Unlocking Autism DNA

Chalapati Sowmya
Department of Computer Science and Engineering
Vignan's Foundation for Science,Technology and Research
Guntur, India
221fa04074@gmail.com

Maridu Bhargavi
Department of Computer Science and Engineering
Vignan's Foundation for Science,Technology and Research
Guntur, India
bhargaviformal@gmail.com

Sikhinam Mercy
Department of Computer Science and Engineering
Vignan's Foundation for Science,Technology and Research
Guntur, India
221fa040113@gmail.com

Koduru Jhansi Suvarchala
Department of Computer Science and Engineering
Vignan's Foundation for Science,Technology and Research
Guntur, India
221fa04140@gmail.com

Shaik Mahmooda Aafreen
Department of Computer Science and Engineering
Vignan's Foundation for Science,Technology and Research
Guntur, India
221fa04143@gmail.com

***Abstract*—Autism Disorder is a neuro development disorder diagnosed through various behavioral conditions and, hence, its early diagnosis facilitates proper care, and because traditional clinical methods are time-consuming, instances are prone to error, so machine learning models like XGBoost, SVM, Logistic Regression, and AdaBoost can be used for earlier detection. In our study, models such as Logistic Regression, AdaBoost, and ensemble methods like XGBoost and Random Forest achieve excellent accuracy in some cases, but the main drawback of previous research work was their lack of security. Mitigating that in our proposed study federated learning is better due to the local processing mechanism, which safeguards the patient data while still further improving the model itself collaboratively.**

Index Terms— Machine Learning,Early Diagnosis,Logistic Regression,Autism Spectrum Disorder,XGBoost,AdaBoost,Predictive Modeling

I. Introduction

Autism Spectrum Disorder is defined as complex neuro developmental condition where individual expresses a considerable level of problems in his social interaction, communication, and behavior. Because outcomes significantly improve if interventions occur when a child is younger, current assessment methods are largely insufficient as they heavily rely on subjective behavioral observation and usually take so long to be completed. The age at which ASD is diagnosed in children is usually two years; however, sometimes it manifests much later, which complicates diagnosis and the availability of access to very basic support services.

Since the causes of ASD are now considered as multifactorial and involve both genetic and environmental factors, the disorder naturally differs in varieties of severities of symptoms.The chances of an accurate diagnosis would lie in a proper assessment of a patient by healthcare professionals, but limitations in traditional methods created interest to use novel approaches that employ technologies for the advancement of detection abilities of ASD.

One of the promising avenues to ASD early prediction is machine learning, which allows easy analysis of complex datasets with a much higher degree of precision than traditional methods. It has facilitated the development of many ML classification models for early identification in the hope of reducing the long-term effects of the disorder on the afflicted. However, these techniques are severely hampered by some significant challenges dealing directly with data privacy and security.These machine learning algorithms and techniques would depend on getting information related to health care.

We therefore advance a model based on Federated Learning: decentralized machine learning, which is an approach to putting data privacy and security. In FL, training of the model happens on individual devices; hence only updates go to the central server, thus eliminating all the issues and problems that come with raw data for an individual to share. This innovative methodology has a great promise in various medical applications, like the detection of neurological disorders, and can revolutionize the detection of ASD, as the whole methodology facilitates collaborative learning while not compromising on confidentiality about a patient.

Our study aims to utilize the FL technology in the diagnosis of ASD among both children and adults. We will attempt to draw on the findings of previous works regarding the criteria for ASD. We shall utilize the distinct models of ML and datasets of ASD Therefore, the central objective is oriented toward training a FL-based metaclassifier that aggregates multiple local machine learning models towards effective and identification of risk factors for ASD, overcoming the drawbacks of traditional ML approaches, while achieving performance comparable to prevailing detection strategies.

II. Literature Survey

N BalaKrishna et al.[1] used SVM, Decision Trees, and Logistic Regression for ASD detection, with SVM achieving the highest accuracy at 93%. The study highlights the need for effective preprocessing to improve performance.Nabila Zaman et al. [2] implemented models like Logistic Regression and KNN, with Naive Bayes achieving 96.23% accuracy, but the model requires greater validation for general applicability.Shirajul Islam et al. [3] emphasized early ASD detection, with KNN achieving 98% accuracy, though limited by small datasets.

Deepa M et al. [4] applied XGBoost and KNN, achieving 98.2% accuracy, though concerns exist regarding dataset bias and generalizability.Kaushik vakadkar et al.[5] developed model using SVM and Random Forest with Logistic Regression reaching 97.15% accuracy, though large datasets were scarce.Astha Baranwal et al. [6] applied LDA and KNN, achieving 72.2% accuracy but stressed the need for larger datasets for reliable predictions.

Ritu Chauhan et al. [7] used RF and SVM, highlighting ethical issues and regulatory challenges in healthcare applications, with RF reaching 74% accuracy.Aishwarya D et al. [8] used Neural Networks and Gradient Boost, achieving up to 99% accuracy in predicting ASD, offering cost-efficient solutions but with some limitations in dataset.Yong Jeon Cheong et al.[9] integrated epigenetic and brain data, with Random Forests achieving 97% accuracy, though the study lacked generalizability across diverse populations.

Konstantinos-Filippos Kollias et al. [10] explored robotics in ASD diagnosis, enhancing emotional support, though sample sizes were small and varied in effectiveness.Shreea Bose et al. [11] applied XGBoost, achieving 100% accuracy but noted the limitation of feature independence affecting model sensitivity.Konstantinos-Filippos Kollias et al. [12] used RNN and SVM, with high classification accuracy, though biases in training data and sample size variability limit its broader applicability.

Bhawana Tyagi et al.[13] used SVM, LR , KNN, CART , LDA and Naive Bayes algorithms for classification and got the best result for Linear Discriminant Analysis with 72.2024% accuracy.Naurin Farooqi et al.[14] used XGB, GBC, ABC, SVM, RFC, DTC ,LR ,KNN and GNB algorithms for classification and got the best result for both RFC and SVM with 95% accuracy.Khushbu Garg et al.[15] used Decision Tree, KNN, Logistic Regression and Naive Bayes algorithms and also applied deep learning to get the best results, also they had done based on image-based data.

III. Methodology

A. Statement of the Problem

The main objective is to determine the prevalence and associated factors of Autism Spectrum Disorder (ASD) among toddlers based on demographic disparities as well as health conditions.

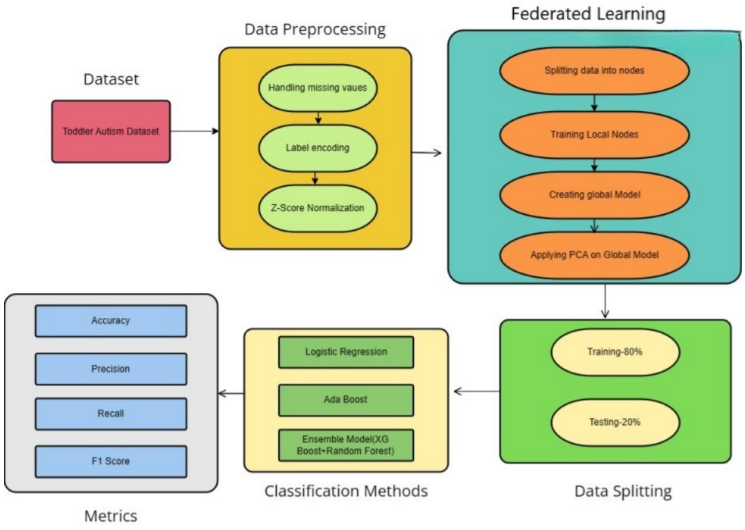

Fig. 1. Flowchart of Methodology

B. Data Collection

The recorded and analyzed patients' medical records, patient survey, and public health database for toddlers whose diagnosed ASD were made of with age and gender and significant health conditions of toddlers, and these sources were made up of the data. In addition to the above, the dataset also includes information regarding behavioral patterns, family history, and developmental milestones-the factors that are required to identify early signs of Autism Spectrum Disorder (ASD). The used dataset is available on Kaggle and was relied upon to both developing prediction modeling and exploratory data analysis to support the development of early detection strategies for toddlers suffering from ASD.

C. Exploratory Data Analysis:

An exploratory data analysis was used to detect and diagnose patterns, trends, or relationships between variables in the dataset. Histogram, scatter plot, and heatmaps were used to check on the distribution of ASD cases in terms of age and gender, including in relation to diverse influencing factors on the diagnosis of Autism Spectrum Disorder. At this stage, the influences of jaundice, heredity, and prior health issues were fully explored and assessed on their contribution to autism incidence rates. Finally, correlation matrices and regression analysis were conducted to further investigate potential interactions among these variables. Based on our detailed analyses, the figures below depict considerable correlations and present factors that may show a higher prevalence in toddlers with autism, which would become worthwhile for the detection and intervention strategies early on.

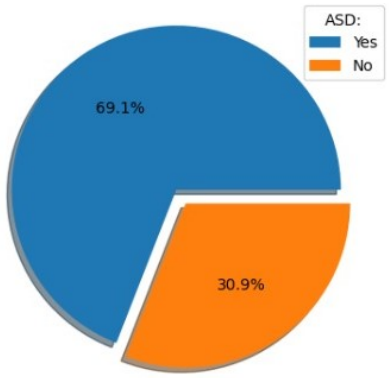

Fig. 2. Pie chart showing 69.1% of people with ASD.

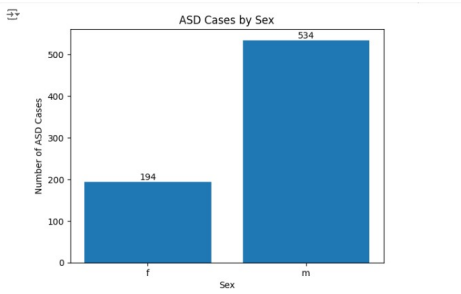

Fig. 3. ASD cases by sex, showing a higher prevalence in males

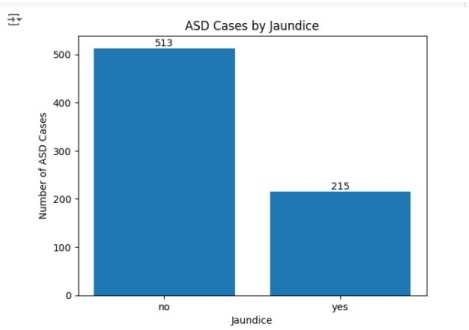

Fig. 4. ASD cases and jaundice prevalence.

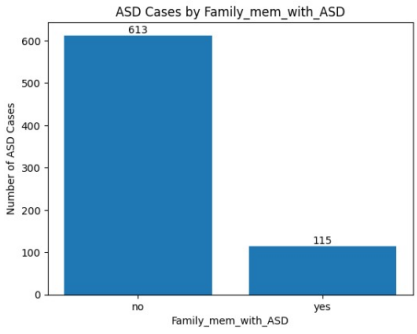

Fig. 5. ASD cases by family members with ASD.

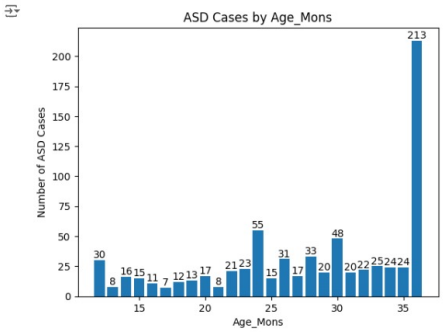

Fig. 6. ASD cases by age (months).

D. Data Preprocessing

Data pre-processing is probably the most critical step before actual analysis or training of the model over the data set. The phase includes the following:

1) Data Cleaning: In the dataset,we checked for the missing values and inconsistent values as well, and it was corrected accordingly:

Detection of Missing Values

For the each column of the dataset, test it to see whether it contains any missing values.Calculate the percentage of missing data for every feature, to evaluate how severe the problem is

Duplicates Elimination

The dataset had duplicated records, we cleaned to ensure that all the records were unique.

Anomalies Entries

The dataset was scanned for anomalies. These could be typographical errors in categorical variables (such as gender entries tagged as ”Male,” ”male,” and ”M” to become made uniform to ”Male”), or error in data type. Entries were standardized using Label encoding.

2) Normalization: Continuous variables were normalized to a common scale, which is very important for machine learning algorithms. The steps followed are as given below:

Z-score Normalization

We have applied normalization technique on the normal-distribution-like data, known as z-score normalization or standardization. It standardizes values of mean is 0 and standard deviation is 1, using a following formula:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

where:

- μ is mean, and
- σ is standard deviation of feature.

This normalization guarantees that all features are equally important while training the model. It plays a crucial role

in some algorithms, especially those using gradient descent-based methods, which are sensitive to feature scales.

E. Federated Learning

We use Federated learning, to ensure privacy and security as it trained models at local sites; thus, no site could access other sites' sensitive information. Some key steps included the following:
Split into nodes:
We have divided our dataset into 3 nodes, to train them locally and created a global model using their updates.
Local Training:
We have trained our nodes locally using logistic regression.It does share the Data they will only share the knowledge gained from training to global model.
Model Update:
Rather than transferring the raw data, each site communicated only model updates - gradients or weights - to a central server. The updates were then aggregated using techniques like Federated Averaging (FedAvg).
Application of PCA:
We had apply PCA on our global model to reduce the redundant features.After applying PCA we choose 10 components out of 17 components.

F. Train-Test Split

An 80-20 split was then applied to the dataset with the splits, thus creating both the training subsets and testing subsets.A former was used to train model while the latter was used for testing the performance of the models.

G. Application of Machine Learning Algorithms

For this purpose, various machinelearning algorithms were trained on the training data with a view to predicting prevalence about ASD:

1) Logistic Regression: Logistic Regression is statistical method for dealing with the problem of binary classification. This method provides the probability of the occurrence of an event, given one or more predictor variables. The logistic function takes the following form:

$$P(X = 1|Y) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_2 + ... + \alpha_n Y_n)}} \quad (2)$$

Where $P(X = 1|Y)$ is probability of positive class, α_0 is intercept, $\alpha_1, \ldots, \alpha_n$ are coefficients, and $Y_1, \ldots, Y_n$ are features.

2) K-Nearest Neighbors (KNN): KNN is non-parametric classifier, which simply assigns class to point based upon the classes of its 'k' nearest neighbors in feature space. The classification rule can be summarized as follows:

$$\hat{y} = \text{argmax}c \sum i = 1^k I(y_i = c) \quad (3)$$

Where $\hat{y}$ is the predicted class, c is a class label, I is an indicator function, and y_i are the classes of the nearest neighbors.

3) Support Vector Machines (SVM): SVM is highly effective classification method that identifies the best hyperplane that separates the data points of distinct classes from each other in high-dimensional space. The decision function can be written as:

$$f(Y) = \text{sign}\left(\sum_{i=1}^{n} \beta_i x_i A(Y, Y_i) + c\right) \quad (4)$$

Where β_i are Lagrange multipliers, x_i are the target labels, A is kernel function, and c is bias term.

4) Decision Trees: Decision Trees are those models that, based on the feature's values, split the data recursively into a tree structure. In this case, every internal node is feature, each edge is decision rule, and each leaf corresponds to an outcome. For instance, a splitting criterion may be noted as Gini impurity or entropy:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2 \quad (5)$$

Where p_i is proportion of instances of class i in dataset T and n is number of classes.

5) Naive Bayes: Naive Bayes is the training classifier adapted from Bayes' theorem, assuming independence between predictors. It is really good for text classification tasks. The posterior probability can be calculated as:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (6)$$

Where $P(Y|X)$ is posterior probability, $P(X|Y)$ is likelihood, $P(Y)$ is prior probability, and $P(X)$ is marginal likelihood.

6) AdaBoost: Adaptive Boosting is ensemble method to combine weak classifiers' outputs in order to produce strong classifier. It adjusts the weights of false classified instances to lay more emphasis on harder cases. The final model prediction is obtained from:

$$H(X) = \sum_{m=1}^{M} \alpha_m h_m(X) \quad (7)$$

Where $H(X)$ is the final prediction, $h_m(X)$ are the weak classifiers, and α_m are the weights assigned to each classifier based on their performance.

7) Ensemble (XGBoost + Random Forest): This model combines the strengths of XGBoost and Random Forest through ensemble learning. XGBoost utilizes gradient boosting to optimize model performance, while Random Forest builds multiple decision trees for robustness. The general formula for XGBoost can be represented as:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{8}$$

Where $F(x)$ is the final prediction, $h_m(x)$ are the trees, and γ_m are the weights.

H. Model Evaluation

The performance of the applied machinelearning models for prediction of prevalence of Autism Spectrum Disorder was tested using a number of different metrics:

1) Accuracy: Accuracy gives an percentage of true instances correctly predicted over the total number of instances. It can therefore be used to gauge how well the model performs. Accuracy is computed by formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

2) Precision: Precision is the ratio of correctly predicted positive cases out of all positive predictions made by the model, which represents the model's ability to avoid false positives. It is very crucial in scenarios where the cost of false positives is very high, such as in medical diagnosis or fraud detection scenarios where incorrect positive prediction may cause heavy penalties. The precision ratio is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

3) Recall: Recall, sometimes referred to as Sensitivity, is the ratio of rightly recognized positive instances, which the model actually is. It is important when missing a positive instance incurs a high cost. The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

4) F1-Score: The F1-score would be the harmonic mean between precision and recall, so these two would be balanced.It also finds a good application when one needs to find a balance between precision that is the ability of a model to avoid false positives, and recall, which is the ability to catch true positives. It is particularly helpful in scenarios where there is an issue of managing imbalanced datasets. Formula for F1-score:

$$F1 = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{12}$$

I. Comparison of existing model and proposed model

In our proposed model the most important thing we have added is federated learning which ensures data privacy of patients by training the data locally and sharing only updates. We have also improved accuracy as compared to those who have worked on our dataset. Kaushik vakadkar et al.[5] developed model using SVM and Random Forest with Logistic Regression reaching 97.15% accuracy .
The current model which we have proposed got an accuracy of 100% for logistic regression, adaboost And ensemble(xgb+rf) models.
Our model has also improved the precision recall and f1 score values.
Here the comparision table

TABLE I
MODEL ACCURACY COMPARISON

Model	**Accuracy**
Logistic regression (existing model)	97.15%
Naïve Bayes (existing model)	94.79%
SVM (existing model)	93.84%
Logistic regression (proposed model)	100%
Ada Boost (proposed model)	100%
Ensemble (XGB + RF) (proposed model)	100%

IV. RESULTS AND ANALYSIS

In this study, we used a dataset available on Kaggle, provided by Dr. Fadi Fayez Thabtah, to predict autism disorder in toddlers using various machine learning algorithms. Here we applied some machine learning algorithms like Logistic Regression, Adaboost, K Nearest Neighbor, Naive Bayes, Support Vector Machine, Decision Tree and ensemble method(XGB +RF).We used all these algoriths to detect the autism in toddlers.We got the highest accuracy for some algorithms to detect the autism. The algorithms such as logistic regression, adaboost,and ensemble method(XGB +RF) achieved an accuracy of 100% in detecting autism in toddlers. additionally we used federated learning for data privacy. The performance of each algorithm is summarized below in the form of confusion matrix:

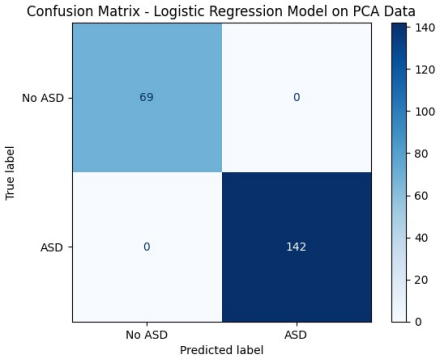

Fig. 7. Confusion Matrix-Logistic Regression

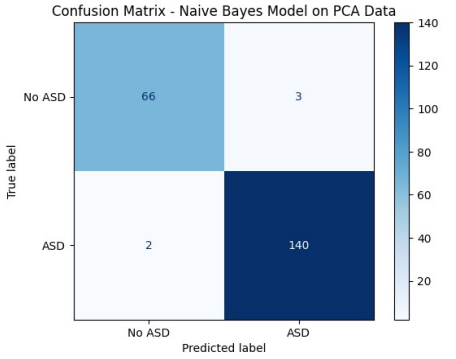

Fig. 8. Confusion Matrix-Naive Bayes

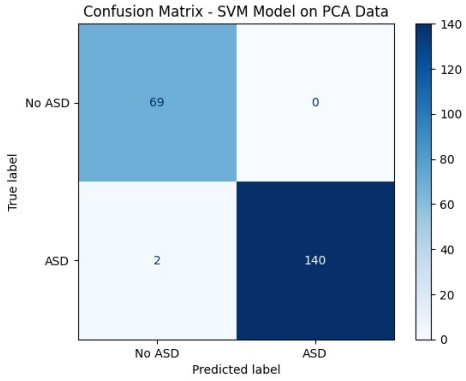

Fig. 9. Confusion Matrix-SVM

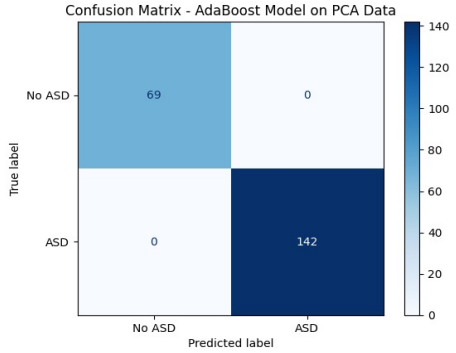

Fig. 10. Confusion Matrix-Adaboost

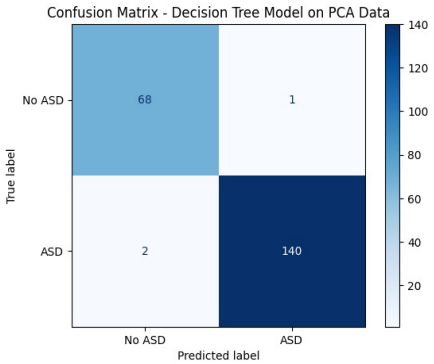

Fig. 11. Confusion Matrix-Decision Tree

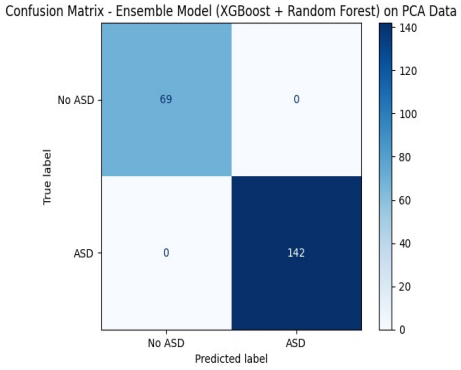

Fig. 12. Confusion Matrix-Ensemble Method(XGB+RF)

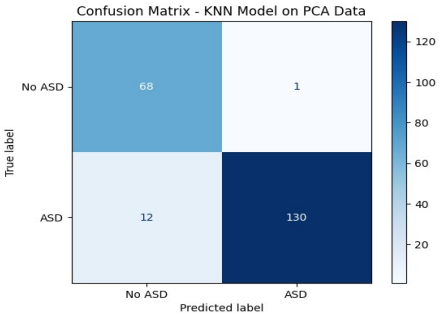

Fig. 13. Confusion Matrix-KNN

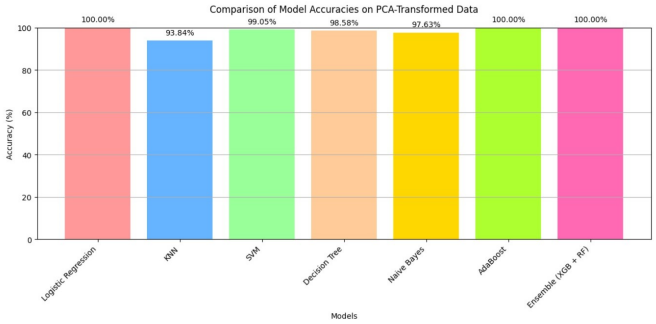

Fig. 14. Comparision of model accuracies.

V. Conclusion

We found that the Logistic Regression model,AdaBoost, Ensemble Method(XGB+RF) was 100% accurate, and thus they remain among the best classifiers for ASD while precision, recall, and F1-score balance is perfect. KNN was somewhat less reliable with accuracy at 93.84% which, although fair, ranked very low amongst the models. SVM model showed more strength at about 99.05%. Accuracy is pretty high and produces low false positives as well as negatives. Decision Tree had established good predictive power at 98.58% accuracy while Naive Bayes was also promising at 97.63% and, therefore generalized well even though the accuracy was low. The three with 100% accuracy were AdaBoost, Logistic Regression and the Ensemble model (XGBoost + Random Forest).

TABLE II
Model Performance Metrics in Percentage

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	100.00%	100.00%	100.00%	100.00%
KNN	93.84%	99.00%	92.00%	95.00%
SVM	99.05%	100.00%	99.00%	99.00%
Decision Tree	98.58%	99.00%	99.00%	99.00%
Naive Bayes	97.63%	99.00%	98.00%	99.00%
AdaBoost	100.00%	100.00%	100.00%	100.00%
XGB + RF	100.00%	100.00%	100.00%	100.00%

References

[1] N. BalaKrishna, M. B. Mukesh Krishnan, S. M. Reddy, S. K. Irfan and S. Sumaiya, "AUTISM Spectrum Disorder Detection Using Machine Learning," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1645-1650, doi: 10.1109/ICACITE57410.2023.10183095.

[2] N. Zaman, J. Ferdus and A. Sattar, "Autism Spectrum Disorder Detection Using Machine Learning Approach," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579522.

[3] S. Islam, T. Akter, S. Zakir, S. Sabreen and M. I. Hossain, "Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411531.

[4] S. K. R. Naik, D. M, R. P B, S. Prakash and U. J. Royal, "Determination and Diagnosis of Autism Spectrum Disorder using Efficient Machine Learning Algorithm," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-5, doi: 10.1109/CONIT59222.2023.10205718.

[5] Vakadkar, K., Purkayastha, D. Krishnan, D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN COMPUT. SCI. 2, 386 (2021). https://doi.org/10.1007/s42979-021-00776-5.

[6] A. Baranwal and M. Vanitha, "Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-7, doi: 10.1109/ic-ETITE47903.2020.186.

[7] R. Chauhan, K. Mehta, Y. Eiad and M. F. Zuhairi, "Prediction of Autism Spectrum Disorder Using AI and Machine Learning," 2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM), Kuala Lumpur, Malaysia, 2024, pp. 1-7, doi: 10.1109/IMCOM60618.2024.10418312.

[8] A. D, C. R. P, N. M and M. K, "Intelligent Autism Disease Prediction System Using Machine Learning," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 1146-1151, doi: 10.1109/ICIRCA57980.2023.10220779.

[9] Y. J. Cheong et al., "Prediction of autism spectrum disorder using epigenetic, brain, and sensory behavioral factors," 2024 12th International Winter Conference on Brain-Computer Interface (BCI), Gangwon, Korea, Republic of, 2024, pp. 1-4, doi: 10.1109/BCI60775.2024.10480486.

[10] K. -F. Kollias, L. M. Maia Marques Torres E Silva, P. Sarigiannidis, C. K. Syriopoulou-Delli and G. F. Fragulis, "Implementation of Robots in Autism Spectrum Disorder Research: Diagnosis and Emotion Recognition and Expression," 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST), Athens, Greece, 2023, pp. 1-4, doi: 10.1109/MOCAST57943.2023.10176588.

[11] S. Bose and P. Seth, "Screening of Autism Spectrum Disorder using Machine Learning Approach in Accordance with DSM-5," 2023 7th International Conference on Electronics, Materials Engineering Nano-Technology (IEMENTech), Kolkata, India, 2023, pp. 1-6, doi: 10.1109/IEMENTech60402.2023.10423494.

[12] K. -F. Kollias, C. K. Syriopoulou-Delli, P. Sarigiannidis and G. F. Fragulis, "The contribution of Machine Learning and Eye-tracking technology in Autism Spectrum Disorder research: A Review Study," 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, Greece, 2021, pp. 1-4, doi: 10.1109/MOCAST52088.2021.9493357.

[13] B. Tyagi, R. Mishra and N. Bajpai, "Machine Learning Techniques to Predict Autism Spectrum Disorder," 2018 IEEE Punecon, Pune, India, 2018, pp. 1-5, doi: 10.1109/PUNECON.2018.8745405.

[14] N. Farooqi, F. Bukhari and W. Iqbal, "Predictive Analysis of Autism Spectrum Disorder (ASD) using Machine Learning," 2021 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2021, pp. 305-310, doi: 10.1109/FIT53504.2021.00063.

[15] K. Garg, N. N. Das and G. Aggrawal, "A Review On: Autism Spectrum Disorder Detection by Machine Learning Using Small Video," 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2023, pp. 1-8, doi: 10.1109/ICCT56969.2023.10076139.