Unlocking Autism: A Machine Learning-Based Approach for Early Diagnosis

Chalapati Sowmya[1], Maridu Bhargavi[2], Sikhinam Mercy[3], Koduru Jhansi Suvarchala[4], Shaik Mahmooda Aafreen[5]

[1,2,3,4,5]Department of CSE, Vignan's Foundation for Science,Technology and Research,, Vadlamudi, 522213, Andra Pradesh, India.

[1]221fa04074@gmail.com.

[2]bhargaviformal@gmail.com.

[3]221fa04113@gmail.com.

[4]221fa04140@gmail.com.

[5]221fa04143@gmail.com.

Abstract

ASD is a highly complex neurodevelopmental disorder with social interaction and communication challenges that diagnosis, in most cases, involves laborious behavioral analysis and can delay much-needed early interventions. The following paper introduces the use of ML in the detection of ASD using models such as Logistic Regression, AdaBoost, and ensemble approaches like XGBoost and Random Forest at an extremely high accuracy rate. It has been recognized that current methods toward data privacy are short of the mark using federated learning(FL) , our solution has enabled model training locally in decentralized devices that guarantee patient confidentiality. Since there's an aggregation of updates rather than actual raw data, there is therefore a higher preservation of privacy while maintaining efficiency at collaborative learning. This experiment shows its FL-based metaclassifier with ASD-specific data with higher diagnostic accuracy, a better model, and significantly greater advancements in safe accurate early detection strategies for ASD.

Keywords: Machine Learning,Early Diagnosis,Logistic Regression,Autism Spectrum Disorder,XGBoost,AdaBoost,Predictive Modeling

1 Introduction

Autism spectrum disorder represents a complex neurodevelopmental condition characterized by a vast array of symptoms, involving marked impairment in social interaction and communication, besides obviously abnormal behavior. Diagnosis generally requires long-term behavioral observations coupled with subjective clinical evaluations and is essentially time-consuming as well as incongruous. Thus, the significant role that timely intervention assumes in molding developmental outcomes may often be delayed. In fact, the traditional methods of diagnostics, though clinically robust, fail to meet the demand of rapid detection with accuracy, an interest in technological solutions is motivated to enhance diagnostic efficiency. This gap in swift and accurate diagnosis underscores the potential of advanced technologies, such as machine learning, to support clinicians in faster, data-driven assessments. Leveraging these tools could significantly enhance diagnostic efficiency, enabling earlier interventions that are vital for developmental progress in individuals with autism spectrum disorder.

Among the promising lines of support for early detection of ASD is machine learning, which can now analyze vast behavioral and demographic data to discover complex patterns for diagnosis. The various machine learning models that have been applied in ASD detection range from Logistic Regression to Support Vector Machines (SVM) and even Decision Trees. Some of the models have demonstrated remarkable accuracy. As depicted in the above discussion, for instance, BalaKrishna et al.[1] SVM, Decision Trees, and Logistic Regression were employed in the ASD classification, and SVM achieved a very high 93% accuracy. Despite such issues as generalization problems of data and ethical factors of health data like privacy, the study seemed to have controlled small or nondiversified datasets.

Another study by Zaman et al.[2] applied Logistic Regression models and K-Nearest Neighbors (KNN), which resulted in reaching an accuracy of 96.23%. Again the authors noted that to enable the applicability of their model in other population and ethnic groups, a further validation of the model becomes important. Such research epitomizes the application capacity of ML algorithms into diagnoses of ASD but severely holds it back in another category: privacy of the information. Traditional ML approaches toward ASD diagnosis usually imply centralization of data and raise a high risk on confidentiality of the patient with all the ethical issues raised during the processing of personal health information.

These concerns can be solved by the strategy known as Federated Learning. FL is a form of decentralized machine learning since the models are trained locally on individual devices rather than relying on a centralized database. In FL, only the model updates are transferred to a central server. This framework makes FL ideal for healthcare applications, as patient data sensitivity is their priority. FL decreases the necessity of data centralization, which makes collaborative learning without compromising privacy an extremely attractive approach for ASD detection.

Applying FL helps to overcome the current privacy and accuracy bottlenecks in diagnosing the disease. Under FL, data is kept on device, and only updates at any point are transmitted where required, thus keeping this information confidential to the patients and enhancing model robustness. The work uses ensembling of models that achieve high precision diagnostics without lowering data on privacy by integrating Logistic

Regression, AdaBoost, and a combination model from XGBoost with the Random Forest.

Our main contributions in this study are as follows:

- Federated Learning Framework: Patient data will not be leaked out since they will be localized on the decentralized devices.
- High-Accuracy Models: It employs an ensemble of optimally optimized ASD diagnosis ML algorithms.
- Privacy-Enhanced Model: sends only aggregated updates ensuring the privacy of healthcare data.
- Performance Benchmarking: It demonstrates higher accuracy and privacy than the best known detection methods of ASD.

The paper follows the following structure, Section 1 introduces the paper, Section 2 gives a literature review, Section 3 describes the proposed workflow and methods in detail, Results and Analysis is given in Section 4 and Conclusion in Section 5.

2 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. 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. Islam et al.[3] emphasized early ASD detection, with KNN achieving 98% accuracy, though limited by small datasets.

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

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

Kollias et al.[10] explored robotics in ASD diagnosis, enhancing emotional support, though sample sizes were small and varied in effectiveness. Bose et al.[11] applied XGBoost, achieving 100% accuracy but noted the limitation of feature independence affecting model sensitivity.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.

3 Methodology

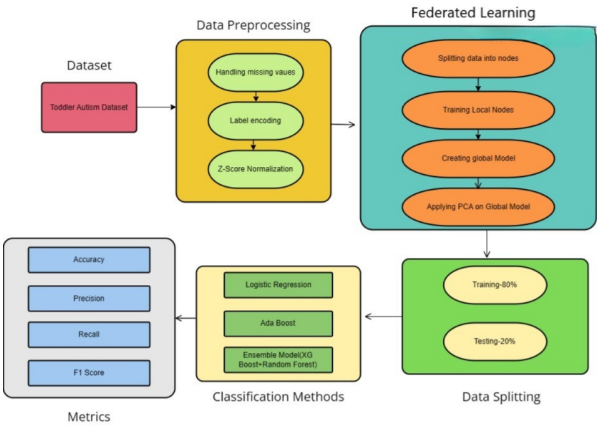

Fig. 1 Flowchart of Metholodgy

3.1 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.

3.2 Data Collection

The patients medical records, a survey of patients, and data from public health databases for toddlers whose diagnoses of ASD were recorded and analyzed, along with their age, gender, and significant health conditions, constituted the sources of the data.

3.3 Exploratory Data Analysis:

We carried out exploratory data analysis in order to point out and diagnose the patterns, trends, or correlations within the data Histograms, scatter plots will be used to present the findings for the distribution of ASD cases by age and gender.

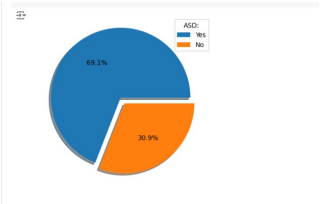

Fig. 2 Pie chart showing 69.1% of people with ASD.

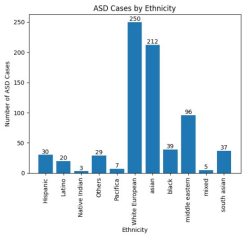

Fig. 3 Bar chart showing ASD cases by ethnicity.

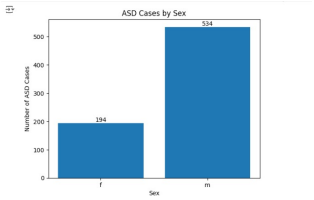

Fig. 4 ASD cases by sex, showing a higher prevalence in males.

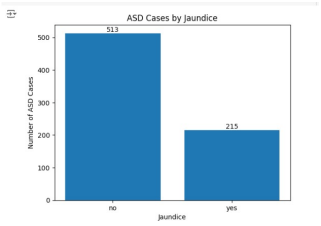

Fig. 5 ASD cases and jaundice prevalence.

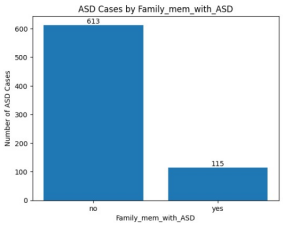

Fig. 6 ASD cases by family members with ASD.

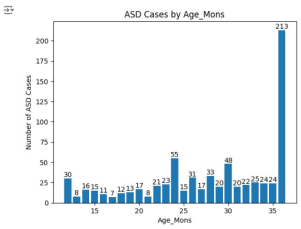

Fig. 7 ASD cases by age (months).

3.4 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:

3.4.1 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.

3.4.2 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.

3.5 Federated Learning

We have used Federated learning, to ensure privacy and security as it trains 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 got 10 components out of 17 components.

3.6 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.

3.7 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:

3.7.1 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:

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

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.

3.7.2 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:

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

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.

3.7.3 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.

3.7.4 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:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2 \tag{4}$$

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

3.7.5 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:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{5}$$

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

3.7.6 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.

3.7.7 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:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{6}$$

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

3.8 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:

3.8.1 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.

3.8.2 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.

3.8.3 Recall

Recall 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.

3.8.4 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 and recall.

3.9 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 1 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%

4 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. The performance of each algorithm is summarized below:

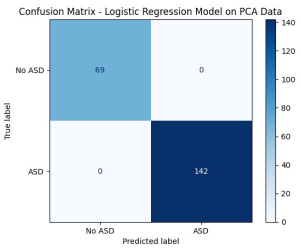

Fig. 8 Confusion matrix for LR

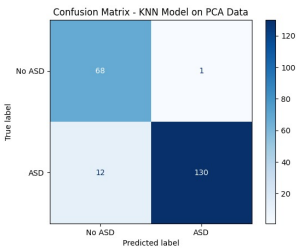

Fig. 9 Confusion matrix for KNN

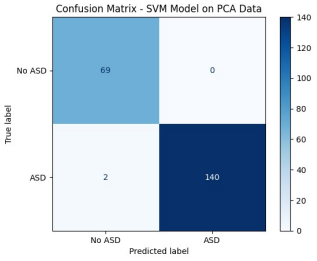

Fig. 10 Confusion matrix for SVM

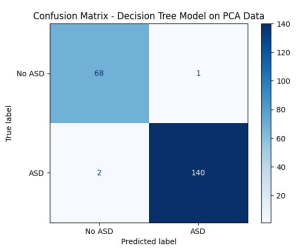

Fig. 11 Confusion matrix for decision tree

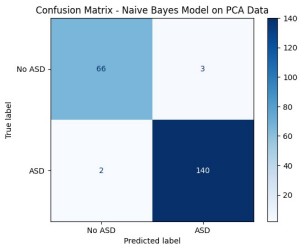

Fig. 12 Confusion matrix for Naive Bayes

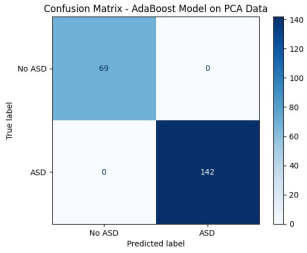

Fig. 13 Confusion matrix for Adaboost

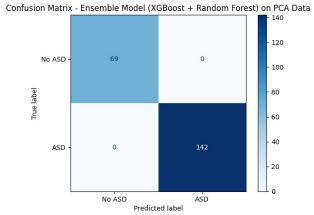

Fig. 14 Confusion matrix for Ensemble(XGBOOST + RF)

5 Conclusion

This study introduces the design of a safe, highly accurate framework for early detection of Autism Spectrum Disorder (ASD) using federated learning combined with ML techniques to achieve 100% accuracy rating. FL will ensure that the patient data is kept decentralized in such a manner that it does not lose the performance in the ability to diagnose.This framework addresses the ethical and regulatory concerns because it is transmitting only the aggregated updates of the models instead of raw data, and this secures collaborative learning in health care. Results are thus obtained which depict improvements much higher than traditional methods both on accuracy and on privacy level, hence bringing a big potential for transformation in diagnosis of ASD.

Table 2 Model Accuracy Comparison

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	98.00	99.00	98.00
AdaBoost	100.00	100.00	100.00	100.00
Ensemble (XGB + RF)	100.00	100.00	100.00	100.00

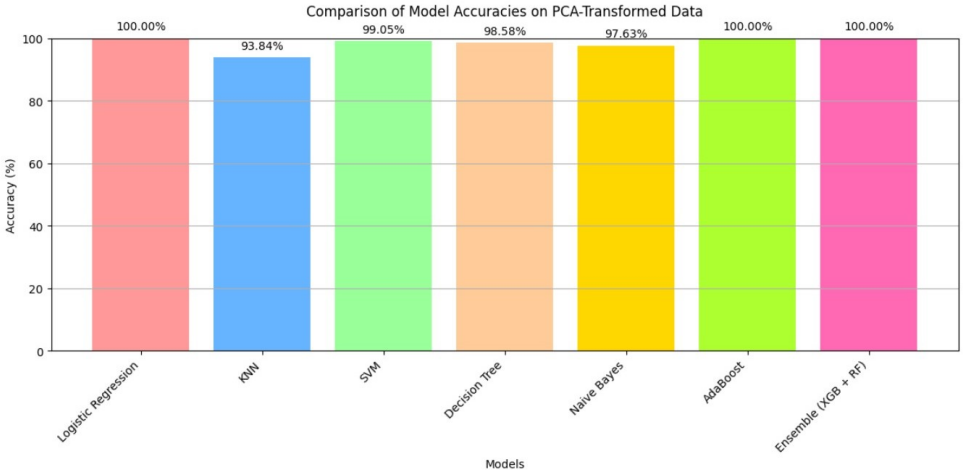

Fig. 15 Comparison of Accuracies

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.