Ensemble Learning Approach for Early Prediction of Autism Spectrum Disorder in Children

Samuel Akpan Robinson[1], Raphael Henshaw Ekpo[2*], Ukeme Donatus Archibong[3], Ini Umoeka[4], Samuel Pius Etim[5]
1 Department of Cyber Security, Faculty of Computing, University of Uyo, Uyo, Nigeria.
2. Department of Computer Science, Lagos State University, of Science and Technology, Ikorodu, Lagos State, Nigeria.
3 Department of Applied Chemical Science Laboratory Technology, Faculty of Science Laboratory Technology, University of Benin, Benin City.
4. Department of Software Engineering, University of Uyo, Uyo, Nigeria.
5. Department of Information Systems, Faculty of Computing, University of Uyo, Uyo, Nigeria.

Abstract—This study presents the development and evaluation of an ensemble learning model for the early prediction of autism spectrum disorder (ASD). By integrating machine learning algorithms, the study aims to improve the detection of ASD-positive cases. Using a dataset comprising behavioral and demographic attributes, comprehensive preprocessing, descriptive analysis, and model evaluations were conducted on random forest (RF) and extreme gradient boosting (XGBoost). The results yield a high accuracy of 82% - 84% in identifying non-ASD cases, with F1-scores nearing 90%. However, the model showed moderate sensitivity in detecting ASD-positive cases due to overlapping features and data imbalance. Despite this limitation, the ensemble model outperformed individual classifiers, signifying its effectiveness for real-world screening applications. The study highlights the importance of ensemble learning, in building robust diagnostic systems. It contributes a practical, reproducible framework for integrating machine learning into neurodevelopmental disorder screening, advocating for its adoption in health informatics.

Index Terms— Autism, disorder, ensemble learning, spectrum.

I. INTRODUCTION

Autism is a neurodevelopmental disorder marked by communication challenges, social deficits, repetitive behaviors, and restricted interests (McDonald *et al*., 2018). The ASD varies in severity and often emerges before age three, with some children showing difficulties in communication between 18 to 24 months. Diagnosis depends on social communication impairments and sensory-motor behaviors (Lord *et al*., 2018). This often linked to co-occurring conditions like gastrointestinal and sleep issues (Ferina *et al*., 2023). Increasing adult diagnoses show ASD's variability and diagnostic challenges (Huang *et al.,* 2020). Methods such as computer-based interventions aid visual learning, structured environments (Ramdoss *et al*., 2012). This improves outcomes as compare to maladaptive traditional diagnostic methods which rely on expert assessments. However, these methods making are time-consuming and subjective. Genetic and neurobiological studies seek biomarkers but face cost and accessibility barriers. Machine learning (ML) offers a promising alternative, leveraging behavioral, eye-tracking, and neuroimaging data to enhance accuracy (Farooq *et al.,* 2023)

Early ASD prediction faces challenges such as limited data size, diagnostic subjectivity, and variability in symptoms. Though some ML models depend on small dataset produces poor predictive outcome, subjective to expert assessments. This may cause misdiagnosis and delayed intervention thus further complicate detection. The challenges of high computational demands, high costs of accessibility, especially in low-income regions contribute to poor detection of ASD. Many caregivers are unaware of early symptoms or screening tools, delaying diagnosis and intervention due to wrong detection tools. This further complicates timely detection. The AI-driven models show promising results but require extensive validation before clinical use. Addressing these issues requires large dataset. ASD prediction models often depend on diverse datasets, including behavioral patterns, clinical records, and genetic information, which are not only vast in size but also frequently imbalanced. This imbalance occurs because the number of children diagnosed with ASD may be lower than those without ASD. This may lead to skewed datasets that can hinder the model's ability to identify patterns accurately (Robinson *et al.*, 2025). Furthermore, the complexity of the data encompassing varied behavioral and cognitive traits adds another layer of difficulty, requiring sophisticated models to handle the intricate relationships between different features.

This issue arises from factors, including the reliability of the measures used to assess ASD, the variability inherent in the data, and the presence of outliers that can distort the model's learning process. Such low correlations make it difficult for conventional machine learning models to extract meaningful insights from the data, as they struggle to identify the key factors that distinguish children with ASD from those without. This lack of clarity between input features and the prediction target further complicates the diagnostic process.

To address these challenges, this study proposes an ensemble learning model that combines the strengths of RF and XGBoost techniques for early ASD prediction. The model addresses the problem of low data correlation by leveraging the complementary strengths to detect subtle patterns in the data and providing a more reliable early diagnosis of ASD in children.

The work is presented in different Sections, while Section 2 is the literature review; Section 3 is the research methodology; Section 4 presents results and discussion, and Section 5 is the conclusion and future work.

II. RELATED WORKS

This study in Farooq *et al*., (2023) used a machine learning and fuzzy logic (FL) approach for the first time to analyze severe ASD in both children. The study gathered over 600 records of affected adults and children from four different ASD datasets across various sources to extract relevant features. The results show that 80% of adults have ASD and 98% of children suffer from ASD. The work is limited to the imbalance of data variables. Various models, including Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), Neural Networks, Random Forest Classifier (RFC), Naïve Bayes (NB), Logistic Regression (LR), and KNN and Convolutional Neural Networks, were used to predict and analysed ASD issues in children, adolescents, and adults, and the results are efficient but limited by high computational cost (Raj & Masood, 2020; Robinson *et al*., 2024). Rogala *et al.* (2023) conducted a review using two different association measures and applied both classical statistical methods and machine learning techniques. The integration of statistical analysis and traditional machine learning methods enhanced the understanding of predictive models based on the spectral or network features of a subject's electroencephalography (EEG) data. This enhances the identification and validation of early detection of ASD symptoms in children. However, limited to high computational cost. Damianos *et al* (2023) provides an overview of the machine learning and artificial intelligence algorithms applied to ASD diagnosis and prediction across different age groups using clinical approaches. Boughattas *et al* (2022) used RF with transfer learning and deep neural networks to achieve ASD classification accuracies of 98.9% and 99.8%, respectively. The findings show that machine learning approaches are efficient in ASD classification compared to traditional autism screening methods.

Shekarro *et al* (2024) considered ASD as a type of neurodevelopmental disorder that is typically identified during early childhood. The authors utilized a predictive ASD model, such as Forest (RF), Bernoulli Naive Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel on the dataset of toddlers. The predictive analysis of ASD is carried out in two stages: First, ASD diagnosis is conducted utilizing classification algorithms. To improve the performance of the ASD prediction model, the voting ensemble learning approach with three classifiers from the base learners is used. The findings show that the proposed ensemble-based prediction models outperform basic Machine Learning (ML) approaches. Though it is computationally intensive. Kavadi *et al.* (2024) investigated the issues faced by the healthcare sector in managing and processing huge volumes of unstructured, real-time medical data. This study employed a big data and machine learning-based medical data classification (BDML-MDCASD) model aimed at improving the effectiveness of ASD diagnosis. Simulation results show that the BDML-MDCASD method outperforms traditional methods, achieving a classification accuracy of 92%, precision of 90%, and recall of 93%. Alqaysi *et al.* (2022) ASD is a complicated neurobehavioral disorder that affects linguistic and behavioral skills as well as communication. The authors gathered information from the IEEE Xplore digital library, Science Direct (SD), Web of Science (WoS), and Scopus databases. A definitive collection of 40 papers based on inclusion and exclusion criteria is compiled from a set of 944 articles published between 2017 and 2021. The chosen publications were categorized according to the evidence's goal and objective. This provides an insight into ASD among children, but does not provide a measure of early detection for prevention. In Alam *et al* (2022), parameters such as inheritance and surroundings, which influence the growth of neurodevelopmental disorders for 36 months of life, were collected and analyzed. The results of ASD diagnoses depend on traditional clinical assessments from the last few decades. These traditional methods are based on massive data collection from multiple respondent responses and the

III. RESEARCH METHODOLOGY

Figure 1 presents the proposed XGBoost-RF-based framework for early prediction of ASD in children. It consists of data collection, preprocessing of ASD data, splitting of data, model building, and model evaluation.

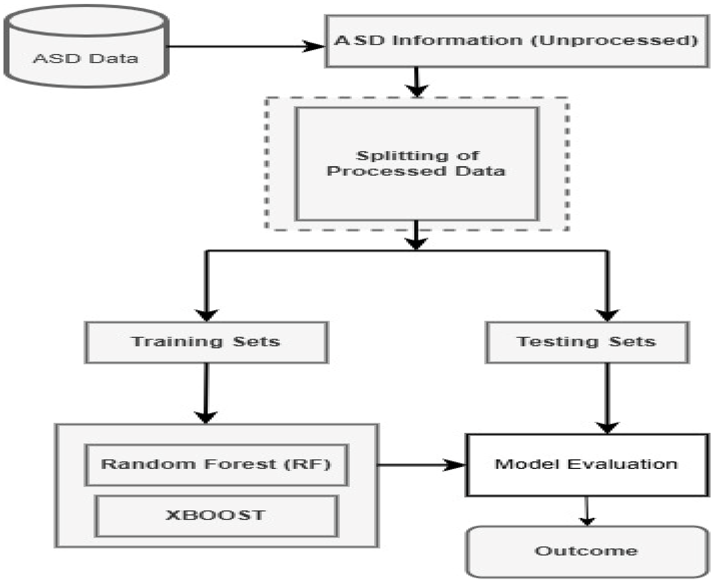

Figure 1: Proposed RF-XGBoost Framework

A. Data Collection

The ASD data was collected from the Kaggle repository consists of 800 rows and 22 fields consisting of, 19 categorical, and 3 numerical variables. These variables contain noisy and missy that to be preprocessed. Table 1 presents a set of features and these features enable a robust framework for training machine learning models aimed at the early identification of ASD in children.

Table 1: Features of ASD in Children

Values	Datatype	Description
Child_ID	String	Unique ID for each child
Age	Integer	Age of the child (2–12 years)
Gender	String	Male or Female
Jaundice	Boolean	Whether the child had neonatal jaundice
Family_ASD_History	Boolean	Any family history of ASD
Language_Delay	Boolean	Delay in speech development
Social_Interaction_Score	Integer	1–10 scale measuring social ability
Communication_Score	Integer	1–10 scale for communication skills
Repetitive_Behavior_Score	Integer	1–10 scale for repetitive behavior
Diagnosed_ASD	String	Final ASD diagnosis (Yes/No)
Date_Recorded	Date	The date when the data was recorded

B. Data Preprocessing

ASD contains missing values that that is preprocessed to remove inconsistent data. To address this, mean imputation is employed, where missing entries in a particular feature are replaced by the average value of the observed data for that feature. This method ensures the dataset's integrity is protected against loss of information during analysis as depicted in Equation 1 (Inyang *et al*., 2021)

$$X_i = \frac{1}{n}\sum_j^n X_{ij} \qquad (1)$$

Where, X_i is the imputed value for the missing observation, $\frac{1}{n}\sum_j^n X_{ij}$ is the sum of all observed values for the feature across n available samples, n is the count of observed values (Imianvan and Robinson. 2024)

C. Correlation Analysis

Correlation analysis determined the degree of relationship between two variables (Imianvan *et al*., 2024). The correlation analysis helps to identify which variables significantly contribute to early diagnosis. The Pearson correlation coefficient, r is used to determine the degree of linear association between two continuous variables and ranges from -1 to +1. A value of +1 indicates a perfect positive linear correlation, -1 indicates a perfect negative correlation, and 0 implies no linear relationship, as shown in Equation 2 (Shrivastava *et al*., 2024)

$$\boldsymbol{r} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\cdot\sqrt{\sum(y_i-\bar{y})^2}} \qquad (2)$$

Where, $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$: Individual data points of variables $X\ and\ Y$, $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$: Mean values of variables $X\ and\ Y, r$: Pearson correlation coefficient**,** $\sum$ is the sum of data points.

D. . XGBOOST Model

The objective function $L(\theta)$ for XGBoost is formulated. This combines with a loss function and regularization to minimize the loss of overall training samples and a regularization as shown in Equation 3

$$L(\theta) = \sum_{i=1}^{n} loss(y_i, \hat{y_i}) + \sum_{j=1}^{m} \Omega(f_j) \qquad (3)$$

Where, y_i is the true value for the i-th sample $\hat{y_i}$ is the predicted value for the i-th sample as depicted, $loss(y_i, \hat{y_i})$ the loss function, which could be mean squared error (MSE) for regression or logistic loss for classification, $\Omega(f_j)$ is the regularization term for the j-th tree, f_j represents the j-th decision tree. The regularization term $\Omega(f_j)$ helps prevent overfitting by penalizing more complex models, as shown in Equation 4

$$\Omega(f_i) = \Upsilon T + \frac{1}{2}\lambda\sum_{k=1}^{T} w_k^2 \qquad (4)$$

Where T is the number of leaves in the j-th tree, ϒ and λ are regularization parameters, w_k is the weight of leaf k as depicted in Figure 2

```
1. Initialize the regularization value.
       Set $\Omega = 0$
2. Compute the penalty for the number of leaves:
       $\Omega \leftarrow \Omega + \Upsilon \cdot T$
3. Initialize the sum of squared weights:
       weight_sum $= 0$
4. For each leaf $k = 1\ to\ T$ :
       a. Compute the square of leaf weight: $w_k^2$
       b. Add to sum: weight_sum $\leftarrow$ weight_sum $+ w_k^2$
5. Compute weight regularization term:
       $\Omega \leftarrow \Omega + \frac{1}{2}\lambda \cdot$ weight_sum
6. Return $\Omega(f_j)$as the total regularization penalty
```

Figure 2: XGBoost Algorithm for ASD

For each boosting iteration, the objective is to optimize the following function with respect to the model's parameters in Equation 5

$$Obj(t) = \sum_{i=1}^{n}\left[g_i * \hat{y_i} + \frac{1}{2}h_i * \hat{y}_i^2\right] + \Omega(f) \qquad (5)$$

g_i and h_i are the gradient and Hessian of the loss function for the i-th sample, $\Omega(f)$ is the regularization term. The prediction $\hat{y_i}$ for an input vector x_i (where x_i includes features) is computed as the sum of the outputs of all decision trees, as shown in Equation 6

$$\hat{y_i} = \sum_{i=1}^{k} f_k\ (x_i) \qquad (6)$$

K is the number of trees, f_k is the prediction of the k-th tree for the input (x_i)

E. Random Forest (RF) Model

The RF is an ensemble learning method used for both classification and regression tasks (Robinson *et al*., 2025). It is a random vector in p-dimensional space, $X = (X1, \ldots, Xp)T$ representing dataset input or predictor variables and a random variable Y representing the real-valued response. The unknown joint distribution, $PXY\ (X, Y)[21]$. The goal is to find a prediction function $f\ (X)$ for predicting Y.

In this study ensemble learning framework that combines RF and XGBoost for the early prediction of ASD is utilized. The motivation for this hybridization stems from the complementary strengths of the two algorithms. Random Forest is a bagging-based ensemble method that builds multiple decision trees using random subsets of data and features. It is highly effective at reducing variance, handling noisy data, and avoiding overfitting, which makes it particularly robust in real-world clinical and behavioral datasets. However, RF may suffer from limited sensitivity when detecting minority classes, such as ASD-positive cases, especially in imbalanced datasets.

On the other hand, XGBoost is a boosting-based algorithm that sequentially builds decision trees, optimizing the model by focusing on difficult-to-classify instances. Its gradient boosting mechanism, coupled with regularization, improves bias reduction and enhances predictive accuracy on complex, non-linear data. Despite these advantages, XGBoost can be sensitive to noise and prone to overfitting when applied to small or imbalanced datasets.

The ensemble RF–XGBoost model leverages RF's stability and robustness with XGBoost strong predictive capability and feature learning capacity. This combination is suited to ASD prediction, where overlapping behavioural and demographic features can complicate classification. The hybrid model addresses three major challenges: class imbalance, overfitting risk, and complex feature interactions.

IV. RESULTS AND DISCUSSION

Figure 3 illustrates the frequency distribution of two categories related to a classification, which are non-ASD (0) and ASD (1) cases. The x-axis labels these categories as 0 and 1. The y-axis measures the count, ranging from 0 to 600. The bar for category 0 extends to approximately 550-600, indicating a higher number of instances, whereas the bar for category 1 reaches around 200, reflecting a much lower count. This distribution highlights an imbalanced dataset, with the majority of instances belonging to category 0, providing a clear visual comparison between the two classes.

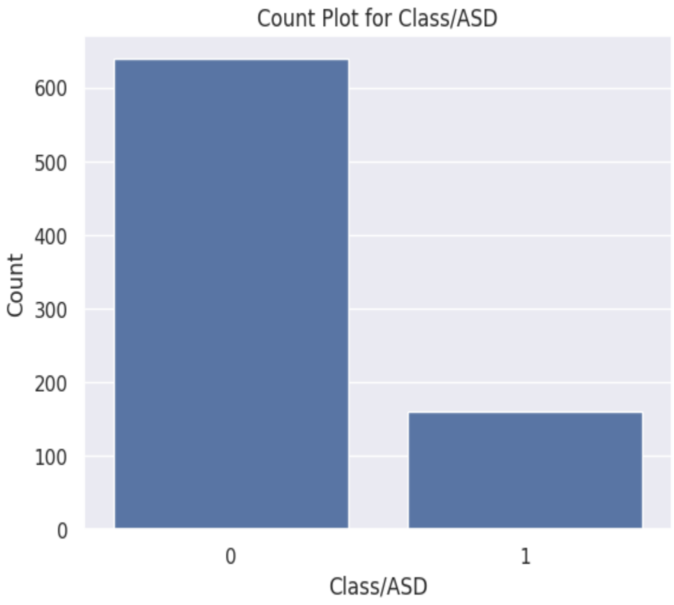

Figure 3: ASD Classification Count

Figure 4 is a composite of multiple histograms, each depicting the distribution of different features related to a dataset. The top row includes histograms for $A1\_Score, A2\_Score, A3\_Score$, and $A4\_Score$, each with counts ranging up to 400 and x-axes scaled from 0 to 1. The bell-shaped distributions with peaks around 0.4 to 0.6 indicate varying concentrations of scores. The second-row features $A5\_Score, A6\_Score, A7\_Score$, and $A8\_Score$, following a similar pattern with counts up to 400 and distributions centered around 0.4 to 0.6. The third row includes $A9\_Score, A10\_Score, age$, and result, where age ranges from 0 to 80 with a peak around 20-30, and result spans from -5 to 15 with a peak around 0-5, both showing skewed distributions. The bottom row contains a single histogram for Class/ASD, with counts up to 600 and a strong peak near 0, with a minor presence near 1. However, the histograms provide a comprehensive overview of the dataset's feature distributions.

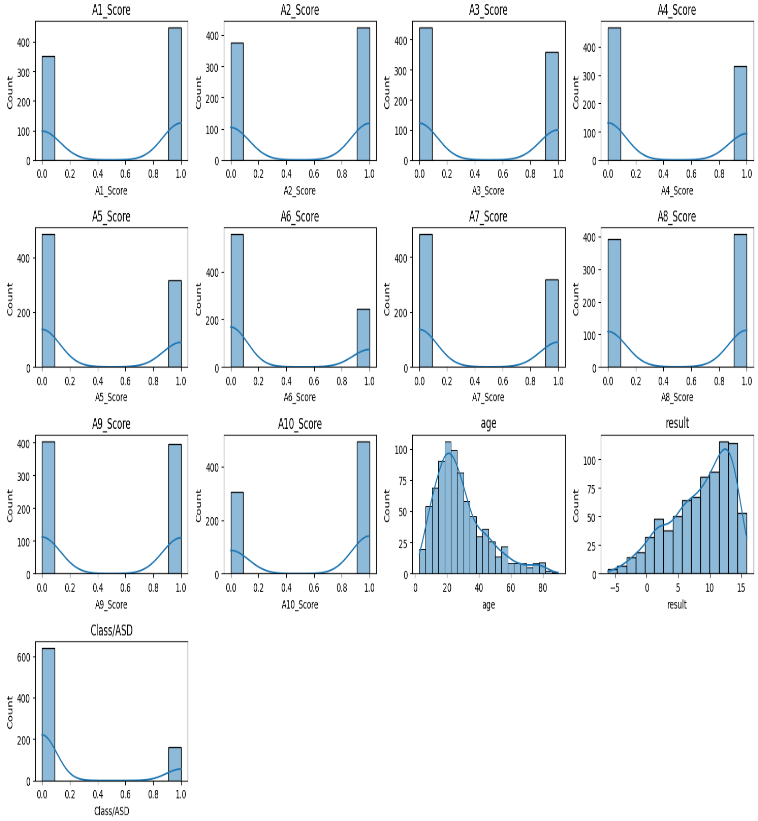

Figure 4: Univariate Distribution Analysis for Early ASD Prediction

Figure 5 is a color-coded matrix that displays the pairwise correlation coefficients between various features of early prediction of ASD. The matrix includes rows and columns representing features

With correlation values ranging from -1.0 (strong negative correlation, shown in dark blue) to 1.0 (strong positive correlation, shown in dark red). The strong positive correlations include those among the A-scores ($A1\_Score\ and\ A2\_Score$ at 1.00), suggesting redundancy or similarity in these measures, while Class/ASD shows moderate positive correlations with result (0.35) and relation (0.34), indicating their potential relevance to ASD classification. Negative correlations are minimal, with the strongest being -0.20 between age and$country_of_res$, suggesting little linear relationship. This heatmap provides a visual insight of feature interdependencies, aiding in feature selection and model development for ASD prediction.

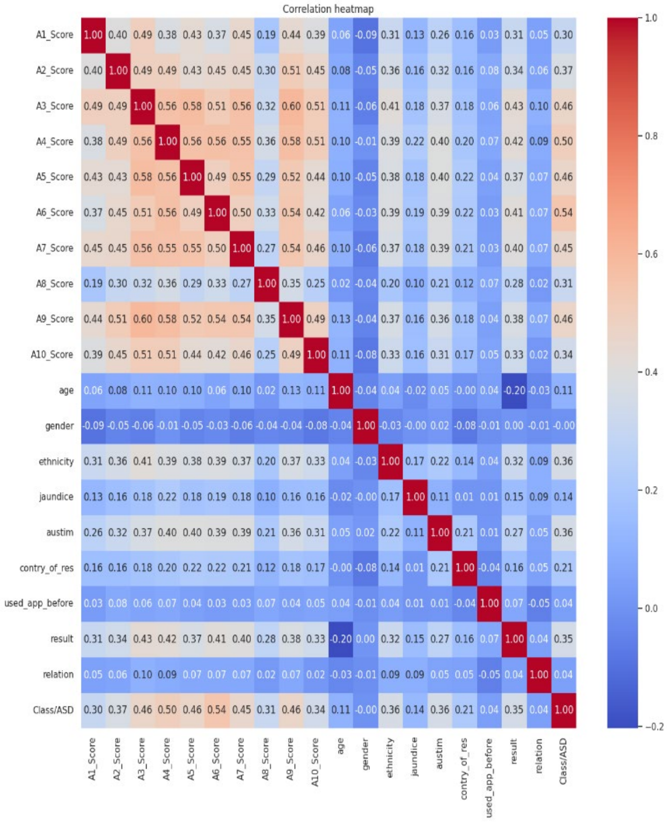

Figure 5: Correlation Heatmap for ASD Prediction Features

A. Data Reduction by Principal Components Analysis (PCA)

Figure 6 shows the principal components used for data reduction for early prediction of ASD. The $x-axis$ represents the principal components, ranging from 1.00 to 3.00. The $y-axis$ shows the eigenvalue, starting at 20.0 and decreasing to 2.5. A green line with data points marks the eigenvalues, with the first component at approximately 19.0, dropping sharply to around 5.0 at the second component, and further declining to about 2.5 at the third, indicating a steep initial drop followed by a gradual decrease. The plot includes a legend noting Eigenvalues ≥ 1, suggesting that only components with eigenvalues greater than or equal to 1 are considered significant, with the first two components capturing the majority of the variance

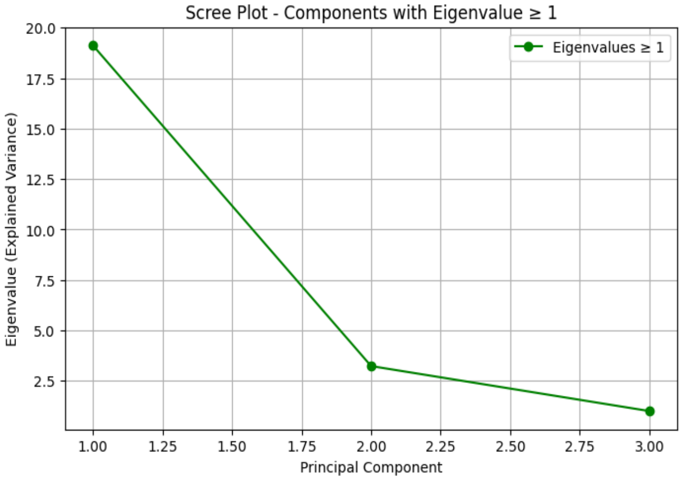

Figure 6: ASD Data Reduction with PCA

B. Models Evaluation

Table 2 presents the performance comparison of the RF and XGBoost classifiers for early detection of ASD. Both models demonstrate strong performance in detecting non-ASD individuals (Class 0). The RF model achieved a precision of 0.88, a recall of 0.94, and an F1-score of 0.91, while XGBoost recorded slightly higher precision (0.89) but lower recall (0.88) and the same F1-score (0.89). The RF achieved a precision of 0.65, a recall of 0.47, and an F1-score of 0.55, while XGBoost recorded lower precision (0.55) but slightly better than $recall$ (0.56) $and\ the\ same\ F1-score$ (0.55) . However, XGBoost outperformed RF with 82% accuracy against 81% accuracy of RF.

Table 2: RF and XGBoost Model Evaluation

Metric	Class	RF	XGBoost
Precision	0	0.88	0.89
	1	0.65	0.55
Recall	0	0.95	0.88
	1	0.55	0.56
F1-Score	0	0.91	0.89
	1	0.55	0.55
Support	0	128	128
	1	32	32
Accuracy	-	0.81	0.82
Macro Avg Precision	-	0.76	0.72
Macro Avg Recall	-	0.70	0.72
Weighted Avg Recall	-	0.84	0.82
Weighted Avg F1-Score	-	0.83	0.82

The confusion matrix in Figure 7 illustrates the performance of the RF model for ASD prediction. The model correctly identified 120 individuals without ASD (true negatives) while misclassifying 8 non-ASD individuals as ASD (false positives). For ASD cases, the model successfully predicted 15 individuals with ASD (true positives), but it failed to identify 17 actual ASD cases (false negatives). This indicates that the model is highly effective at ruling out non-ASD cases but demonstrates weaker performance in detecting ASD cases. The low number of false positives suggests that few non-ASD individuals are misclassified, improving sensitivity to correctly detect more ASD cases is essential for enhancing the model's practical usefulness in clinical and screening.

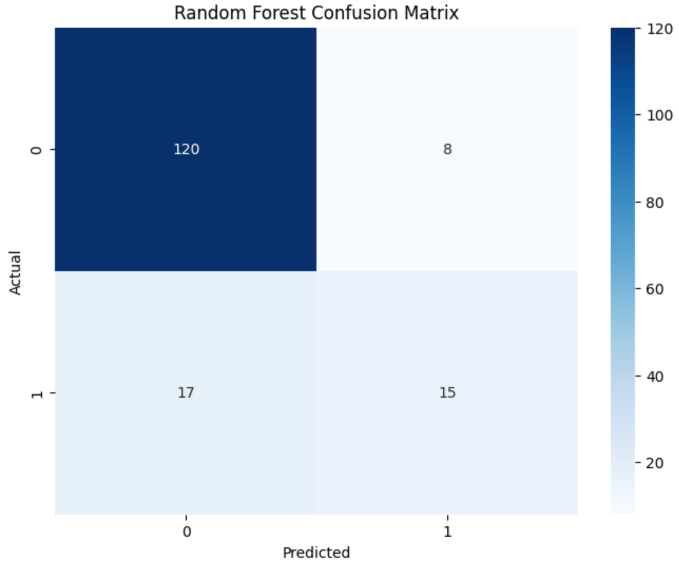

Figure 7: RF Confusion Matrix for ASD Prediction

Figure 8 is the confusion matrix illustrates the performance of the XGBoost model in the early detection of ASD in children. The results show that the model correctly identified 113 children without ASD (True Negatives) and 18 children with ASD (True Positives). However, it also misclassified 15 children without ASD as having ASD (False Positives) and failed to identify 14 children with ASD (False Negatives). Based on these results, the accuracy of the model is approximately 81.9%, indicating that the model correctly classified the majority of cases. The precision for ASD detection is 54.5%, meaning that when the model predicts ASD, it is correct slightly more than half of the time. The recall (sensitivity) for ASD detection is 56.3%, suggesting that the model detects just over half of the children who truly have ASD. On the other hand, the specificity is high at 88.3%, reflecting the model's strong ability to identify children without ASD. However, while the XGBoost model demonstrates good accuracy and strong specificity, its moderate sensitivity highlights the need for improvement in detecting actual ASD cases.

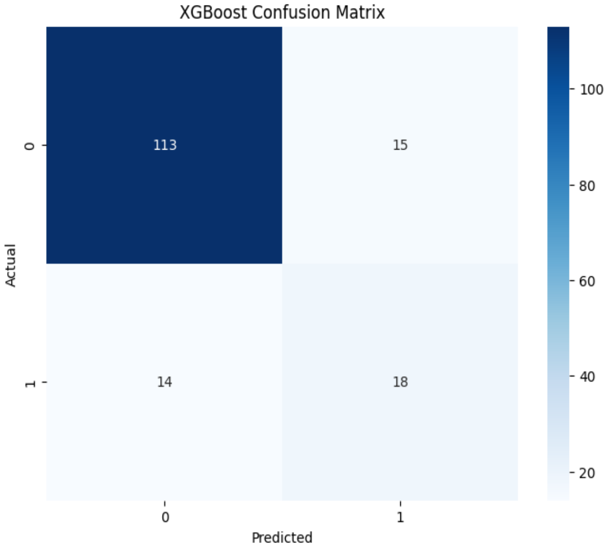

Figure 8: XGBoost Confusion Matrix for ASD Prediction

Figure 9 show ROC of the XGBoost model for the prediction of ASD in children. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 – Specificity) across different decision thresholds. The orange curve represents the model's performance, while the diagonal dashed line serves as a baseline indicating random guessing. The XGBoost model achieved an AUC value of 0.86, which demonstrates strong discriminatory ability between children with and without ASD. The AUC of 0.86 implies that the model has an 86% probability of correctly distinguishing a randomly selected child with ASD from one without ASD. The curve's position above the baseline shows that the model maintains high sensitivity while limiting false positives.

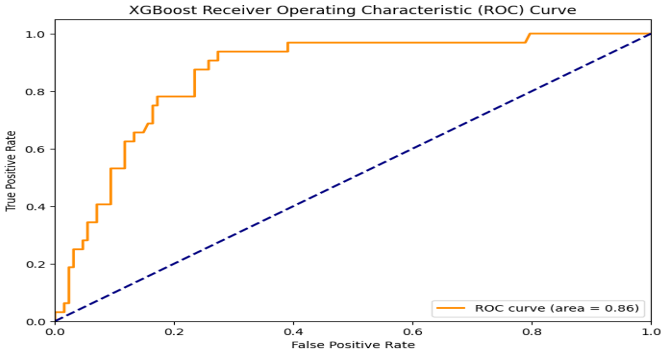

Figure 9: XGBoost ROC Curve for ASD Prediction

C. Hybridization of RF and XG Boost Models

Table 3 is the model's performance using classification metrics, including precision, recall, F1-score, and support, for both non-ASD (Class 0) and ASD (Class 1) categories. For Class 0 (non-ASD), the model demonstrates robust predictive capability. With a precision of 0.89, it correctly identifies 89% of individuals predicted to be non-ASD. The recall value of 0.91 indicates that 91% of the actual non-ASD cases were successfully classified. These strong precision and recall values yield a high F1-score of 0.90, based on 128 support cases. This shows that the model is highly reliable in identifying individuals who are not on the autism spectrum. However, for Class 1 (ASD), the model’s performance is relatively moderate. The precision of 0.59 shows that just under 60% of predicted ASD cases were correct, while the recall of 0.53 suggests that only 53% of true ASD cases were identified. This imbalance between precision and recall results in an F1-score of 0.56, with a smaller support size of 32. The reduced performance in detecting ASD-positive cases may be attributed to class imbalance, where the ASD group is underrepresented compared to the non-ASD group.

Further evaluation yields 0.84, signifying that the model capacity to predict both classes accurately. The macro average, which equally weighs both classes, records moderate results with a precision of 0.74, a recall of 0.72, and an F1-score of 0.73. Meanwhile, the weighted average, which accounts for the larger proportion of non-ASD cases, aligns with the overall accuracy, showing 0.83 across all metrics. This indicates that the hybrid RF-XGBoost model is highly effective at identifying non-ASD and ASD individuals

Table 3: RF-XGBoost Model Evaluation

$Class$	$Precision$	$Recall$	$F1 - Sc$	
0	0.89	0.91	0.90	128
1	0.59	0.53	0.56	32
Accuracy			0.84	160
Macro Avg	0.74	0.72	0.73	
Weighted Avg	0.83	0.83	0.83	

The confusion matrix in Figure 10 illustrates the classification performance of the hybrid RF-XGBoost model for ASD prediction. Out of 128 actual non-ASD cases (Class 0), the model correctly predicted 116 instances as non-ASD (true negatives), while it misclassified 12 cases as ASD (false positives). For ASD cases (Class 1), the model correctly identified 17 instances (true positives) but failed to detect 15 actual ASD cases, classifying them instead as non-ASD (false negatives). This outcome shows that the model is highly effective at identifying non-ASD individuals, with very few false positives. However, the performance in detecting ASD cases is moderate, as nearly half of the true ASD cases were misclassified. This indicates a limitation in sensitivity for ASD detection, which may be due to class imbalance where ASD instances are fewer compared to non-ASD cases. However, the model demonstrates strong accuracy for the majority class (non-ASD), but improvement is required in recognizing ASD-positive individuals to enhance early detection capability. This visualization provides insight into the model's accuracy and misclassification patterns for ASD prediction.

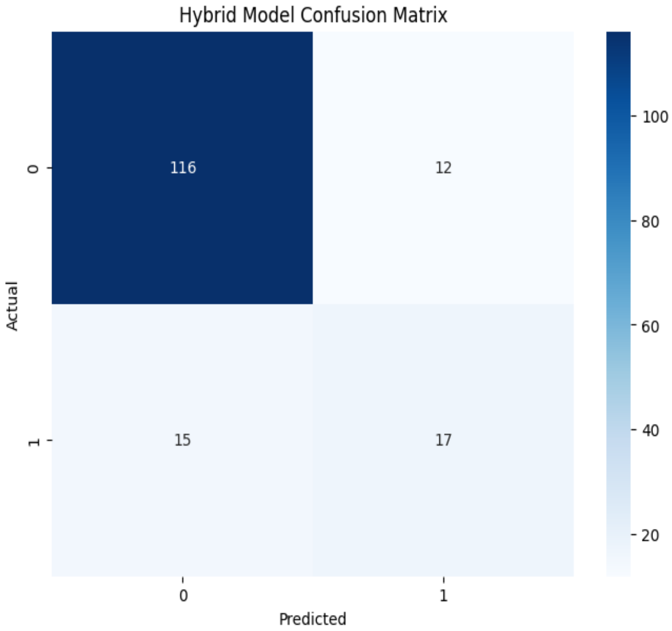

Figure 10: Hybrid Model Confusion Matrix for ASD Prediction

Figure 11 show ROC and AUC recorded at 0.86. This value indicates strong model performance, as it significantly exceeds the baseline score of 0.5, which represents random guessing. The graph effectively demonstrates the model’s ability to achieve a balanced trade-off between sensitivity and specificity in predicting ASD.

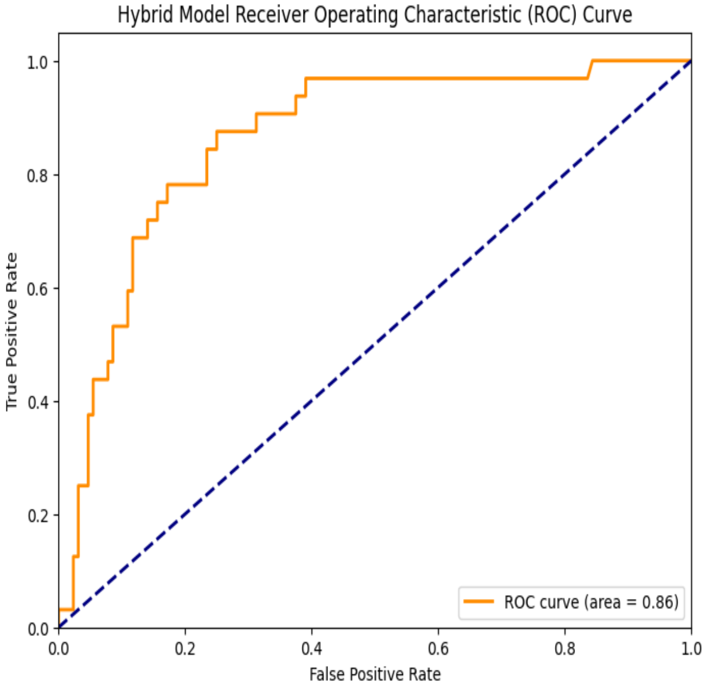

Figure 11: Ensemble Model ROC Curve for ASD Prediction

D. Experimentation Results with Existing Methods

Table 4 shows that deep learning (DNN) and big data-driven approaches (BDML-MDCASD) produce very high accuracy. However, demand for large-scale, high-quality datasets and high computational resources, which are often not available in real-world ASD screening, particularly in low-resource or clinical settings. Fuzzy logic and standard ML models (ML+FL) struggle with class imbalance and overlapping behavioral features, leading to reduced sensitivity in detecting ASD-positive cases. The proposed RF–XGBoost hybrid model offers a balanced trade-off between accuracy, sensitivity, and interpretability. The experimentation results show RF robustness against noise and XGBoost gradient-boosted optimization, the model handles moderate-sized, imbalanced datasets more effectively than deep learning approaches. It achieves a strong AUC (0.86), showing reliable discriminatory ability between ASD and non-ASD cases. Maintains computational efficiency, making it feasible for real-world clinical deployment. Provides interpretability of feature contributions (through feature importance scores), which supports clinical decision-making, something less transparent in DNN-based methods. The RF–XGBoost is the more practical and reliable choice in this study, due to its performance, such as interpretability, and applicability in ASD early detection.

Table 4: Model Comparison

Study	Method	Dataset	Accuracy	F1-Score	AUC / ROC	Remarks
Proposed work	RF + XGBoost	Kaggle	84%	0.73	0.86	Balanced sensitivity; robust to noisy, small-scale data
Boughattas & Jabnoun (2022)	DNN	ASD benchmark	99%	0.98	0.97	Very high accuracy but requires large, clean datasets
Kavadi et al. (2024)	BDML-MDCASD	Medical big data	92%	0.92	0.92	Effective on big data, computationally expensive
Qureshi et al. (2023)	ML + FL	Mixed datasets	80%	-	0.8	Suffers from class imbalance and limited generalizability

V CONCLUSION

This study successfully demonstrated that the RF-XGBoost hybrid model significantly improves the early prediction of ASD and non-ASD in children. By integrating behavioral, demographic, and clinical data, the model achieved reliable performance in distinguishing between ASD and non-ASD cases. The findings also highlighted a performance gap in detecting ASD-positive cases, primarily due to class imbalance and feature similarities across groups. However, the hybrid model outperformed individual classifiers in overall classification effectiveness, underscoring its potential for real-world screening applications. Unlike conventional models, the hybrid RF-XGBoost framework leverages the combined predictive strength of multiple algorithms to address common challenges such as overfitting, low sensitivity, and data imbalance. The inclusion of behavioral and demographic indicators within the model architecture provides a more holistic approach to ASD screening. Moreover, this study contributes a reproducible methodology and valuable insights that future researchers and practitioners can adapt to enhance predictive models not only for ASD but also for other neurodevelopmental disorders. Future research should focus on extending the hybrid machine learning framework by incorporating deep learning techniques and real-time behavioral data obtained from wearable or mobile devices. Studies could also investigate the model's performance across different age groups, cultural contexts, and clinical settings to improve its generalizability. Also, comparative analyses involving ASD and other neurodevelopmental disorders may provide a broader understanding of symptom overlaps and diagnostic boundaries.

ACKNOWLEDGEMENT

The authors are grateful to Professor Udoinyang G. Inyang, Direrctor Tetfund Center of Excellence of Computational Intelligence, University of Uyo, Uyo, Nigeria, for the assistance in the thorough supervision of the work.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION STATEMENT

Samuel Akpan Robinson[1]: Conceptualization, Methodology, Supervision, Project Administration. Raphael Henshaw Ekpo[2]: Writing - original draft, Writing. Ukeme Donatus Archibong[3]: Review, editing, and Visualization. Kingsley Joseph[3]: Validation, Investigation.

Declaration Statement: AI tools such as Grammarly and ChatGPT were used for sentence construction and checking of grammar.

References

M. E. McDonald and F. D. DiGennaro Reed, "Distinguishing science and pseudoscience in the assessment and treatment of autism spectrum disorder," in Assessment of Autism Spectrum Disorder, 2nd ed., 2018, pp. 415–441, ISBN 9781462533107.

C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "autism spectrum disorder," The Lancet, vol. 392, no. 10146, pp. 508–520, 2018, doi: 10.1016/S0140-6736(18)31129-2.

J. Ferina, M. Kruger, U. Kruger, D. Ryan, C. Anderson, J. Foster, and J. Hahn, "Predicting problematic behavior in autism spectrum disorder using medical history and environmental data," Journal of Personalized Medicine, vol. 13, no. 10, p. 1513, 2023, doi: 10.3390/jpm13101513.

Y. Huang, S. R. Arnold, K. R. Foley, and J. N. Trollor, "Diagnosis of autism in adulthood: A scoping review," Autism, vol. 24, no. 6, pp. 1311 – 1327, 2020, doi: 10.1177/1362361320903128.

S. Ramdoss, W. Machalicek, M. Rispoli, A. Mulloy, R. Lang, and M. O'Reilly, "Computer-based interventions to improve social and emotional skills in individuals with autism spectrum disorders: A systematic review, " Developmental Neurorehabilitation, vol. 15, no. 2, pp. 119−135, 2012, doi: 10.3109/17518423.2011.651655.

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

S. A. Robinson, A. A. Imianvan, E. C. Igodan, E. A. Dan, K. U. Joseph, and L. A. Dickson, "Machine learning approach for path loss prediction in urban drive 5G network environments," International Journal of Microwave & Optical Technology, vol. 20, no. 4, 2025.

S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques, " Procedia Computer Science, vol. 167, pp. 994–1004, 2020, doi: 10.1016/j.procs.2020.03.399.

S. A. Robinson, A. E. Udoh, E. A. Dan, P. U. Ejodamen, K. U. Joseph, and D. G. Asuquo, "Early depression prediction among Nigerian university students using adaptive neuro-fuzzy inference system (ANFIS), " Journal of Advances in Mathematics and Computer Science, vol. 39, no. 2, pp. 1–10, 2024.

J. Rogala, J. Żygierewicz, U. Malinowska, H. Cygan, E. Stawicka, A. Kobus, and B. Vanrumste, "Enhancing autism spectrum disorder classification in children through the integration of traditional statistics and classical machine learning techniques in EEG analysis," Scientific Reports, vol. 13, no. 1, p. 21748, 2023, doi: 10.1038/s41598-023-49048-7.

Imianvan, A. A., Robinson, S. A., Asuquo, D. E., George, U. D., Dan, E. A., Ejodamen, P. U., & Udoh, A. E. (2024). Enhancing Job Recruitment Prediction through Supervised Learning and Structured Intelligent System: A Data Analytics Approach. *Journal of Advances in Mathematics and Computer Science*, *39*(2), 72-88.

L. Damianos, C. Vlachas, K. F. Kollias, N. Asimopoulos, and G. F. Fragulis, "Machine learning methods for autism spectrum disorder classification," in AIP Conference Proceedings, vol. 2909, no. 1, AIP Publishing, Nov. 2023, doi: 10.1063/5.0182539.

N. Boughattas and H. Jabnoun, "Autism spectrum disorder (ASD) detection using machine learning algorithms," in Int. Conf. Smart Homes and Health Telematics, Cham: Springer, 2022, pp. 225–233, doi: 10.1007/978-3-031-09593-1_18.

M. Shekarro, S. Hassanzadeh, and R. Kellems, "Identification of autism spectrum disorder by parents: a retrospective-comparative study of the role of early behavioral signs, developmental and demographic characteristics, " Current Psychology, vol. 43, pp. 2403 – 2424, 2024, doi: 10.1007/s12144-023-04458-8.

M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, "Diagnosis-based hybridization of multimedical tests and sociodemographic characteristics of autism spectrum disorder using artificial intelligence and machine learning techniques: A systematic review," International Journal of Telemedicine and Applications, vol. 2022, no. 1, p. 3551528, 2022.

S. Alam, P. Raja, and Y. Gulzar, "Investigation of machine learning methods for early prediction of neurodevelopmental disorders in children," Wireless Communications and Mobile Computing, vol. 2022, no. 1, p. 5766386, 2022.

U. G. Inyang, S. A. Robinson, F. F. Ijebu, I. J. Udo, and C. O. Nwokoro, "Optimality assessments of classifiers on single and multi-labelled obstetrics outcome classification problems, " International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, 2021.

A. Imianvan and S. Robinson, "Optimizing 5G IoT connectivity for path loss reduction through ant colony optimization and support vector regression," in 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), IEEE, 2024, pp. 1–10.

A. Imianvan, S. A. Robinson, D. E. Asuquo, U. D. George, E. A. Dan, P. U. Ejodamen, and A. E. Udoh, " Enhancing job recruitment prediction through supervised learning and structured intelligent system: A data analytics approach," Journal of Advances in Mathematics and Computer Science, vol. 39, no. 2, pp. 72–88, 2024.

Shrivastava, A. Kotiyal, M. I. Habelalmateen, A. Rana, V. A. Devi, B. D. Rao, and S. Bansal, "Leveraging XGBoost for predictive analytics in healthcare: enhancing disease diagnosis," in 7th Int. Conf. Contemporary Computing and Informatics (IC3I), IEEE, Sept. 2024, vol. 7, pp. 1666–1672.

S. A. Robinson, A. A. Imianvan, E. C. Igodan, E. A. Dan, K. U. Joseph, and L. A. Dickson, "Machine learning approach for path loss prediction in urban drive 5G network environments," International Journal of Microwave & Optical Technology, vol. 20, no. 4, 2025.

U. G. Inyang, I. J. Eyoh, S. A. Robinson, and E. N. Udo, "Visual association analytics approach to predictive modelling of students' academic performance," International Journal of Modern Education & Computer Science, vol. 11, no. 12, pp. 1–13, 2019.