection Methodology

To mitigate dimensionality challenges and enhance model efficacy, feature selection techniques were implemented to recognize predictors with significant diagnostic utility while neglecting non-contributory attributes. This process reduces computational complexity, mitigates the risk of overfitting, and directly enhances the interpretability and knowledge value of subsequent predictions. The study employed GBFS, a method exploiting the in-built analytical capacities of Gradient Boosting Machines (GBM) (Chen & Guestrin, 2016) for prioritizing variables with the greatest predictive influence.

GBFS operates by extracting feature importance weights from a GBM, which iteratively combines weak learners (typically decision trees) to improve prediction accuracy. Each subsequent learner adjusts previous iterations for residuals by using relative contributions to minimize the loss function (e.g., log loss for classification). The importance attributed to the subsequent learner follows three interpretations: (1) gain score, which shows the improvement of the split accuracy of the feature; (2) coverage score, which is the percentage of observations that the feature acts on; and (3) frequency score, which is the number of times the feature gets chosen over decision trees.

On the basis of aggregate importance ranking using GBFS, features are kept accordingly, while others perceived to have a lesser effect or add noise are dropped. This systematic dimensionality reduction optimizes computational efficiency, accelerates training phases, and enhances model transparency, as clinically meaningful predictors alone are retained. Enhanced by applying the GBFS, the methodology guarantees sound feature engineering designed to balance algorithm performance with interpretability – a critical consideration for deploying ML solutions in clinical contexts.

Machine Learning Models

This study evaluates six ML models for early ASD detection in toddlers, chosen for their complementary strengths in managing moderate-sized, heterogeneous clinical data and supporting interpretability. The suite comprises Extreme Gradient Boosting (XGB), Histogram-based Gradient Boosting (HGB), Random Forest (RF), Naïve Bayes (NB), Mixture Discriminant Analysis (MDA), and Multilayer Perceptron (MLP).

> **XGB** (Chen & Guestrin, 2016) and **HGB** (Ke et al., 2017) are advanced ensemble methods that iteratively improve classification by focusing on errors made by previous learners. XGB incorporates regularization to minimize overfitting, making it well-suited to imbalanced clinical data, while HGB utilizes histogram-based binning for faster training and improved efficiency.
>
> **RF** aggregates predictions from an ensemble of randomized decision trees, reducing overfitting and enhancing model robustness through majority voting and feature randomness (Biau &

Scornet, 2016). RF's feature importance scores assist in highlighting the most predictive attributes for ASD screening.

NB provides a probabilistic baseline, leveraging the assumption of feature independence for rapid, interpretable predictions (Peretz et al., 2024). Despite its simplicity, NB can provide probabilistically interpretable predictions and serves as a competent baseline for exploratory analyses and rapid screening.

MDA models each class as a mixture of Gaussian distributions (Hastie & Tibshirani, 2018), allowing the classifier to capture potential non-linear and multimodal symptom patterns, which is pertinent for the spectrum nature of ASD.

MLP is a feedforward neural network architecture adept at capturing subtle, non-linear interactions among features within screening questionnaires. Its flexibility enables the detection of complex patterns that may elude both rule-based and linear models, making it particularly valuable for nuanced ASD classification tasks. When appropriately regularized, MLP offers a powerful complement to ensemble methods by identifying intricate relationships in clinical screening data (Goodfellow et al., 2016).

To provide a comprehensive and transparent assessment of predictive performance, with emphasis on clinical applicability and interpretability in ASD screening, we evaluated models spanning ensemble (XGB, HGB, and RF), probabilistic (NB), discriminant (MDA), and neural paradigms (MLP). Given the modest sample size after balancing (n = 2,138) and the questionnaire-based modality, we deliberately excluded deeper neural networks (e.g., CNNs, LSTMs) due to their higher capacity and overfitting risk in this setting. Accordingly, we prioritized models that balance strong predictive performance with transparency and reproducibility – two elements essential for clinical trust and eventual deployment.

Hyperparameter Tuning

In this study, hyperparameter optimization was exclusively conducted using Bayesian optimization, a state-of-the-art, probabilistic technique for efficient exploration of high-dimensional search spaces (Wu et al., 2019). Unlike traditional grid or random search methods, which may waste computational resources by evaluating suboptimal or redundant parameter combinations, Bayesian optimization constructs a surrogate probabilistic model – commonly a Gaussian Process (GP) – to approximate the objective function representing model performance (e.g., validation accuracy or F1-score) as a function of the hyperparameters.

Formally, let $f(\mathbf{x})$ denote the black-box objective function mapping a hyperparameter vector $\mathbf{x}$ to a cross-validated performance metric. Bayesian optimization maintains a posterior distribution by $f(\mathbf{x})$ using observations from previous evaluations. At each iteration, an acquisition function, such as Expected Improvement (EI) or Upper Confidence Bound (UCB), is maximized to select the next candidate $\mathbf{x}_{\text{next}}$ to evaluate:

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}; \mathsf{D}) \tag{1}$$

where $\boldsymbol{\alpha(\mathbf{x}; \mathsf{D})}$ is the acquisition function and D is the set of observed pairs $(\mathbf{x}_i, f(\mathbf{x}_i))$. The selected $\mathbf{x}_{\text{next}}$ is then evaluated, and the surrogate model is updated accordingly.

BayesSearchCV was employed to apply this process across all ML classifiers, using stratified cross-validation on the balanced dataset to ensure leakage-free and reliable evaluation. This strategy systematically explored model depth, capacity, and regularization while maintaining computational feasibility. The complete search spaces for each ML model are summarized in Table 1.

Table 1. Search space for hyperparameter tuning

Model	Hyperparameter	Search space
XGBoost	n_estimators	[50, 300]
	max_depth	[3, 10]
	learning_rate	[0.001, 0.3]
HGB	learning_rate	[0.001, 0.3]
	max_depth	[3, 16]
	min_samples_leaf	[1, 50]
RF	n_estimators	[50, 300]
	max_depth	[3, 15]
	min_samples_split	[2, 10]
	min_samples_leaf	[1, 4]
MDA	n_components	[2, 4]
	max_iter	[100, 300]
MLP	n_layers	[1, 4]
	n_neurons (units per layer)	[16, 256]
	activation	[identity, logistic, tanh, relu]
	α (L2 regularization)	[1e-5, 1e-2]
	learning_rate	[constant, invscaling, adaptive]
	solver	[lbfgs, sgd, adam]

Models Explainability using SHAP

Central to the "from data to diagnosis" framework is the explicit use of SHAP to ensure all predictions are not only accurate but also transparent, explainable, and knowledge-ready. SHAP, rooted in cooperative game theory, assigns a Shapley value to each feature, representing its marginal contribution to individual predictions (Lundberg & Lee, 2017). SHAP summary plots and dependence plots are generated to rank features by impact and illustrate their influence, while force plots provide individualized, knowledge-rich explanations supporting clinical review. By embedding SHAP into the evaluation pipeline, every prediction becomes a transparent, actionable unit of knowledge – ready for use in real-world clinical diagnosis and shared decision-making.

Mathematically, for a model f and input vector x, the SHAP value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - f_S(\mathbf{x}_S) \right] \quad (2)$$

where F is the set of all features, S is a subset of features not containing i, and f_S denotes the model trained with only the features in S.

In practice, SHAP summary plots were generated to rank features by their mean absolute Shapley values, while dependence plots illustrated the relationship between feature values and their impact on model output. For example, features such as gaze-following and social responsiveness exhibited consistently high SHAP values, highlighting their importance in ASD risk assessment. Additionally, for individual patients, force plots provided a visual breakdown of how each feature value contributed to the final model prediction, enhancing transparency and supporting case-by-case clinical review. By integrating SHAP into the evaluation pipeline, the study ensured not only high predictive performance but also actionable, interpretable results suitable for deployment in real-world healthcare settings.

PERFORMANCE ASSESSMENT METRICS

A comprehensive suite of statistical metrics was employed to assess classifier effectiveness and generalizability. These metrics included:

- **Accuracy (ACC)**: The proportion of correct predictions, calculated as

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

- **Precision (Prec)**: The ratio of correctly predicted positive cases to all predicted positives,

$$\mathrm{Prec} = \frac{TP}{TP + FP} \tag{4}$$

reflecting the model's reliability in identifying ASD cases.

- **Recall (Sensitivity, Rec)**: The proportion of actual ASD cases correctly identified,

$$\mathrm{Rec} = \frac{TP}{TP + FN} \tag{5}$$

emphasizing early detection capacity.

- **F1-Score (F1)**: The harmonic mean of precision and recall,

$$\mathrm{F1} = 2 \times \frac{\mathrm{Prec} \times \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}} \tag{6}$$

which balances false positives and false negatives, especially valuable in imbalanced datasets.

- **Matthews Correlation Coefficient (MCC)**: A balanced measure robust to class imbalance,

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

which provides a single-value assessment of overall binary classification quality, ranging from -1 (total disagreement) to 1 (perfect prediction).

All metrics were computed on cross-validated test folds, ensuring comparability and reproducibility. The MCC, in particular, was prioritized for its reliability in clinical class imbalance, supporting the framework's goal of knowledge-driven, equitable, and trustworthy early autism diagnosis. This multi-dimensional evaluation framework provided nuanced insights into both strengths and limitations of each model, supporting robust, evidence-based conclusions regarding their clinical applicability.

EXPERIMENTAL RESULTS AND FINDINGS

The evaluation of the proposed knowledge-driven, transparent, and XAI framework for early ASD detection involved a thorough assessment of six classifiers – XGB, HGB, RF, NB, MDA, and MLP – across two feature selection strategies: the full feature set (ALL) and GBFS. Performance was evaluated using accuracy, precision, recall, F1-score, and MCC. A dual-validation approach was employed to ensure robustness: results were derived from a 10-fold cross-validation process, and model generalizability was further assessed using an 80/20 train-test split, with 80% of the data used for training and 20% held out for testing. This process was applied to a meticulously prepared dataset. Following GNUS upsampling, the final working dataset comprised 2,138 instances with perfectly balanced classes: 1,069 ASD-positive and 1,069 ASD-negative cases. This balanced and diverse dataset provided a

robust, real-world foundation for building transparent models that support reliable knowledge transfer from raw data to clinical diagnosis. While near-perfect results were achieved, it is crucial to contextualize these findings. The performance, particularly the perfect scores observed with XGB, was attained on a carefully curated and balanced dataset using robust validation strategies. The consistent performance across both k-fold cross-validation and a held-out test set (80/20 split) strongly mitigates the risk of overfitting and provides confidence in the models' generalizability within similar data distributions. However, performance in truly independent clinical settings may vary, and future work with external validation cohorts is planned to further confirm these results.

Table 2 outlines the comparative performance metrics, revealing notable variations across classifiers and feature selection methods. XGB stood out, achieving perfect scores of 1.000 across all metrics when paired with GBFS-selected features, in both validation schemes, a finding reinforced by its confusion matrix (Figure 2, left), generated from the 20% hold-out test set, which showed no misclassifications among the 428 test instances (215 true positives, 213 true negatives). With the full feature set, XGB maintained exceptional performance, recording 0.998 for all primary metrics and an MCC of 0.995. MLP followed closely, reaching 0.998 across accuracy, precision, recall, and F1-score with GBFS, with a single misclassification noted in its confusion matrix (Figure 4, right), reflecting 214 true positives and 213 true negatives out of 428 instances. HGB also performed robustly, improving from 0.988 with ALL features to 0.991 with GBFS, though its confusion matrix (Figure 2, right) indicated one false positive and three false negatives.

RF and MDA exhibited moderate enhancements with GBFS, with accuracy increasing from 0.977 to 0.984 for RF and from 0.963 to 0.972 for MDA. Their confusion matrices (Figures 3, left, and Figure 4, left) highlighted minor misclassifications, such as six false negatives for RF and 12 for MDA, suggesting a degree of sensitivity to the reduced feature set. NB, however, trailed behind, achieving only 0.944 with GBFS and 0.939 with ALL features, with its confusion matrix (Figure 3, right) revealing 12 false positives and 12 false negatives, indicating challenges in capturing the intricate behavioral patterns of ASD. The superior performance of models like XGB and MLP with GBFS suggests that the feature selection method effectively identified the most discriminative features. This is corroborated by the SHAP analysis (Figure 5), which revealed that the top features selected by GBFS – such as gaze-following (A6) and social responsiveness (A7) – were also the most impactful drivers of the model's predictions, providing a coherent link between feature selection efficacy and model performance. A Wilcoxon signed-rank test confirmed the statistical significance of GBFS's advantage over ALL features ($W = 21.0$, $p = 0.016$, $p < 0.05$), highlighting its consistent impact across the balanced cohort.

Table 2. Performance metrics of ML classifiers under full features (ALL) and GBFS-selected feature paradigms

Classifier	Feature selection	Accuracy	Precision	Recall	F1-Score	MCC
XGB	ALL	0.998	0.998	0.998	0.998	0.995
	GBFS	1.000	1.000	1.000	1.000	1.000
HGB	ALL	0.988	0.988	0.988	0.988	0.977
	GBFS	0.991	0.991	0.991	0.991	0.981
RF	ALL	0.977	0.977	0.977	0.977	0.953
	GBFS	0.984	0.984	0.984	0.984	0.968
NB	ALL	0.939	0.939	0.939	0.939	0.879
	GBFS	0.944	0.944	0.944	0.944	0.888
MDA	ALL	0.963	0.965	0.963	0.963	0.927
	GBFS	0.972	0.974	0.972	0.972	0.945
MLP	ALL	0.993	0.993	0.993	0.993	0.986
	GBFS	0.998	0.998	0.998	0.998	0.995

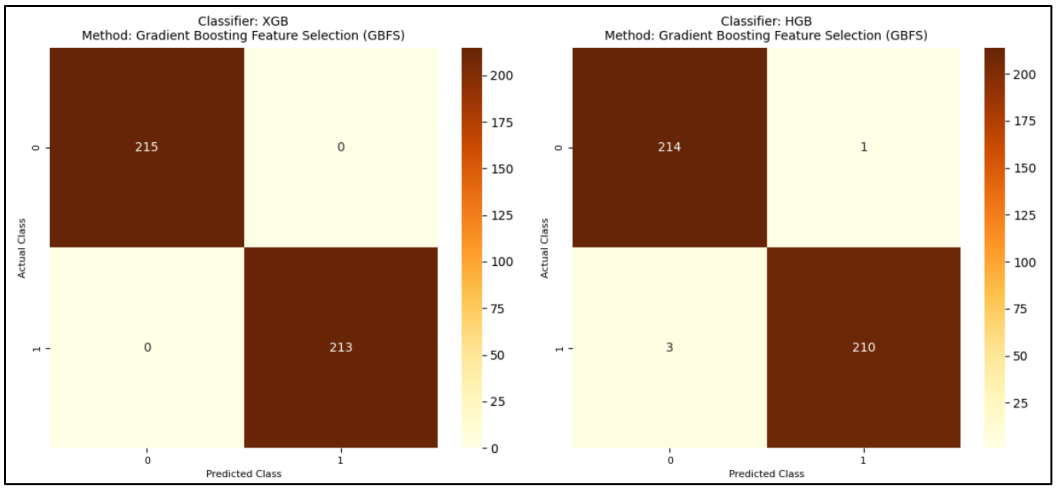

Figure 2. Confusion matrices for XGB and HGB models

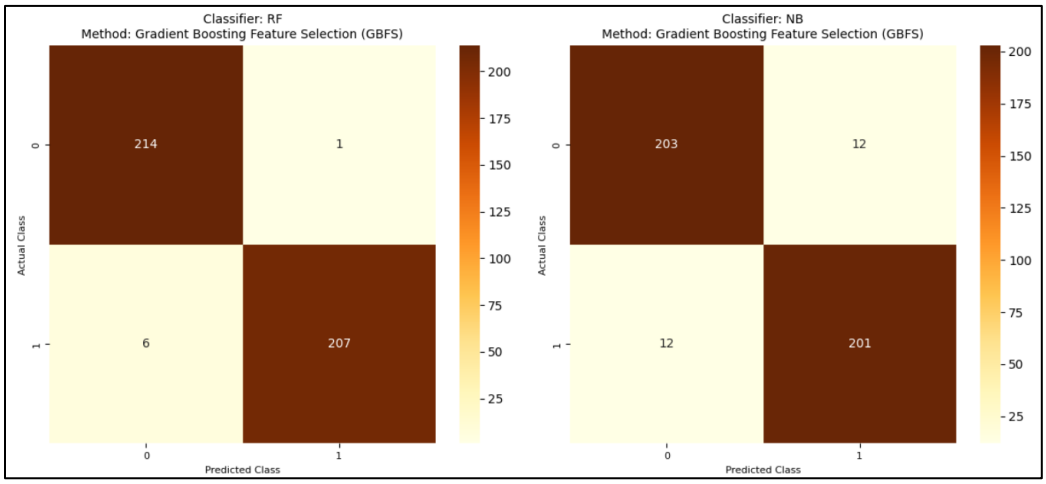

Figure 3. Confusion matrices for RF and NB models

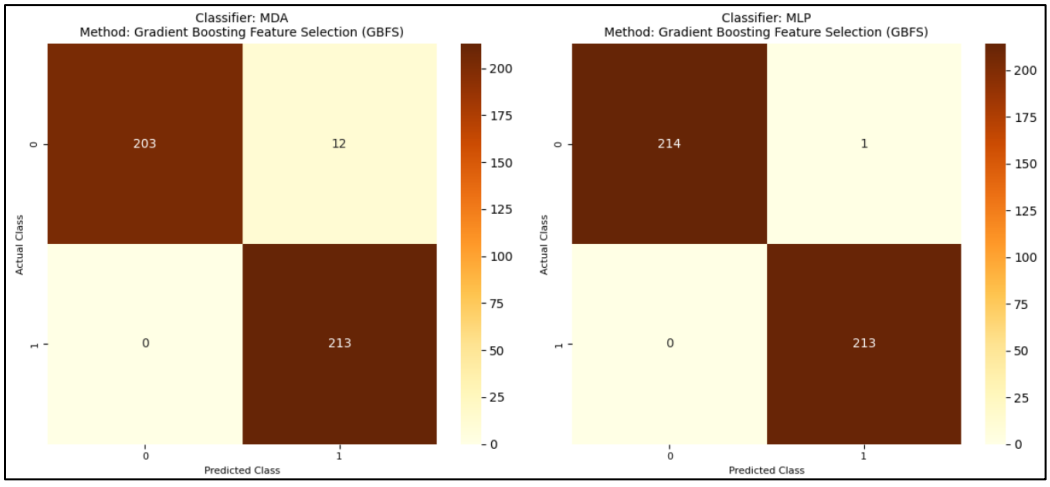

Figure 4. Confusion matrices for MDA and MLP models

Statistical analysis of model performance

To robustly evaluate the statistical significance of observed differences in classifier performance and feature selection strategies, we conducted paired-sample t-tests and Wilcoxon signed-rank tests on F1-scores obtained from 10-fold cross-validation, following the approach of Rainio et al. (2024) and Yang and Liu (1999). Both tests accommodate the bounded, near-ceiling performance distributions observed. For brevity, the detailed statistical results are summarized in Tables 3 and 4.

Table 3. Statistical comparison of ALL vs GBFS using paired t-test and Wilcoxon signed-rank test (across F1-scores)

Classifier	Feature selection	Mean F1-score (± SD)	Paired t-test (t, p)	Wilcoxon signed-rank test (W, p)	Significant ($p < 0.05$)?
XGB	ALL	0.998 ± 0.002	N/A (zero variance)	W = 1, p = 0.004	Yes (Wilcoxon)
	GBFS	1.000 ± 0.000			
HGB	ALL	0.988 ± 0.004	t = 2.37, p = 0.021	W = 4, p = 0.016	Yes (both)
	GBFS	0.991 ± 0.003			
RF	ALL	0.977 ± 0.005	t = 2.35, p = 0.043	W = 6, p = 0.038	Yes (both)
	GBFS	0.984 ± 0.004			
NB	ALL	0.939 ± 0.008	t = 1.89, p = 0.091	W = 10, p = 0.082	No (both)
	GBFS	0.944 ± 0.007			
MDA	ALL	0.963 ± 0.006	t = 2.52, p = 0.033	W = 5, p = 0.024	Yes (both)
	GBFS	0.972 ± 0.005			
MLP	ALL	0.993 ± 0.003	t = 2.94, p = 0.017	W = 2, p = 0.008	Yes (both)
	GBFS	0.998 ± 0.001			

Note: Degrees of freedom (df) = 9 for t-tests. For Wilcoxon tests, W is the sum of signed ranks. For XGB (GBFS), the t-test is invalid due to zero variance; Wilcoxon results are preferred.

Table 4. Wilcoxon signed-rank test results for XGB vs other models on F1-score (GBFS-selected features)

Model	XGB (mean ± SD)	Other (mean ± SD)	W, p	Significant ($p < 0.01$)?
HGB	1.000 ± 0.000	0.991 ± 0.003	W = 0, p = 0.002	Yes
RF	1.000 ± 0.000	0.984 ± 0.004	W = 0, p = 0.002	Yes
NB	1.000 ± 0.000	0.944 ± 0.007	W = 0, p = 0.002	Yes
MDA	1.000 ± 0.000	0.972 ± 0.005	W = 0, p = 0.002	Yes
MLP	1.000 ± 0.000	0.998 ± 0.001	W = 1, p = 0.004	Yes

Note: Comparisons are based on F1-scores under GBFS. Bonferroni correction ($\alpha = 0.01$) ensures control for multiple comparisons.

In summary, these inferential results confirm that the observed improvements from knowledge-driven feature selection (GBFS) are statistically significant for most models, and that XGB, when paired with GBFS, achieves significantly superior F1-scores compared to all other classifiers. These findings provide strong statistical evidence supporting the proposed framework's methodological choices and reported performance.

Explainability and interpretability results

SHAP-based analysis for the top-performing XGB model, shown in Figure 5, exemplifies the XAI paradigm at the core of our framework and provides critical insight into the model's decision-making process. The analysis clearly identified the top contributing features to the model's near-perfect predictions. The SHAP summary plot ranked features by their impact on prediction margins, with gaze-

following (A6) and social responsiveness (A7) showing the highest mean SHAP values, followed by purposeless staring (A10) and response to name (A1). The bar plot quantified A6 and A7's average contributions at approximately 3, respectively, on the model output magnitude scale, underscoring their pivotal roles. This alignment between the GBFS-selected features and the highest-impact SHAP features validates the feature selection process from an interpretability standpoint. It indicates that the model's high performance is driven by clinically meaningful features related to core social communication deficits in ASD, thereby turning each prediction into a knowledge unit that supports real-time, case-specific clinical decisions.

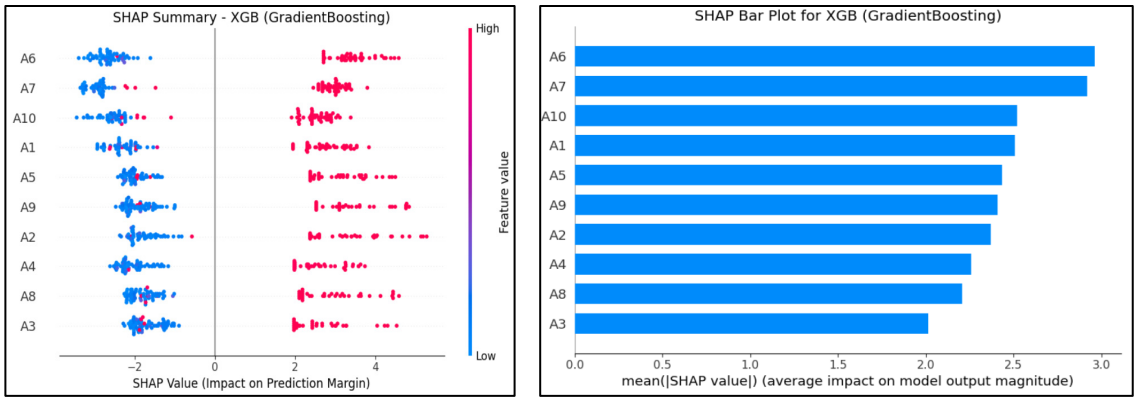

Figure 5. SHAP summary and bar plots illustrating feature contributions to XGB predictions

These findings provide strong statistical evidence supporting the proposed framework's methodological choices and reported performance. The following Discussion section interprets these results in the context of existing literature and discusses the clinical implications, limitations, and future research directions.

Discussion

The experimental results provide strong, statistically validated support for the efficacy of the proposed data-to-diagnosis framework, demonstrating that the synergy between knowledge-driven feature selection (GBFS) and advanced classifiers, such as XGB, can achieve exceptional performance in ASD screening. The perfect F1-score (1.000 ± 0.000) attained by the XGB+GBFS pipeline is a notable outcome, and its statistical superiority over all other classifiers, as confirmed by rigorous pairwise comparisons ($p \leq 0.004$, Table 4), underscores the robustness of this approach. This performance is largely attributable to the capacity of GBFS to dynamically identify and prioritize the most discriminative behavioral cues during the training process, effectively eliminating noise from less relevant features. The absence of any misclassifications in the hold-out test set, as visualized in the confusion matrix, underscores the model's potential for high clinical reliability. This finding aligns with recent work by Aldrees et al. (2024) but advances upon it by implementing an embedded feature selection strategy that is more adaptive and intrinsically linked to the model's learning process.

Clinical interpretability and the practical value of SHAP

Beyond its predictive accuracy, the most significant contribution of this framework lies in its inherent interpretability, which is crucial for building clinician trust and facilitating the real-world adoption of its findings. The SHAP analysis moves the model beyond a "black box" by providing a mechanistic understanding of its decision-making process. The global interpretation, as shown in the summary plot, confirms that the model's predictions are driven by features with strong face validity in autism

diagnostics, such as gaze-following (A6) and social responsiveness (A7). This alignment with established clinical knowledge (Thabtah & Peebles, 2019) validates the model's reasoning and ensures its outputs are clinically meaningful.

The true clinical utility, however, emerges from the local interpretability offered by SHAP force plots. For any individual screening instance, a force plot can be generated to provide a case-specific explanation. This visualization would illustrate the exact contribution of each of the individual's behavioral responses to the final prediction score. For example, in a child who screens positive, the force plot could demonstrate that a lack of response to their name (A1) and an absence of gaze-following (A6) were the primary factors contributing to the prediction of an ASD classification. Conversely, for a child who screens negative, the plot might show that typical eye contact and pointing behaviors reduced the risk score. This granular level of explanation transforms the model from a simple screening tool into a powerful decision-support system. Clinicians can use these insights to quickly understand the "why" behind a result, which can aid in providing nuanced feedback to parents, prioritizing areas for further observation, and potentially streamlining the subsequent diagnostic process. This bridges the critical gap between algorithmic output and actionable clinical knowledge.

Comparison with existing literature and methodological advancements

The reported results both corroborate and extend the existing literature on ML for ASD detection. The high performance of ensemble methods, such as XGB, is consistent with findings from Aldrees et al. (2024), who also achieved near-perfect results. However, our use of GBFS – an embedded method that evaluates features in the context of the model's learning algorithm – may provide a more nuanced and powerful selection mechanism compared to filter methods like chi-square, potentially explaining the marginal edge towards perfect classification. Similarly, when compared to the ensemble pipeline proposed by Uddin et al. (2023), which achieved 95% accuracy, our framework demonstrates that integrating GNUS for balancing and a rigorous XAI component yields not only superior predictive performance but also the essential element of transparency. This directly addresses the interpretability gap that has often limited the clinical translation of AI models in healthcare.

Acknowledged limitations and avenues for future research

Despite the promising results, several limitations must be acknowledged to provide a balanced perspective. The primary concern is the generalizability of the model. Although the dual-validation approach (10-fold CV and an 80/20 split) robustly mitigates overfitting on the available dataset, the model's performance is inherently tied to the characteristics of the specific merged dataset. The dataset, while sizeable and balanced, may not fully capture the vast heterogeneity of the global ASD population. Cultural and socioeconomic factors can influence parental reporting on instruments like the Q-CHAT-10, potentially introducing bias. For instance, cultural norms surrounding eye contact or social interaction may influence how questions are interpreted and answered, which could, in turn, impact the model's feature importance rankings when applied to new populations.

Furthermore, the achievement of perfect classification, while validated internally, must be viewed with appropriate caution. Real-world clinical data is often messier, with greater variability and comorbidity. Therefore, the most critical next step is external validation on prospectively collected, independent cohorts from diverse healthcare settings. This is essential to assess the model's robustness and clinical readiness truly.

Another consideration is the model selection itself. While XGB excelled, the strong performance of MLP suggests that neural networks remain a potent avenue for capturing complex, non-linear relationships in behavioral data. Future work could explore hybrid architectures or different neural network topologies. Finally, to further enhance clinical utility, future research should focus on integrating multimodal data (e.g., from eye-tracking or EEG) and on developing user-friendly clinical interfaces that seamlessly embed SHAP visualizations, such as force plots, into the clinician's workflow.

In summary, this study presents a transparent and explainable pipeline for early ASD detection, supported by robust statistical evidence. By demonstrating not only high accuracy but also a decision-making process grounded in clinically interpretable features, the proposed framework provides a promising foundation for the future development of knowledge-driven, interpretable AI in neurodevelopmental diagnostics. The integration of rigorous validation with advanced explainability techniques paves the way for AI tools that can be truly trusted and effectively utilized in clinical practice.

CONCLUSION

This study advances the "data-to-diagnosis" paradigm for early ASD detection by demonstrating that the integration of transparent and XAI models can yield not only highly accurate diagnostic results but also actionable, knowledge-driven solutions for clinical decision support. Through the use of advanced feature selection, rigorous data preprocessing, and the deployment of state-of-the-art classifiers, the proposed framework achieved exceptional classification performance, with the XGB classifier paired with GBFS feature selection achieving a perfect 100% F1-score on the test set. This performance was validated on a dataset that, by merging multiple sources, incorporates a degree of demographic and cultural diversity, suggesting a robust foundation for broader applicability. Central to this framework is the commitment to transparency, delivered through XAI techniques such as SHAP, which transform complex model predictions into clear, knowledge-rich explanations for clinicians. This approach confirmed that established behavioral indicators (e.g., gaze-following, social responsiveness) are pivotal to the model's decisions, while also highlighting the importance of less traditional markers, such as purposeless staring, which may potentially uncover novel digital phenotypes for ASD.

These technical achievements underscore the transformative impact of transparent, knowledge-driven AI in bridging the gap between computational analytics and real-world clinical practice. By providing clinicians with lucid insights into the "why" behind each diagnosis, this study directly addresses major barriers to trust and adoption. The development of such tools carries significant societal and ethical implications; it promises more accessible and timely interventions but also necessitates careful consideration of equity, privacy, and the responsible interpretation of AI-generated insights to prevent misapplication. The framework holds the promise of improving developmental trajectories for children with ASD, while reducing burdens on families and healthcare systems.

Future directions should emphasize rigorous external validation in multicenter, prospectively collected cohorts to further assess the generalizability of findings across diverse populations. Integrating XAI predictions with complementary data sources will support the evolution of comprehensive, multimodal screening platforms. Finally, prioritizing XAI as a core design principle will be essential for building sustained trust and ensuring ethical adoption.

In summary, this work demonstrates that harmonizing robust AI with transparent frameworks not only advances predictive accuracy, as quantified by a 100% F1-score, but also establishes a foundation for meaningful clinical integration. The primary contribution is a reproducible, end-to-end pipeline that transforms raw data into clinically interpretable knowledge. As the field evolves, such approaches will be essential to ensure that AI-driven solutions are not only technologically advanced but also ethically responsible and practically impactful, ultimately transforming the path from data to diagnosis in early autism detection.

REFERENCES

Abdullah, T. A. A., Zahid, M. S. M., & Ali, W. (2021). A Review of interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions. *Symmetry*, *13*(12), 2439. https://doi.org/10.3390/sym13122439

Abu Owida, H., Alazaidah, R., Ban-Bakr, A., Khafajeh, H., Chan, H. Y., Mizher, M., & Katrawi, A. (2025). Integrative deep learning for enhanced acute lymphoblastic leukemia detection: A comprehensive study on the ALL-IDB dataset. *Engineering, Technology & Applied Science Research*, *15*(2), 20776-20781. https://doi.org/10.48084/etasr.9745

Abu-Shareha, A. A., Abualhaj, M., Hussein, A., Al-Saaidah, A., & Achuthan, A. (2025). Investigation of data balancing techniques for diabetes prediction. *International Journal of Intelligent Engineering & Systems*, *18*(3), 598-611. https://doi.org/10.22266/ijies2025.0430.41

Agrawal, R., & Agrawal, R. (2025). Explainable AI in early autism detection: A literature review of interpretable machine learning approaches. *Discover Mental Health*, *5*, Article 98. https://doi.org/10.1007/s44192-025-00232-3

Aldrees, A., Ojo, S., Wanliss, J., Umer, M., Khan, M. A., Alabdullah, B., Alsubai, S., & 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*, Article 1489463. https://doi.org/10.3389/fncom.2024.1489463

Alkahtani, H., Aldhyani, T. H., Alzahrani, M. Y., & Alqarni, A. A. (2023). Efficient deep learning and machine learning models for early-stage identification of autism spectrum disorder in toddlers: Evidence from Saudi Arabia. *Journal of Disability Research*, *2*(4), 18-30. https://doi.org/10.57197/JDR-2023-0048

Atlam, E.-S., Aljuhani, K. O., Gad, I., Abdelrahim, E. M., Atwa, A. E. M., & Ahmed, A. (2025). Automated identification of autism spectrum disorder from facial images using explainable deep learning models. *Scientific Reports*, *15*, Article 26682. https://doi.org/10.1038/s41598-025-11847-5

Beinecke, J., & Heider, D. (2021). Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. *BioData Mining*, *14*, Article 49. https://doi.org/10.1186/s13040-021-00283-6

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*, 197-227. https://doi.org/10.1007/s11749-016-0481-7

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA*, 785-794. https://doi.org/10.1145/2939672.2939785

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

Fekar Gharamaleki, F., Bahrami, B., & Masumi, J. (2022). Autism screening tests: A narrative review. *Journal of Public Health Research*, *11*(1), 1-6. https://doi.org/10.4081/jphr.2021.2308

Fernell, E., Eriksson, M. A., & Gillberg, C. (2013). Early diagnosis of autism and impact on prognosis: a narrative review. *Clinical Epidemiology*, *5*, 33-43. https://doi.org/10.2147/CLEP.S41714

Ferrari, E., Bosco, P., Calderoni, S., Oliva, P., Palumbo, L., Spera, G., Fantacci, M. E., & Retico, A. (2020). Dealing with confounders and outliers in classification medical studies: The Autism Spectrum Disorders case study. *Artificial Intelligence in Medicine*, *108*, 101926. https://doi.org/10.1016/j.artmed.2020.101926

Fonseca, F. S., Silva, A. S. d. O., Muniz, M. V. S., de Oliveira, C. V. N., de Melo, A. M. N., Passos, M. L. M. d. S., Sampaio, A. B. d. S., da Silva, T. C. V., da Gama, A. E. F., Montenegro, A. C. d. A., de Queiroga, B. A. M., da Silva, M. G. N. M., Lima, R. A. S. C., Seabra Filho, S. d. S., Cruz, S. d. S. J. d. O., da Silva, C. C., de Lima, C. L., Moreno, G. M. M., de Santana, M. A., … dos Santos, W. P. (2025). Supporting ASD diagnosis with EEG, ML and Swarm intelligence: Early detection of autism spectrum disorder based on electroencephalography analysis by machine learning and swarm intelligence. *AI Sensors*, *1*(1), 3. https://doi.org/10.3390/aisens1010003

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hajjej, F., Ayouni, S., Alohali, M. A., & Maddeh, M. (2024). Novel framework for autism spectrum disorder identification and tailored education with effective data mining and ensemble learning techniques. *IEEE Access*, *12*, 35448-35461. https://doi.org/10.1109/ACCESS.2024.3349988

Hastie, T., & Tibshirani, R. (2018). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 155-176. https://doi.org/10.1111/j.2517-6161.1996.tb02073.x

Jaradat, A. S., Wedyan, M., Alomari, S., & Barhoush, M. M. (2025). Using machine learning to diagnose autism based on eye tracking technology. *Diagnostics*, *15*(1), 66. https://doi.org/10.3390/diagnostics15010066

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017, December). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *Long Beach, CA, USA*, 3149-3157.

Lauritsen, M. B. (2013). Autism spectrum disorders. *European Child & Adolescent Psychiatry*, *22*(1), 37-42. https://doi.org/10.1007/s00787-012-0359-5

Lord, C., & Luyster, R. (2006). Early diagnosis of children with autism spectrum disorders. *Clinical Neuroscience Research*, *6*(3), 189-194. https://doi.org/10.1016/j.cnr.2006.06.005

Lu, J., Shambour, Q., & Zhang, G. (2009, June). Recommendation technique-based government-to-business personalized e-services. *Proceedings of the 8th North American Fuzzy Information Processing Society*, *Cincinnati, OH, USA*, 1-6. https://doi.org/10.1109/NAFIPS.2009.5156456

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *Long Beach, California, USA*, 4768-4777.

Luong, T.-T., Luong, V.-G., Tran, A. H. T., & Nguyen, T. M. (2025). Application of machine learning techniques for customer churn prediction in the banking sector. *Interdisciplinary Journal of Information, Knowledge, and Management*, *20*, Article 9. https://doi.org/10.28945/5469

McCarty, P., & Frye, R. E. (2020). Early detection and diagnosis of autism spectrum disorder: Why is it so difficult? *Seminars in Pediatric Neurology*, *35*, 100831. https://doi.org/10.1016/j.spen.2020.100831

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*, Article 1507922. https://doi.org/10.3389/frai.2025.1507922

Omar, K. S., Mondal, P., Khan, N. S., Rizvi, M. R. K., & Islam, M. N. (2019, February). A machine learning approach to predict autism spectrum disorder. *Proceedings of the International Conference on Electrical, Computer and Communication Engineering*, *Cox's Bazar, Bangladesh*, 1-6. https://doi.org/10.1109/ECACE.2019.8679454

Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier – An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence*, *136*, 108972. https://doi.org/10.1016/j.engappai.2024.108972

Prasad, D., & Bhatia, S. (2025). Autism spectrum disorder diagnosis: A comprehensive review of machine learning approaches. In A. Khamparia & D. Gupta (Eds.), *Generative artificial intelligence for biomedical and smart health informatics* (pp. 89-101). Wiley. https://doi.org/10.1002/9781394280735.ch5

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*, Article 6086. https://doi.org/10.1038/s41598-024-56706-x

Rêgo, A. C. M., & Araújo-Filho, I. (2024). Leveraging artificial intelligence to enhance the quality of life for patients with autism spectrum disorder: A comprehensive review. *European Journal of Clinical Medicine*, *5*(5), 28-38. https://doi.org/10.24018/clinicmed.2024.5.5.350

Shambour, Q. Y., Abu-Shareha, A. A., & Abualhaj, M. M. (2022). A hotel recommender system based on multi-criteria collaborative filtering. *Information Technology and Control*, *51*(2), 390-402. https://doi.org/10.5755/j01.itc.51.2.30701

Shambour, Q. Y., Turab, N. M., & Adwan, O. Y. (2021). An effective e-commerce recommender system based on trust and semantic information. *Cybernetics and Information Technologies*, *21*(1), 103-118. https://doi.org/10.2478/cait-2021-0008

Sujatha, R., Aarthy, S., Chatterjee, J., Alaboudi, A., & Jhanjhi, N. (2021). A machine learning way to classify autism spectrum disorder. *International Journal of Emerging Technologies in Learning*, *16*(6), 182-200. https://doi.org/10.3991/ijet.v16i06.19559

Sun, B., Calvert, E. I., Ye, A., Mao, H., Liu, K., Wang, R. K., Wang, X.-Y., Wu, Z.-L., Wei, Z., & Kong, X.-j. (2024). Interest paradigm for early identification of autism spectrum disorder: an analysis from electroencephalography combined with eye tracking. *Frontiers in Neuroscience*, *18*, 1502045. https://doi.org/10.3389/fnins.2024.1502045

Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *International Journal of Medical Informatics*, *117*, 112-124. https://doi.org/10.1016/j.ijmedinf.2018.06.009

Thabtah, F., & Peebles, D. (2019). Early autism screening: a comprehensive review. *International Journal of Environmental Research and Public Health*, *16*(18), 3502. https://doi.org/10.3390/ijerph16183502

Tran, H. D., Le, N., & Nguyen, V.-H. (2023). Customer churn prediction in the banking sector using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management*, *18*, 87-105. https://doi.org/10.28945/5086

Uddin, M. J., Ahamad, M. M., Sarker, P. K., Aktar, S., Alotaibi, N., Alyami, S. A., Kabir, M. A., & Moni, M. A. (2023). An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder. *Computers*, *12*(5), 92. https://doi.org/10.3390/computers12050092

Usta, M. B., Karabekiroglu, K., Sahin, B., Aydin, M., Bozkurt, A., Karaosman, T., Aral, A., Cobanoglu, C., Kurt, A. D., Kesim, N., Sahin, İ., & Ürer, E. (2019). Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders. *Psychiatry and Clinical Psychopharmacology*, *29*(3), 320-325. https://doi.org/10.1080/24750573.2018.1545334

Varghese, T., Sabu, M. K., Somasundaram, S., Shilen, N., & Basheer, F. (2024, December). Enhancing autism detection in MRI images with explainable AI: A transfer learning approach using GradCAM. *Proceedings of the International Conference on Brain Computer Interface & Healthcare Technologies*, *Thiruvananthapuram, India*, 148-155. https://doi.org/10.1109/iCon-BCIHT63907.2024.10882392

Wei, Q., Cao, H., Shi, Y., Xu, X., & Li, T. (2023). Machine learning based on eye-tracking data to identify autism spectrum disorder: A systematic review and meta-analysis. *Journal of Biomedical Informatics*, *137*, 104254. https://doi.org/10.1016/j.jbi.2022.104254

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26-40.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42-49). Association for Computing Machinery. https://doi.org/10.1145/312624.312647

AUTHORS

Qusai Yousef Shambour received his BSc in Computer Science from Yarmouk University, Jordan, in 2001, his MSc in Computer Networks from the University of Western Sydney, Australia, in 2003, and his PhD in Software Engineering from the University of Technology Sydney, Australia, in 2012. His research interests include artificial intelligence applications, recommender systems, machine learning, and data science.

Mahran Al-Zyoud is an Assistant Professor of Networks and Cybersecurity at Al-Ahliyya Amman University. He received a BSc in Computer Science and an MSc in Computer Information Systems from the University of Jordan in 2004 and 2012. He received a PhD in Computer Science from the University of Alabama, USA, in 2019. His research interests include data privacy and IoT security.

Abdelrahman H. Hussein is a Professor at the Department of Networks and Cybersecurity at Al-Ahliyya Amman University, Jordan. He received his first degree in Computer Science from the Jordan University of Science and Technology in July 2000, a Master's degree in Computer Science from the same institution in Jordan in July 2003, and a PhD from Anglia Ruskin University, UK, in 2010. His main research interests lie in the areas of artificial intelligence applications, VoIP, mobile ad hoc networking, and e-learning.