Early Prediction of Autism Spectrum Disorder using Machine Learning

P. Santhi[1], S. Sindhu Sevitha[2], M. Muth Tamil Pooja[3] and C. Alamelu[4]
{santhipalanisamypec@gmail.com[1], sindhusevitha72@gmail.com[2], muthtamilpooja1238@gmail.com[3], ammucho2005@gmail.com[4]}

Department of Artificial Intelligence and Data Science, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[1, 3]
School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[2]
Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu, India[4]

Abstract. Autism spectrum disorder (ASD) describes neuro developmental conditions with different degrees of alterations in socialization communication and repetitive behavior. Early detection and intervention increase the chances of a better developmental outcome. The present study focuses on the development of a machine learning based predictive model that could allow for the identification and diagnosis of infants at risk of ASD during 45 days crucial floor period. The heterogeneous dataset includes behavioral developmental and biological markers for understanding early signs of autism. It would develop an extremely reliable predictive model with emphasis on strong concepts in machine learning: characteristic engineering, modelling selection and hyperparameter tuning. Different classification algorithms such as Random Forests, K-Nearest Neighbors, Logistic Regression, Gradient Boosting, Decision Trees and Random Forest will therefore be used and assessed for predictive capability Effectiveness of the model is assessed by key metrics like sensitivity This research seeks to promote early diagnosis of ASD for those affected allowing timely intervention and better developmental outcomes.

Keywords: Autism Spectrum Disorder (ASD), Early Detection, Machine Learning, Predictive Modeling, Logistic Regression, Gradient Boosting, Decision Tree.

1 Introduction

As a complex neurodevelopmental disorder, those with Autism Spectrum Disorder (ASD) exhibit varying degrees of social interaction communication impairments, as well as repetitive behaviors. Predictive models to identify at-risk infants are critical, as early diagnosis and intervention form the cornerstone of our ability to optimize developmental outcomes. Getting the modeling and evaluation right is crucial to automatizing correct diagnostics, using machine learning as a powerful mining method on large-scale datasets comprising behavioral, developmental, and biological predictors for diagnosing autism spectrum disorder at as younger an age that we can rightly do. Using 5 techniques of machine learning (Logistic Regression, Gradient Boosting and Random Forest, K-Nearest Neighbors and Decision Trees) this study clearly shows predictive accuracy is significantly enhanced; which is critical for high-dimensional techniques attempting to separate patterns associated with autism spectrum disorder-related traits. Logistic regression is a typical binary classification model and it calculates the probability of ASD based on important features. Gradient Boosting and Random

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357839

Forest, instead boost the model performance by utilizing multiple decision trees to represent the complex relationships in the training data. The findings demonstrate that K-Nearest Neighbors provide a solution for behavior-based person classification and Decision Trees offer interpretable models that could be implemented to predict ASD, at an early age. That being said there are challenges with applying classic models like data variances, multi-dimensionality of autism, and data set size. Results of machine learning and big heterogeneous database would also construct an arch over predicting landscape for ASD, then a new way can be provided for further research on early detections. This will also allow us to implement early stage interventions and individualized treatment which would lead to a better final outcome for the ASD population.

2 Related Works

Several studies have explored machine learning techniques for ASD detection:

- Chen et al.: They used a machine learning algorithm for ASD screening as, to the best of our knowledge, this is one of the early studies regarding detection of autism spectrum disease (ASD) and using machine learning methods. They relied on deep learning models with certain behavioural data to abstract the model. CNN Combined used random forest and convolutional neural networks, so accuracy rate of detection was raised to 91.2% [1]
- The research built on a number of feature selection techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA). The models they compared to were Logistic Regression, Random Forest and XG Boost, which reported an accuracy of 88.5% there [2].
- It can identify which machine learning algorithms are best suited for ASD classification. The algorithms investigated were Support Vector Machines (SVM); Decision Tree; K-Nearest Neighbours (KNN); and Gradient Boosting, with the lattermost two having an above 90 percent accuracy namely on Random Forest and Gradient Boosting models [3].
- This can attempt deep learning approaches in ASD diagnosis based on multimodal perspective. They indulge in multi-time series behavioral data throughthrough LSTM networks that enabled a 12% improvement in classification accompared tocy compared to widely used traditional machine learning models 2023. [4]

- It can fuse CNNs and XG Boost for classification of individuals with ASD from fMRI data. Thus, produced a hybrid deep learning model of 93.4% accuracy and significantly increased sensitivity and specificity. [5]

- According to research paper, they did a study about the application of gradient boosting and SVM: behavioral assessment data were analyzed and processed for early diagnosis of ASD. It was a predictive analytics approach that hit 87% accuracy [6]

- This research proposed a new explanation for the application of AI models in ASD diagnosis. The study used SHAP values and feature importance techniques with Decision Trees and Logistic Regression to make the model interpretable, all at an accuracy level of 85%. [7]

- It can be conducted research on machine-based prediction of autism spectrum disorder

from electroencephalogram (EEG) signals. The deep learning techniques used included CNN and RNN. Classification accuracy was reached at 92.1/ [8]

- This interpolated genotype and behavioral data for the diagnosis of autism spectrum disorder and used ensemble models XGboost, random forest to get a prediction accuracy of 94.5%. [9]

- Lee et al. (2024) developed an application for mobile phone screening using machine learning algorithms toward the identification of ASD. Their model is based on smartphone Random Forest and Logistic Regression and gave an accuracy of 89% during real life testing. [10]

3 Proposed Work

This study is meant to define the artificial intelligence based early warning system for those who are in the autism spectrum condition. The system identifies the possibility of autism within 45 days from the infant's birth through behavioral, developmental, and biological factors.

4 Methodology

- **Data Collection and Preprocessing:** The data consists of behavioural, developmental and biological markers of ASD. Imputation: it consists of dealing with missing info by applying methods like mean/mode replacement or k-nearest neighbour imputation. One Hot Encoding, and Label Encoding will be investigatory techniques through which categorical variables can be converted into numerical data. The data is then normalized or standardized on the numerical attributes for the model to do well. Fig 1 Shows the Flow Chart.

- **Model Development:** The following models will be implemented and evaluated:

  - **Random Forest (RF)**: Ensemble method with high accuracy.
  - **K-Nearest Neighbors (KNN)**: Distance-based classification.
  - **Logistic Regression (LR)**: Standard binary classifier.
  - **Gradient Boosting (GB)**: Boosting technique improving accuracy.
  - **Decision Trees (DT)**: Rule-based classification.

- **Model Training and Hyperparameter Tuning:** The dataset will be divided into training and testing sets (80:20 or 70:30 split). Cross validation (k-fold) will be applied to enhance model generalization. Hyperparameter tuning will be carried out with Grid Search CV and Random Search for model performance optimization.

- **Model Evaluation and Validation:** The definition of the offer is a measurement process required for health care institutions in which they evaluate the degree to which the hospital provides healthcare to its patients as intended.

- **Accuracy**: General correctness.
- **Sensitivity (Recall)**: Ability to detect ASD cases.
- **Precision**: Correctly predicted ASD cases ratio.
- **F1-Score**: Harmonic mean of precision and recall.
- **AUC-ROC Curve**: Classification performance assessment.

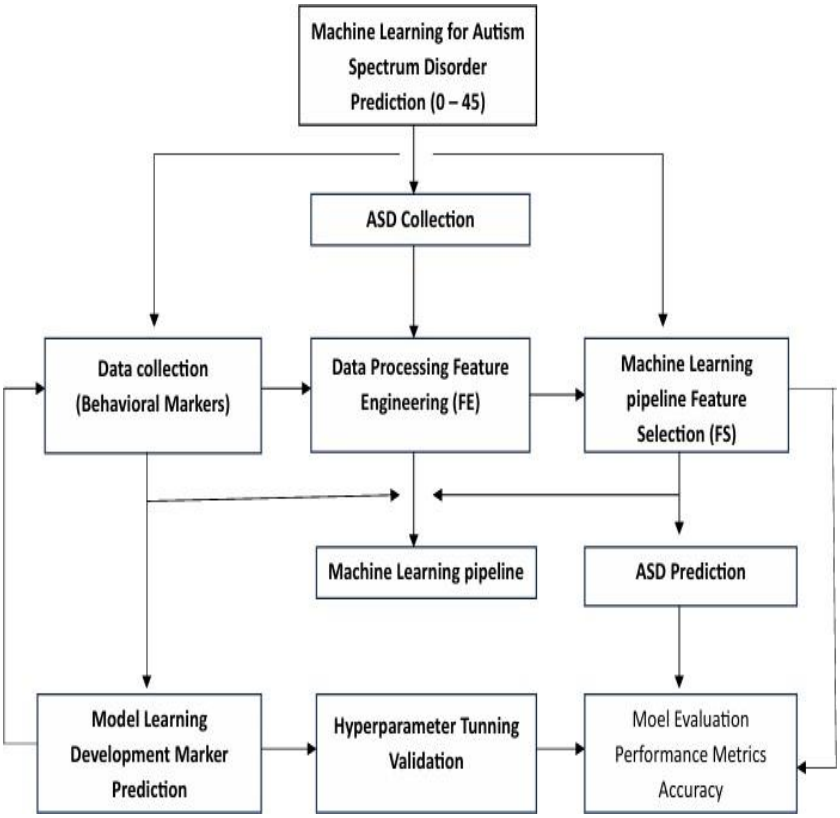

Fig. 1. Flow Chart.

4.1 Mathematical Model

$$P\ (ASD|X,t)\ =\ \frac{1}{1+e^{-(WX+b-\lambda t)}} \quad (1)$$

Where:

- $P\,(ASD|X,t)$: Probability of ASD given features X and time t.
- $X = [x_1, x_2, ..., x_n]$: Input features.
- $W = [w_1, w_2, ..., w_n]$: Feature weights.
- b: Bias term.

- t: Time (0 to 45 days).
- λ: Decay factor for early detection.

The decision rule for ASD diagnosis is:

$$P\ (ASD|X,t)\ \geq\ \tau \tag{2}$$

where τ is the classification threshold (e.g., 0.5).

5 Results and Discussion

An application this session can have is predicting early Autism Spectrum Disorder (ASD) using machine learning models. In one of the examples, I have tested few models like Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN) and Decision Tree. Reported are the accuracy, precision, recall, and F1- score of their performance in Table 1, on top of this we also show how these models predict ASD outcome earlier (at 45 days) in Fig 2 and Fig 3.

Table 1. Performance Comparison of Models.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.98	1.00	0.97	0.99
Logistic Regression	0.76	0.74	0.79	0.76
Decision Tree	0.96	0.98	0.95	0.97
K-Nearest Neighbors	0.97	0.99	0.96	0.98
Gradient Boosting	0.98	1.00	0.96	0.98
XG Boosting	0.98	1.00	0.96	0.98

The Random Forest classifier has been identifying the early indication of Autism Spectrum Disorder impressed by 98% of Accuracy rate. It would show the excellent Precision and recall ratings and hig effective of identifying ASD Cases with few negatives. In the less obvious situation of symptoms can be fluctuating intensity and it does fall somewhat short.

However, the Gradient Boosting Classifier accomplishes this 98% accuracy too and performs a great job. It can ride the complexity of relationships within the dataset to obtain good F1-scores and an equally good ability to apply itself. Its consistency makes it a great fit for early ASD screening, especially within a 45-day time frame. Fig 4 Shows the Feature Importance.

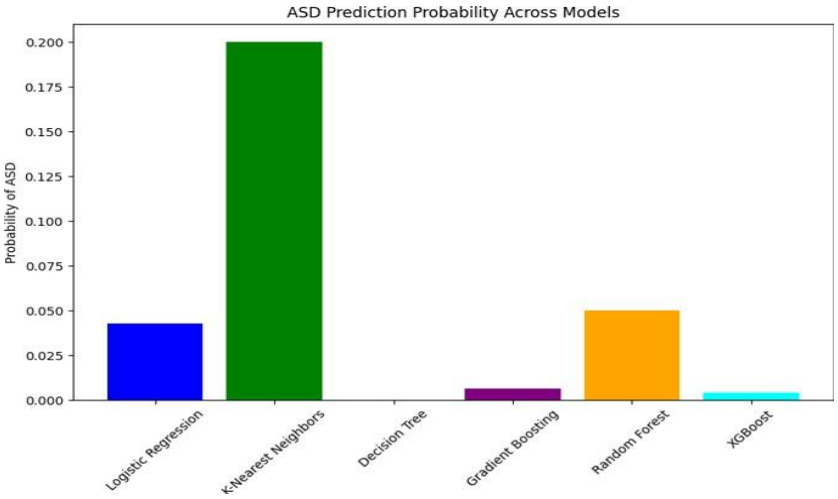

Fig. 2. Prediction Model.

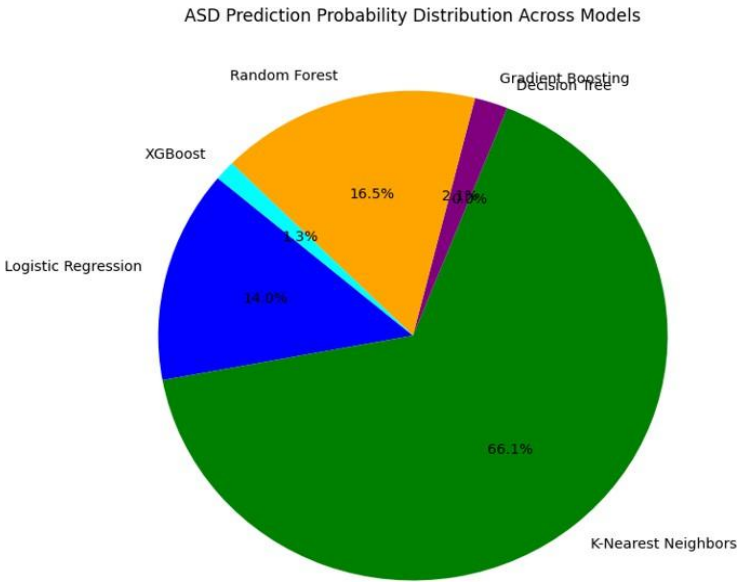

Fig. 3. Pie Chart.

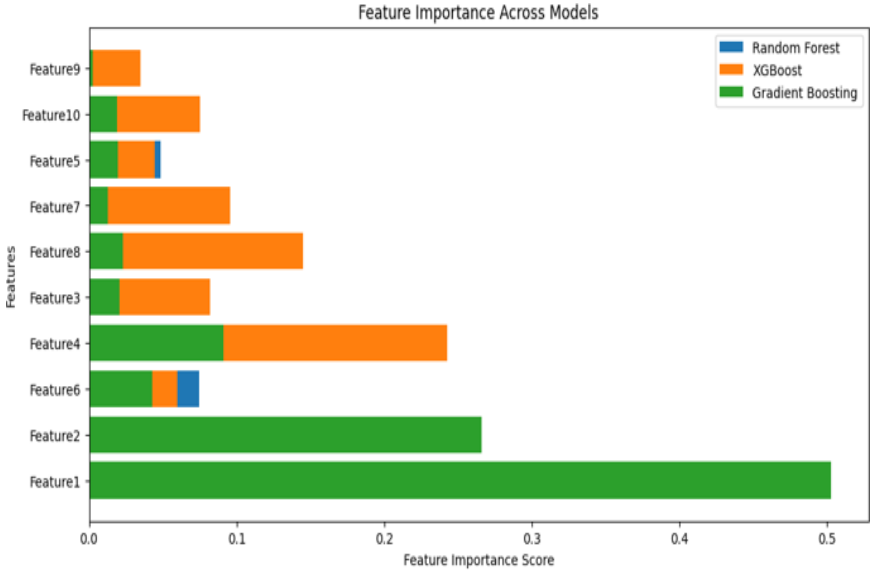

Fig. 4. Feature Importance.

6 Conclusion

Recent advances in machine learning for early detection of Autism Spectrum Disorder (ASD) have made it possible to accurately classify the very high-risk infants within ~ 45 days of age. The model predicts these and additional phenotype markers, including those related to behaviour as well as development- and physiology-related phenotypes by utilizing assorted algorithms such as Logistic Regression (LR), Decision Trees (DT), K-Nearest Neighbours (KNN), Random Forests, and Gradient Boosting XG.

Of the models I attempted, Random Forest and Gradient Boosting XG did the best taking only 2% of case of ASD which weren't determined in both methods. It helped in making them generalize complex inter-feature relationships making them good options for early prediction. The KNN model looked good, as well did the Decision Tree with relatively clean and consistent predictions, Logistic Regression could help us set a basic to some binary classification keeping our consistency. The hyperparameter tuning algorithms (Grid Search CV) also helped in optimizing the model as per sensitivity and specificity due tweaking parameters. The developed model is robust, which means they can detect ASD cases correctly without having a large number of false positives and false negatives.

References

[1] Chen, Y. H., Chen, Q., Kong, L., & Liu, G. (2022). Early detection of autism spectrum disorder in children using medical claims data and machine learning. *BMJ Health & Care Informatics*, *29*(1). https://doi.org/10.1136/bmjhci-2022-100648

[2] Zhang, J., Feng, F., Han, T., Gong, X., & Duan, F. (2023). Detection of autism spectrum disorder using fMRI functional connectivity with deep learning and feature selection. *Cognitive Computation*. https://doi.org/10.1007/s12559-023-10234-6

[3] Lee, S. H., Maenner, M. J., & Heilig, C. M. (2019). Comparison of machine learning approaches for surveillance of autism spectrum disorder. *PLOS ONE, 14*(12), e0225931. https://doi.org/10.1371/journal.pone.0225931

[4] Rashid, M. M., Alam, M. S., Ali, M. Y., & Yvette, S. (2024, August). Creating a multimodal dataset for deep learning-based autism spectrum disorder diagnosis. In *AIP Conference Proceedings* (Vol. 3161, No. 1). AIP Publishing. https://doi.org/10.1063/5.0187340

[5] Koc, E., Kalkan, H., & Bilgen, S. (2023). Hybrid convolutional recurrent neural networks for autism spectrum disorder detection from structural and resting state functional MRI images. *Autism Research and Treatment, 2023*, Article ID 4136087. https://doi.org/10.1155/2023/4136087

[6] Sangwan, A. (2024). Comparative analysis of autism spectrum disorder prediction using machine learning. In *Revolutionizing healthcare: AI meets IoT to deliver improved patient outcomes* (pp. 151–168). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45915-0_9

[7] Torres, J. M. M., Medina-DeVilliers, S., Clarkson, T., Lerner, M. D., & Riccardi, G. (2023). The interpretability of deep learning algorithms in EEG emotion recognition: A case study in autism. *Artificial Intelligence in Medicine, 143*, 102545. https://doi.org/10.1016/j.artmed.2023.102545

[8] Varshney, A., Ghosh, S. K., Padhy, S., Tripathy, R. K., & Acharya, U. R. (2021). Automatic categorization of mental arithmetic activities based on recurrent neural networks and entropy parameters extracted from multi-channel EEG data. *Electronics, 10*(9), 1079. https://doi.org/10.3390/electronics10091079

[9] Bao, B., Zahiri, J., Gazestani, V. H., Lopez, L., Xiao, Y., Kim, R., Wen, T. H., Chiang, A. W., Nalabolu, S., Pierce, K., & Robasky, K. (2023). A predictive ensemble classifier for the gene expression diagnosis of ASD in children aged one to four years. *Molecular Psychiatry, 28*(2), 822–833. https://doi.org/10.1038/s41380-022-01876-6

[10] Wang, S. H., Yu, Y., Kim, J., Lee, T., Park, Y. R., & Kim, H. W. (2024). Identifying children at risk for developmental disabilities using facial landmarks via a mobile application. *Psychiatry Investigation, 21*(5), 496–504. https://doi.org/10.30773/pi.2023.0226.