2025 28th International Conference on Computer and Information Technology (ICCIT)
19-21 December 2025, Cox's Bazar, Bangladesh

Improved Autism Identification Using Machine Learning Models : Emphasized Behavioral Analysis

1st Mahbub Akanda
Department of Computer Scicence and Engineering
Jamalpur Science and Technology University
Jamalpur-2012, Bangladesh
mahbubsakib13@gmail.com

2nd Md. Sydur Rahman
Department of Computer Scicence and Engineering
Jamalpur Science and Technology University
Jamalpur-2012, Bangladesh
sydur.cse@jstu.ac.bd

***Abstract*—Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects communication, behavior, and social interaction. Early and accurate identification is essential for timely intervention, yet traditional diagnostic procedures are often costly, time-consuming, and dependent on clinical expertise making them less accessible in resource-constrained settings. This study introduces a behavior-driven, machine learning based screening framework that utilizes publicly available Q-CHAT-10 behavioral datasets to identify ASD in both children and adults. After extensive preprocessing, including class balancing with SMOTE, feature analysis, and dimensionality reduction, seven machine learning models were evaluated: LightGBM, Random Forest, Support Vector Machine, Decision Tree, KNN, Logistic Regression, and Naive Bayes. Performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. LightGBM achieved the highest performance with 99.59% accuracy and a near-perfect AUC-ROC of 0.9999, followed by Random Forest with 98.93% accuracy. The results demonstrate that ensemble-based approaches significantly outperform traditional classifiers, offering robust and scalable alternatives for early ASD screening. This study highlights the potential of machine learning to support clinical decision-making, reduce screening time and cost, and improve access to ASD identification particularly in rural and underserved regions.**

***Index Terms*—Autism detection, Machine learning, KNN, Logistic Regression, Random Forest, LightGBM, Naive Bayes, Decision Tree, SVM, Behavior Analysis.**

I. Introduction

Autism spectrum disorder (ASD) involves repetitive and/or restricted interests and stereotyped behaviors as well as enduring deficits in social interaction. Recent worldwide evidence suggests that the rates of ASD are still rising, highlighting a critical need for early and accurate diagnosis in order to prompt timely provision of targeted interventions associated with improved developmental outcomes [1]. Conventional methods of diagnosing ASD, including clinical interviews and the Autism Diagnostic Observation Schedule (ADOS), are subjective, time-consuming and depend on expert decision making [2]. Accessibility of these conventional methods is generally an issue, more so in rural or resource-poor communities. Consequently, there has been an increasing interest in the use of quantitative and computational tools to aid in the process of screening and diagnosis for ASD [3]. We compare a number of supervised machine learning (ML) algorithms for ASD detection in this study including LightGBM, Random Forest, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, and Naive Bayes. This study builds on what we already know about digital health. It shows how machine learning can make it easier and better to find Autism Spectrum Disorder early on. These computer models can spot initial signs of ASD and support standard ways of finding it, mainly in places without many resources.

The main contributions of this research are:

- Early Detection Support: Created a machine learning framework that can accurately identify autism at an early stage, allowing for timely interventions.
- Affordable and scalable: Proposed models that can be used in mobile or low-resource healthcare settings to lower the cost and time it takes to make a diagnosis.
- Clinical Relevance: Enhanced the incorporation of machine learning into neurodevelopmental disorder screening, facilitating more prompt, dependable, and equitable autism diagnosis.
- High Accuracy Models: Showed that ensemble approaches like LightGBM and Random Forest can find ASD with almost clinical accuracy.
- Practical Alternatives: Tested simpler models (Logistic Regression, SVM) for cases when speed, ease of understanding, and low computer capacity are critical.

II. LITERATURE REVIEW

A related research was conducted in Bangladesh utilizing the Autism Barta App and field questionnaires in Savar to identify behavioral and demographic characteristics that influence the risk of ASD, such as gender and birth time. An analysis of over 600 data revealed regional disparities in autism prevalence, which assisted in the early diagnosis of ASD [4]. Convolutional Neural Networks (CNNs) were used for transfer learning, using the VGG-19 architecture. The model's 84.67% accuracy rate demonstrated how effective pre-trained networks are for image-based ASD identification [5]. In order to improve KNN-based classification, another study used the Binary Firefly Algorithm for feature selection. Following noise reduction, the model's accuracy increased from 87.67% to 93.84%, with similar improvements seen for Naive Bayes and Decision Tree classifiers. [6]

979-8-3315-7867-1/25/$31.00

In order to identify autism through facial features, a deep learning method in [7] used the MobileNet, Xception, and InceptionV3 architectures. Although researchers pointed out that more intricate architectures could further enhance performance, MobileNet demonstrated superior accuracy.
In a different study, the Xception model was also tested against VGG16, and while it achieved 91% accuracy compared to 78%, it had higher computational costs and the potential to overfit [8]. A research on age-specific models in behavioral-based classification employed scaling approaches (normalization, quantile transform) to assess classifiers like AdaBoost (98.6% accuracy for toddlers) and LDA (98.08% for adults) [9]. Another survey-based research found that KNN surpassed SVM and Naive Bayes, with an accuracy rate of 98% [10]. Further research focused on classifiers for integrated Kaggle and UCI datasets. Despite its limitations owing to a lack of feature interaction modeling, logistic regression had the greatest average accuracy (98.59%) [11]. SVM and Logistic Regression both achieved 71% accuracy in dialogue data analysis using machine learning, indicating a moderate level of success. However, biases were introduced by data from online sources [12].

III. Dataset Description

The Q-Chat-10 screening questionnaire served as the basis for the dataset used in this study, which was gathered from publicly accessible sources, mostly Kaggle. Combining binary answers to 10 behavioral questions (A1–A10), demographic information (age, sex), and medical history (birth jaundice, family history of ASD), it comprises 6,075 samples and 15 attributes. Whether a subject has received an ASD diagnosis (YES) or not (NO) is indicated by the target variable "Class" which is presented in Table I. The dataset is ideal for supervised machine learning algorithms due to its thorough representation of behavioral features.

TABLE I
Dataset Attributes for Autism Screening

Column	Description	Type	Values
A1–A10	Autism screening question responses	Integer	0, 1
Age	Participant age	Integer	4–76
Sex	Gender	Categorical	m, f
Jaundice	Jaundice at birth	Categorical	yes, no
Family_ASD	Family history of ASD	Categorical	yes, no
Class	ASD diagnosis result	Categorical	YES, NO

IV. Methodology

This research presents a machine learning (ML) approach for diagnosing autism spectrum disorder (ASD) correctly and early using behavioral data. The methodology's main phases include preprocessing, dimensionality reduction, model training, evaluation, and dataset selection. Fig. 1 in the article shows a schematic overview of the approach.

A. Data Pre-processing

The first step in getting raw data ready for machine learning is to preprocess it. It means cleaning the data by dealing

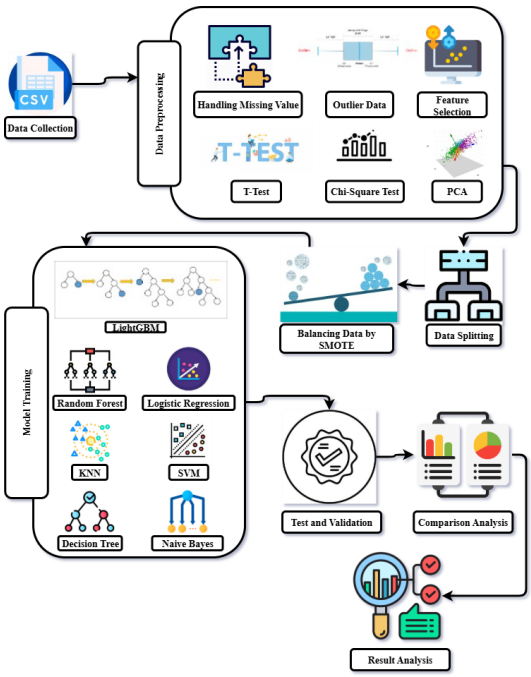

Fig. 1. Step by Step Workflow of Our Research.

with missing values and outliers, changing categorical text into numbers, and scaling characteristics so that they all fall within the same range. To make sure the model doesn't favor one class over another, techniques like SMOTE are employed to balance datasets that aren't balanced. PCA and other dimensionality reduction approaches can make the data easier to work with by merging related features. This whole process takes messy, real-world data and turns it into a clean, structured format. This makes a strong and reliable base for making accurate and useful prediction models. The process can be broken down into the following steps:

1) Handling Missing Values: It helps to detect missing values in each column by counting the number of null entries as the samples are represented in Table II. This allows us to identify quickly which parts of the dataset are incomplete and need further cleaning. Recognizing these gaps early is important because missing data can negatively impact the accuracy and reliability of any analysis or model built on the dataset. By addressing these issues during preprocessing, we ensure the data remains consistent, complete, and ready for the next stages of research.

TABLE II
Missing Values Summary

Column Name	Missing Values
A1 – A10, Age	0
Sex	0
Jauundice	0
Family_ASD	0
Class	0

2) Correlation Matrix and Feature Importance: The correlation matrix is a strong tool for understanding the relationship between different features present in our dataset.The features are listed along both the x-axis and y-axis, including A1 through A10, as well as Age, Sex, Jaundice, Family_ASD, Class, and Cluster. Each cell in the lower triangular part of the matrix shows the Pearson correlation coefficient between two corresponding features, both numerically and through color intensity. Positive correlations are shown in shades of red, negative correlations in shades of blue, and weak or no correlations are represented with lighter, almost white colors. A color bar on the right side of the Fig. 2 indicates the strength of the correlation, ranging from -1.0 (strong negative correlation) to 1.0 (strong positive correlation).

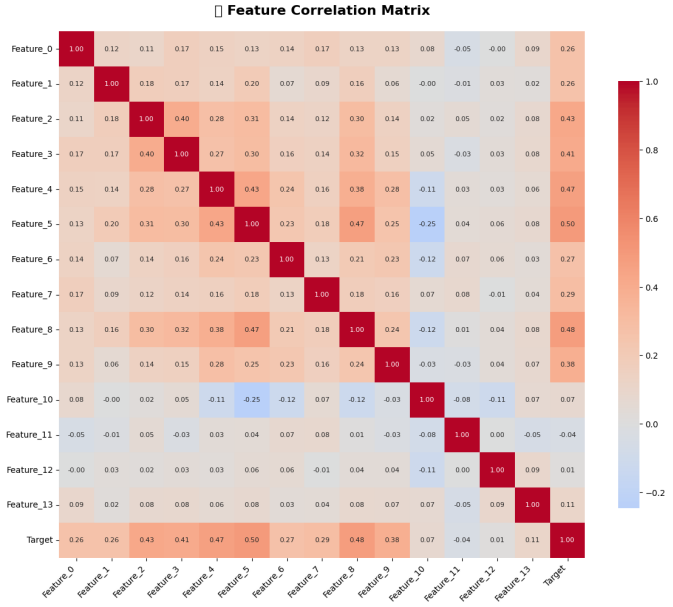

Fig. 2. Correlation Matrix

B. Features Analysis

The Random Forest and the ANOVA F-score were two strong tools utilized in this research to find the best predictive features. The Random Forest looks at how useful a feature is in complicated relationships, while the ANOVA F-score looks at how strong a feature is on its own. There is a substantial agreement among them that "Age," "A6," and "A9" are the best predictors. This is solid, multi-faceted proof that these are the most dependable indications for an accurate assessment, cutting through the noise to show the essential signals in the data.

1) T-test and Chi-Square Test: The statistical tests showed clear links between certain patient traits and the ASD categorization. A T-test validated that age is a major factor, revealing a considerable disparity in the average age between the diagnosed and non-diagnosed cohorts. Moreover, Chi-square tests revealed that an individual's sex and the presence of a familial history of ASD are significantly correlated with the classification outcome. In contrast, a history of jaundice was determined to have no statistically significant correlation with the diagnosis, indicating that it is a less pertinent component in this specific predictive model.

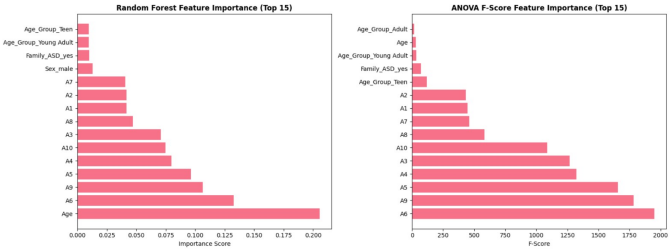

Fig. 3. Random Forest and ANOVA F-Score Feature Importance.

TABLE III
COMPARISON BETWEEN T- TEST AND CHI-SQUARE TEST

Test Type	Feature vs. Class	Test Statistics	P-value	Effect Size	Statistical Significance	Practical Significance
T-test	Age vs. Class	t = 5.80	0.0000	Cohen's d = 0.159	Highly Significant	Very Small
Chi-square	Family$_ASDvs.Class$	² = 69.63	0.0000	Cramer's V = 0.107	Highly Significant	Small
Chi-square	Sex vs. Class	² = 8.48	0.0036	Cramer's V = 0.037	Significant	Negligible
Chi-square	Jaundice vs. Class	² = 2.72	0.0989	Cramer's V = 0.021	Not Significant	Negligible

2) Dimentionality Reduction and Principal Component Analysis: Dimensionality reduction makes complex datasets easier to work with by turning many original features into a smaller collection of new, composite features known as primary components. This approach helps with the "curse of dimensionality," which happens when there are too many characteristics that can slow down models, make them more likely to overfit, and hide important patterns. Standardizing

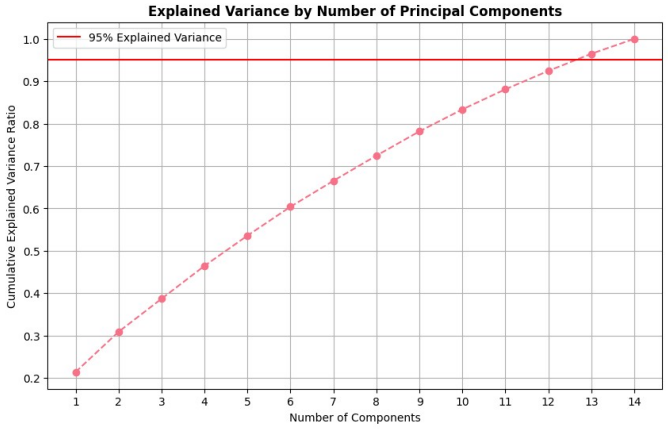

Fig. 4. Variance by Number of Principal Components

the data is the first step in the analysis. This is important to make sure that the size of the features doesn't affect the results. Then, Principal Component Analysis (PCA) is used to find new, orthogonal axes that capture the most variance in the data. The scree plot shows that the relevance of each component drops off quickly, with the initial few holding most of the information. A major discovery is that 13 principle components are enough to capture 95% of the original data's variability like in Fig. 4. This means that the 14 original

features can be combined into 13 components with very little loss of information.

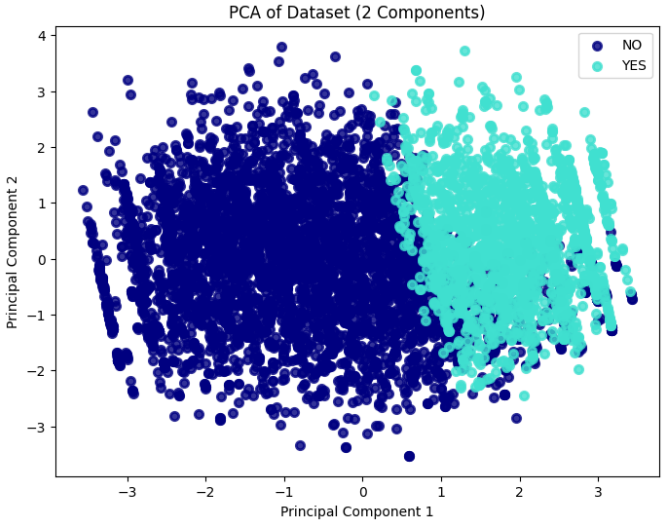

Fig. 5. PCA of Dataset.

C. Handling Class Imbalance

The first dataset had a big class imbalance, with 6,075 total samples and around 2.4 times as many ”NO” (no ASD) cases as ”YES” (ASD) cases. This difference could have led to training models that were biased toward the majority class and did a bad job of finding the important minority class. To fix this, the SMOTE method was used, which smartly makes fake examples for the minority class. The intervention worked quite well, changing the dataset from its original unbalanced state to a fully balanced set of 8,542 samples. The outcome was a 1:1 balance between ”NO” and ”YES” cases, which made the model training more fair and improved its capacity to find patterns for both outcomes which is demonstrated in Fig. 6.

D. Classification

In supervised machine learning, a dataset is split into training and testing sets to evaluate a classifier. The training set teaches the model to learn patterns, while the testing set

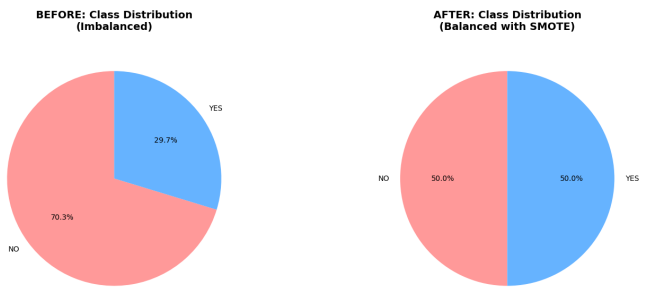

Fig. 6. Imbalance and Balance Class Distribution.

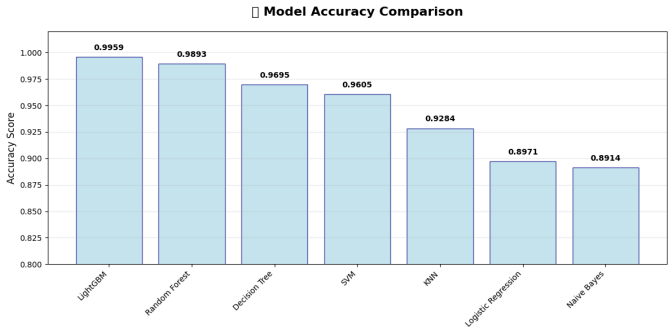

Fig. 7. Accuracy Comparison.

not seen during training checks how well it generalizes [13]. We have splitted our dataset 70% training and 30% testing, which helps us preventing overfitting and ensures reliable performance on new data. After that we have applied seven classifiers.

V. Evaluation and Results

In order to evaluate the efficacy of the machine learning models utilized for the detection of Autism Spectrum Disorder (ASD), we employed a number of common evaluation metrics that are frequently employed in classification problems. A thorough understanding of each model’s ability to differentiate between cases with and without autism is provided by these metrics. The following statistical performance metrics were obtained.

- **Sensitivity (Recall):** The ability of the classifier to correctly identify positive instances.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

- **Specificity:** The ability of the classifier to correctly identify negative instances.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

- **Accuracy:** The overall proportion of correctly classified instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In terms of accuracy, LightGBM had the best predicted accuracy at 99.59%, which was far better than all the other models. Then came Random Forest (98.93%), Decision Tree (96.95%), and SVM (96.05%). The other models KNN (92.84%), Logistic Regression (89.71%), and Naive Bayes (89.14%) were less and less accurate visible in Fig. 7.

The model evaluation shows a clear ranking of performance, with LightGBM showing outstanding competence on all metrics. It had the highest precision at 99.17%, which means that its positive predictions were quite reliable. It also had a recall rate of 99.45%, which means that it could find almost all of the relevant cases. It’s top F1 score of 99.31% shows that this balance is correct. Random Forest came in

a close second with 98.60% precision, 97.78% recall, and a 98.19% F1 score. Decision Tree and SVM made up a strong middle tier, while KNN, Logistic Regression, and Naive Bayes made up the lower-performing category, with their scores for precision, recall, and F1 all being less than 92%, 92%, and 89%, respectively which has been presented in Fig. 8.

LightGBM wouldn't just be winning if this were a race; it

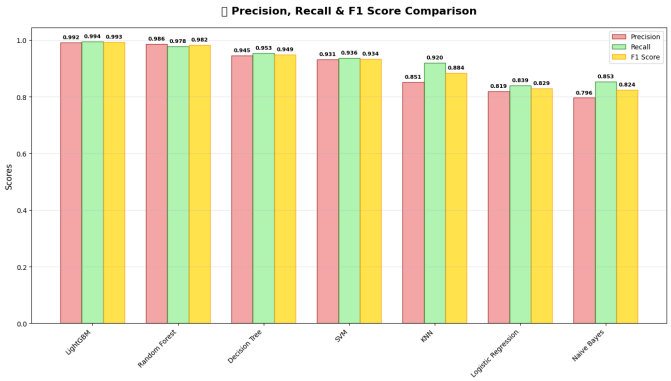

Fig. 8. Precision, Recall and F1-Score Comparison.

would be lapping the other cars. The radar chart makes this clear by showing that LightGBM's performance makes a big, dominant shape that reaches the outer edges of every metric. Next up is Random Forest, which is a strong and reliable competitor. Its shape is noticeably smaller but still strong. The Decision Tree, on the other hand, is still a solid third place, and its smaller size on the chart makes it clear how far it needs to go to catch up. The figure in Fig. 9 makes it clear right away which is doing the best compared to others.

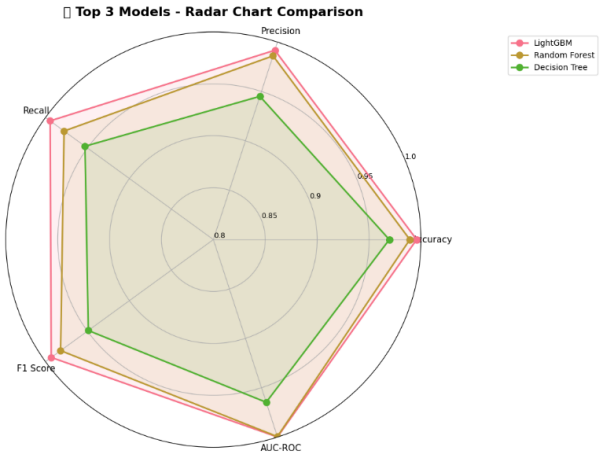

Fig. 9. Top 3 Models : Radar Chart Comparison.

To see how effectively each classifier found autism spectrum disorder (ASD), I employed a lot of performance indicators, such as Accuracy, Precision, Recall, F1 Score, and AUC-ROC. Table IV shows these metrics, which provide information on the diagnostic capacity of each model that is both generic and unique to each class.

TABLE IV
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
LightGBM	**0.9959**	**0.9917**	**0.9945**	**0.9931**	**0.9999**
Random Forest	0.9893	0.9860	0.9778	0.9819	0.9995
SVM	0.9605	0.9311	0.9363	0.9337	0.9922
Decision Tree	0.9695	0.9451	0.9529	0.9490	0.9647
KNN	0.9284	0.8513	0.9197	0.8842	0.9765
Logistic Regression	0.8971	0.8189	0.8393	0.8290	0.9575
Naive Bayes	0.8914	0.7959	0.8532	0.8235	0.9460

The AUC-ROC comparison shows the same thing as the numbers do. LightGBM has the best AUC-ROC score of 0.9999, which means it can almost flawlessly tell the difference between classes. Random Forest comes in second with 0.9995, which also shows great separation. Figure 10 shows that the AUC values for KNN (0.9765), Logistic Regression (0.9575), Decision Tree (0.9647), and Naive Bayes (0.9460) are all considerably over the "Excellent" level of 0.9. With an AUC of 0.9922, SVM also works quite well. Overall, LightGBM and Random Forest are the top models, while all evaluated models demonstrate good classification abilities and reliable predictions.

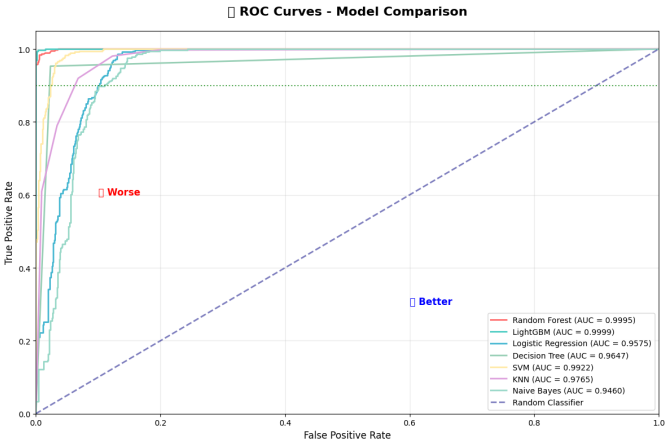

Fig. 10. ROC Curves Comparison.

VI. K-FOLD CROSS VALIDATION

On seven classification models, we have used a five-fold cross-validation technique. In that instance, we utilized 20% of the data for validation and 80% of the data for training. The steps are carried out five times, as shown in Fig. 11

A. LightGBM Performance Matrix

From the stated Table V LightGBM outperforms all other models with nearly perfect scores across all metrics. The standard deviation is also very low, indicating exceptional consistency.

VII. CONCLUSION

This research presents a machine learning methodology that aims to predict autism spectrum disorder(ASD) through the

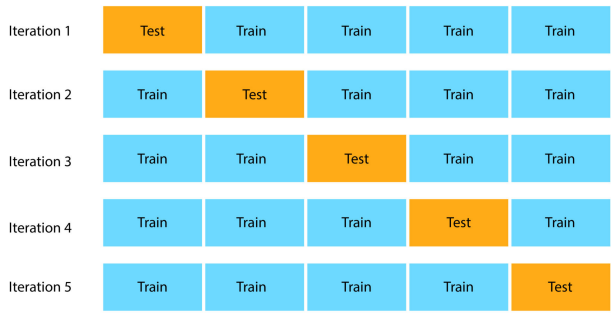

Fig. 11. Illustration of 5-Fold Cross-Validation. In each iteration, one fold is used as the test set (orange), and the remaining folds are used for training (blue).

TABLE V
LIGHTGBM - FOLD-WISE PERFORMANCE

	Accuracy	Precision	Recall	F1	AUC-ROC
Fold 1	0.995	0.997	0.986	0.992	1.000
Fold 2	0.991	0.977	0.995	0.986	1.000
Fold 3	0.992	0.986	0.986	0.986	1.000
Fold 4	0.995	0.991	0.991	0.991	1.000
Fold 5	0.991	0.985	0.982	0.984	0.999
Mean	0.993	0.987	0.988	0.988	1.000
Std Deviation	0.002	0.004	0.006	0.005	0.000

analysis of surveyed data related to behavioral and developmental characteristics. Following pre-processing steps such as encoding, addressing missing values, feature analysis by T-test and Chi-Square Test, feature selection is performed using PCA, seven classifiers namely LightGBM, Random Forest, SVM, Decision Tree, KNN, Logistic Regression, and Naive Bayes are applied. Using a stratified train–test splitting, both LightGBM and Random Forest demonstrated the highest accuracy along with nearly perfect ROC-AUC scores, while the other models exhibited commendable performance but with marginally elevated mis-classification rates. Comparative analysis revealed that ensemble models offer the most reliable predictions, functioning as a data-driven resource for the early detection of ASD to assist clinicians and carers.

VIII. FUTURE RESEARCH DIRECTIVES

Although this study illustrates the potential of machine learning in predicting Autism Spectrum Disorder (ASD), numerous avenues for future research exist to enhance robustness, scalability, and practical applicability. Expanding the data set to include larger and more diverse populations in different regions would improve the generalization and accuracy of the model. Regional-specific analysis could reveal cultural or environmental factors, supporting localized ASD prediction models, while studying ASD traits in various professions might expose occupational patterns that improve classification accuracy. Further, a deeper comparison of advanced machine learning and deep learning techniques—such as XGBoost, convolutional neural networks (CNNs), or hybrid models—may deliver additional performance gains. Incorporating longitudinal or time-series data could enable tracking of the progression of ASD traits, allowing earlier detection and more personalized interventions. Finally, improving the explainability of the model and integrating these methods with clinical tools would help build clinician trust and support adoption in healthcare settings.

REFERENCES

[1] F. Thabtah, "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward," *Informatics for Health and Social Care*, vol. 44, no. 3, pp. 278–297, 2019.

[2] S. K. Tiwari, S. A. Dey, and S. S. Tiwari, "Autism spectrum disorder detection using machine learning techniques: A review," *Journal of Physics: Conference Series*, vol. 1950, 2021, doi: 10.1088/1742-6596/1950/1/012030.

[3] D. Bone, M. P. Goodwin, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of Autism and Developmental Disorders*, vol. 45, no. 5, pp. 1121–1136, 2015, doi: 10.1007/s10803-014-2268-6.

[4] M. S. Satu, F. F. Sathi, M. S. Arifen, M. H. Ali, and M. A. Moni, "Early detection of autism by extracting features: A case study in Bangladesh," in *Proc. 2019 Int. Conf. Robotics, Electrical and Signal Processing Techniques (ICREST)*, IEEE, 2019, pp. 400–405.

[5] S. Jahanara and S. Padmanabhan, "Detecting autism from facial image," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 7, no. 2, pp. 219–225, 2021.

[6] R. Vaishali and R. Sasikala, "A machine learning based approach to classify autism with optimum behaviour sets," *International Journal of Engineering & Technology*, vol. 7, no. 4, p. 18, 2018.

[7] Z. A. Ahmed, T. H. Aldhyani, M. E. Jadhav, M. Y. Alzahrani, M. E. Alzahrani, M. M. Althobaiti, F. Alassery, A. Alshaflut, N. M. Alzahrani, and A. M. Al-Madani, "[Retracted] Facial features detection system to identify children with autism spectrum disorder: Deep learning models," *Computational and Mathematical Methods in Medicine*, vol. 2022, no. 1, p. 3941049, 2022.

[8] A. Rashid and S. Shaker, "Autism spectrum disorder detection using face features based on deep neural network," *Wasit Journal of Computer and Mathematics Sciences*, vol. 2, no. 1, pp. 74–83, 2023.

[9] S. M. Hasan, M. P. Uddin, M. Al Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," *IEEE Access*, vol. 11, pp. 15038–15057, 2022.

[10] S. Islam, T. Akter, S. Zakir, S. Sabreen, and M. I. Hossain, "Autism spectrum disorder detection in toddlers for early diagnosis using machine learning," in *Proc. 2020 IEEE Asia-Pacific Conf. Computer Science and Data Engineering (CSDE)*, IEEE, 2020, pp. 1–6.

[11] T. Akter, M. I. Khan, M. H. Ali, M. S. Satu, M. J. Uddin, and M. A. Moni, "Improved machine learning based classification model for early autism detection," in *Proc. 2021 2nd Int. Conf. Robotics, Electrical and Signal Processing Techniques (ICREST)*, IEEE, 2021, pp. 742–747.

[12] P. Mukherjee, S. Sadhukhan, M. Godse, and B. Chakraborty, "Early detection of autism spectrum disorder (ASD) using traditional machine learning models," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[13] M. S. Rahman and B. Ahmed, "Predicting psychiatric disorder from the classified psychiatric characteristics using machine learning algorithm," *European Journal of Information Technologies and Computer Science*, vol. 2, no. 4, pp. 5–10, 2022.