Improving Early ASD Diagnosis in Pediatric Populations Using Ensemble Learning Approaches

Priyanshi Mulwani [1*], Manisha Bhende[2], Swati Sharma [3]

[1,2,3] Dr. D.Y. Patil School of Science & Technology, Dr. D.Y. Patil Vidyapeeth, Pune, India

1* priyanshi.mulwani@dpu.edu.in
2 manisha.bhende@dpu.edu.in
3 swati.sharma@dpu.edu.in

Abstract. Autism spectrum disorder(ASD) is a neurological condition defined by repetitive mannerisms, communication impairments, and social interaction issues. For treatments to be effective and for children to lead easy lives, early sickness detection is crucial. Because symptoms can manifest in a variety of ways, traditional diagnostic techniques are frequently time-consuming, subjective, and limiting. This research leverages machine learning and ensemble methods to improve the early detection of autism in pediatric populations. During our work flow, a dataset including behavioral,demographic and clinical parameters was analyzed using models like Logistic Regression,Decision Tree and Naïve Bayes, achieving baseline accuracies upto 91.41%. A Voting Classifier ensemble, combining these models with hyperparameter tuning,improved accuracy to 95.95%. Advanced ensemble methods such as AdaBoost,CatBoost, XGBoost and Random Forest were further evaluated, with CatBoost and XGBoost demonstrating superior precision, recall and overall performance.These results underscore the ability of ensembling techniques in addressing the limitations of traditional ASD diagnostics, offering a scalable,accurate and efficient approach for early intervention of autism focusing on children.

Keywords: Autism Spectrum Disorder, Ensembling, Boosting

1 Introduction:

A neurological and developmental illness ,autism spectrum disorder influences behavior and how people interact with others in society. Early diagnosis during critical development years (ages 1-5) is vital as it enables well timed interventions that considerably improve outcomes for affected children. However ,diagnosing ASD at an early age is challenging due to the heterogeneity of symptoms, dependence on subjective assessments and variability in access of trained clinicians.

Using conventional methods based on behavioural observations, autism spectrum disorder (ASD) is usually diagnosed. These diagnostic techniques can be deceptive and time-consuming[1]. It affects around 1% of people worldwide, and its symptoms often appear during the formative phases following birth [2]. It is projected that 75 million individuals worldwide suffer from it, and since research on the condition started in the 1960s, the number of cases has steadily climbed [3].Advancements in Artificial

S. K. Swarnkar et al. (eds.), *Proceedings of International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICSHIT 2024)*, Advances in Intelligent Systems Research 190,
https://doi.org/10.2991/978-94-6463-704-5_13

Intelligence [4] and Machine Learning [5] [6] offer promising tools for improving ASD diagnosis. The ability of ensemble learning, which combines the predictions of multiple models, to increase accuracy, reduce variability, and handle complex data makes it stand out among the others.

By employing machine learning and deep learning models to analyse text inputs, particularly from social media platforms, the study explores the usage AI in the diagnosis of autism spectrum disorder with the goal of refining the value of life for the affected population.[7].ML models are chosen over other approaches because they often work with the relationships between different brain areas. Compared to other models, machine learning algorithms are known for their higher prediction accuracy. By investigating its genetics and developing treatments, machine learning can improve autism diagnosis techniques [8].

2 Data Exploration

This study looks at several machine learning algorithms using the dataset that was gathered from Kaggle in order to identify ASD in children at an early age. As seen in Fig. 1 through Fig.4, data is analysed [9] from a variety of angles. There are total 2226 records of children with various parameters as listed below in Table 1.

Table 1. Parameter Explanation

Parameter	Description	Type
A1 to A10	Response depending on the screening technique	Binary
Age_Years	Mention age of child	Number
Sex	Specifying gender	String
Ethnicity	Mention ethnicity of child	String
Jaundice	Whether child had jaundice?	Boolean
Family_member _with_ASD	Does the close family member have ASD?	Boolean

Who_completed_the_test	Who is conducting test (Family Member/Health care professional)	String
ASD_traits	Has ASD behaviour	Boolean

Based on the total number of records in the dataset, Fig. 1 shows the distribution of ASD cases among the collected records. ASD affects 61.6% of the population, according to observations.

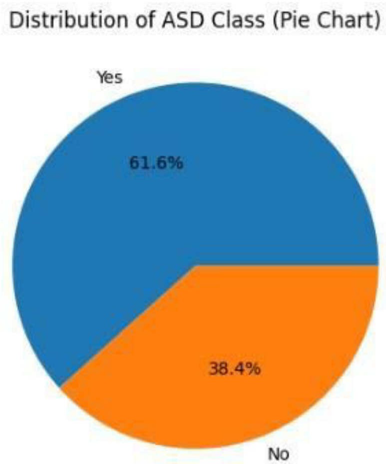

Fig. 1. Distribution of ASD cases

Fig. 2 focuses on distribution of ASD cases on basis of ethinicity in the dataset.

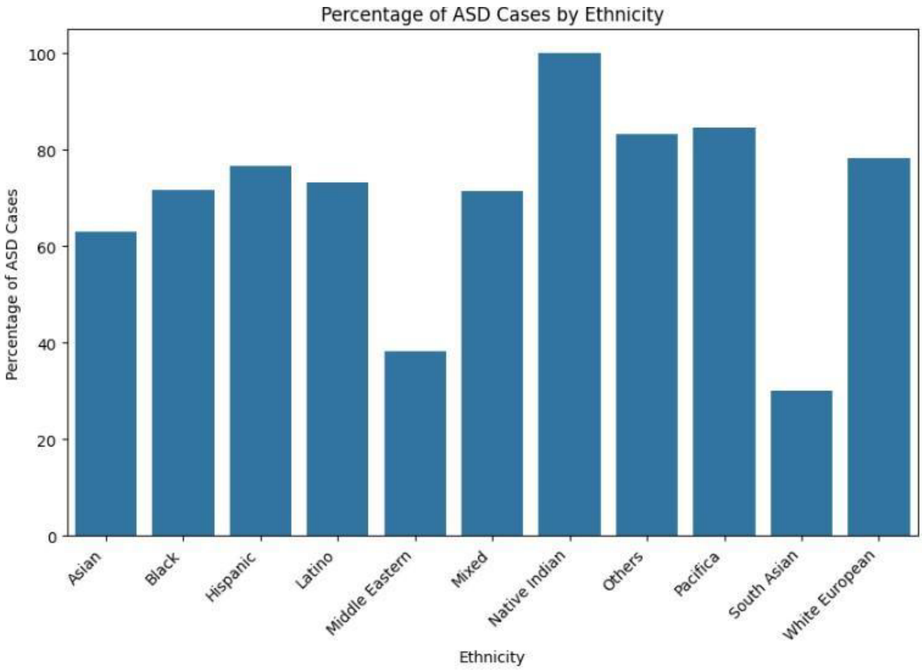

Fig. 2. Percentage of ASD cases based on Ethnicity

The proportion of ASD patients with jaundice is examined in Fig. 3. According to the study, neonatal jaundice may increase a child's risk of having ASD and may be associated with the disorder [10] [11].

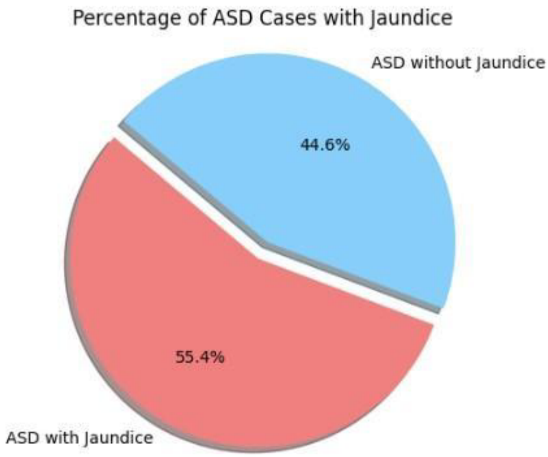

Fig. 3. ASD cases with Jaundice

The percentage distribution of ASD patients by age and gender is displayed in Fig. 4. The age group of one to three years old is found to make up the majority of the population.

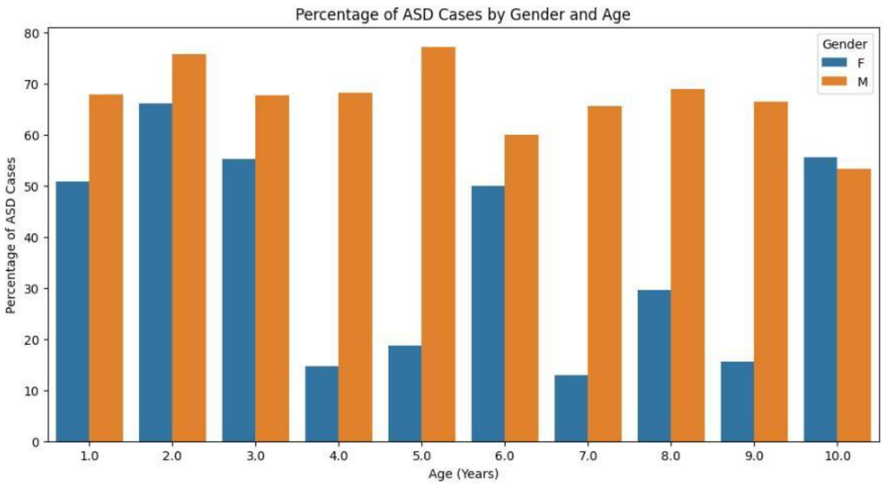

Fig. 4. Gender and Age distribution of ASD cases

3 Methodology:

In statistical modelling and machine learning, assembling methods [12] are ways to integrate many models to enhance robustness, decrease overfitting, and improve prediction performance in comparison to a single model.Several ensemblers are employed in this research work to take advantages of several models in order to improve generalisation. ASD diagnosis using the suggested methods is shown in Fig. 5. Preparing data for analysis or modelling requires both data loading and basic preparation by cleaning and feature extraction process. Data consistency is ensured by converting columns to the appropriate data type (such as numeric) and managing missing values. To facilitate testing and training, data is divided into an 80:20 ratio.

Subsequently, an ensemble of machine learning models, comprising XGBoost, AdaBoost, Random Forest, CatBoost and Voting Classifier, is employed to independently make predictions. All these models are evaluated and compared as illustrated in later sections. This pipeline provides a comprehensive approach to ASD diagnosis, leveraging the power of ensemble learning to mitigate overfitting and improve generalization.

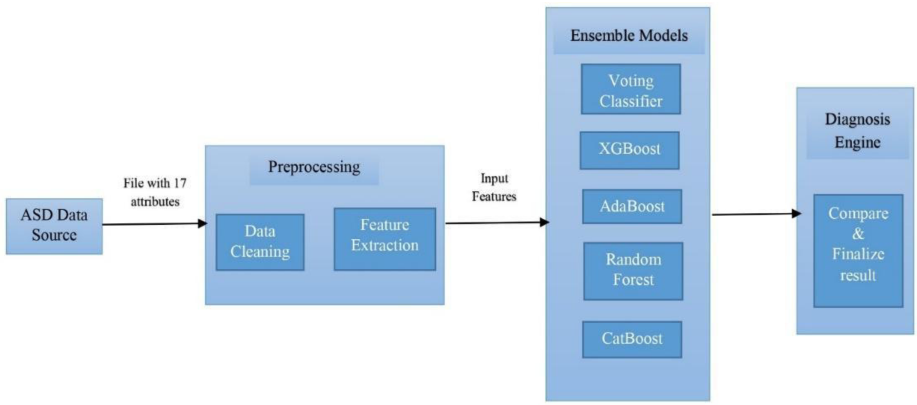

Fig. 5. Proposed Methodology

4 Results & Discussion:

4.1 Voting Classifier

Early diagnosis of ASD in children is accomplished by a variety of machine learning approaches [13] [14]. First, we used a range of machine learning algorithms like Decision Tree, Logistic Regression, and Naïve Bayes, to assess and predict the experiment data. These algorithms are used to assess autism spectrum disorder and make predictions from both qualitative and quantitative perspectives. The accuracy of the models was noted as depicted in Table 2.

Table 2. Machine Learning Models with Accuracy

Model	Accuracy
Logistic Regression	89.67 %
Decision Tree	91.41 %
Naïve Bayes	63.02 %

Additionally, the Naive Bayes model's hyperparameters were adjusted using `GridSearchCV`. For optimal results, the 5-fold cross-validation approach was used to decide the optimal value of the `var_smoothing` parameter, which came out to be 0.1.Best Cross-Validation Score is 0.8459. Thus, Naive Bayes Accuracy was 86.52 % (hypertuned). Using these base models, an ensembler was generated for further optimization. It combines the predictions of the Logistic Regression, Decision Tree and the tuned Naive Bayes classifier using a Voting Classifier with `soft` voting as illustrated in Fig. 6.

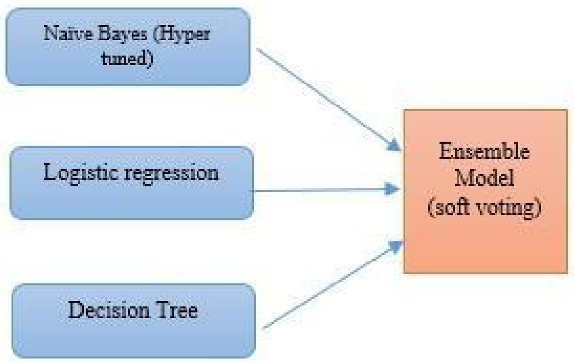

Fig. 6. Ensemble Model

Fig. 7 shows the Confusion Matrix for Voting Classifier Ensemble model which includes hypertuned Naive Bayes , Logistic Regression and Decision Tree model .

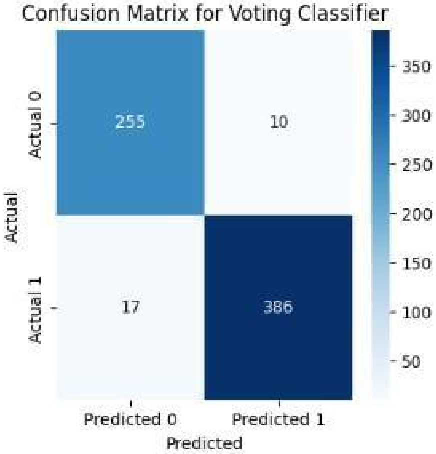

Fig. 7. Confusion Matrix : Ensemble Model (Voting Classifier)

The accuracy of this ensembled classifier was achieved as 95.95 %.

4.2 Adaboost:

In order to produce a powerful classifier, the ensemble learning method AdaBoost (Adaptive Boosting) [15] combines several weak learners, usually decision trees. By iteratively modifying the weights of incorrectly categorised samples, it enables later models to concentrate on more challenging cases. AdaBoost is renowned for increasing accuracy by combining basic base learners with weak models to get a weighted final prediction. Although it might be sensitive to outliers and noisy data, it works well for classification tasks. It may not work as well on complicated datasets as more sophisticated boosting techniques like XGBoost or LightGBM, even though it is less likely to overfit than certain algorithms.The Confusion Matrix appears in Fig. 8 when Adaboost is used on the dataset we have provided.

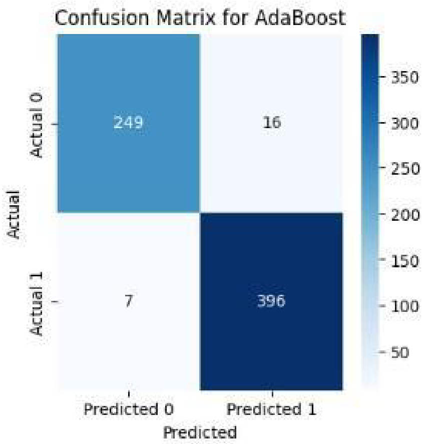

Fig. 8. Confusion Matrix : AdaBoost

4.3 CatBoost:

One gradient boosting technique that is believed to handle categorical data well and with little preprocessing is CatBoost (Categorical Boosting) [16].It uses techniques like ordered boosting and target-based encodings to reduce overfitting and improve accuracy. Known for its ease of use, high performance, and GPU support, CatBoost handles missing and noisy data robustly while requiring little parameter tuning. After applying Catboost algorithm on preprocessed dataset, Confusion Matrix is as shown in Fig. 9.

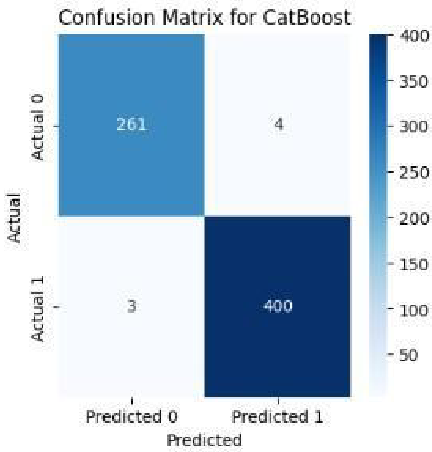

Fig. 9. Confusion Matrix : CatBoost

4.4 XGBoost:

A high-performance machine learning technique based on gradient boosting, XGBoost (Extreme Gradient Boosting) [17] is intended for classification and regression applications. By minimising errors from earlier models and employing strategies like regularisation (L1/L2) to lessen overfitting, it creates models in a sequential fashion, optimising performance. XGBoost [18], which is well-known for its effectiveness, scalability, and capacity to manage big datasets, facilitates parallelised computing, manages missing data well, and offers insights into the significance of features. Despite its great predicted accuracy, it may be computationally demanding and necessitates precise parameter adjustment.The confusion matrix that results from applying the XGBoost algorithm to the provided dataset (after cleaning) is displayed in Fig. 10.

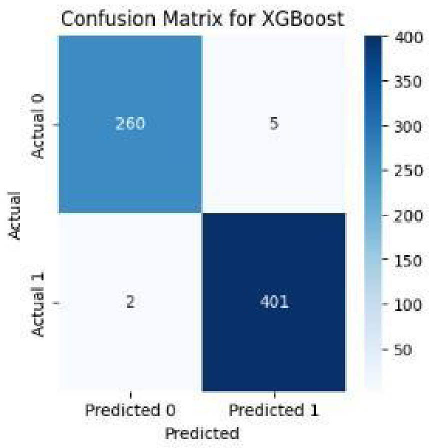

Fig. 10. Confusion Matrix : XGBoost

4.5 Random Forest:

The Random Forest algorithm [12] is an ensemble approach for classification and regression that uses random selections of data and characteristics to generate several decision trees. The predictions of these decision trees are then combined by majority voting or averaging. To lessen overfitting and enhance generalisation, it makes use of feature randomisation and bagging (bootstrap sampling). Random Forest is commonly utilised in fields like medical diagnostics because of its high accuracy, resilience to missing data, and capacity to convey feature significance. The Confusion Matrix is displayed in Fig. 11 after being applied to the provided dataset with the supplied settings.

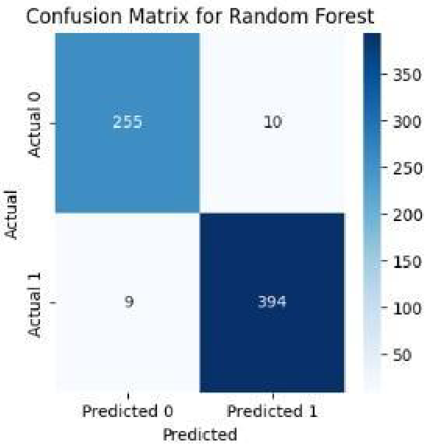

Fig. 11. Confusion Matrix : Random Forest Algorithm

	Model	Accuracy	AUC Score	Precision	Recall	F1-Score
0	Voting Classifier	0.959581	0.995562	0.974747	0.957816	0.966208
1	AdaBoost	0.965569	0.992546	0.961165	0.982630	0.971779
2	CatBoost	0.989521	0.999588	0.990099	0.992556	0.991326
3	XGBoost	0.989521	0.999157	0.987685	0.995037	0.991347
4	Random Forest	0.971557	0.997018	0.975248	0.977667	0.976456

Fig. 12. Evaluation of all models

Fig.12 and Fig.13 shows the comparison of these ensembling methods based on Precision, Recall, F1 Score , Accuracy and ROC Curves. CatBoost and XGBoost appear to be the top-performing models with accuracy score of 98.95 %. They also have high precision and recall scores, indicating that they are both good at recognizing true positive cases and avoiding false positives.

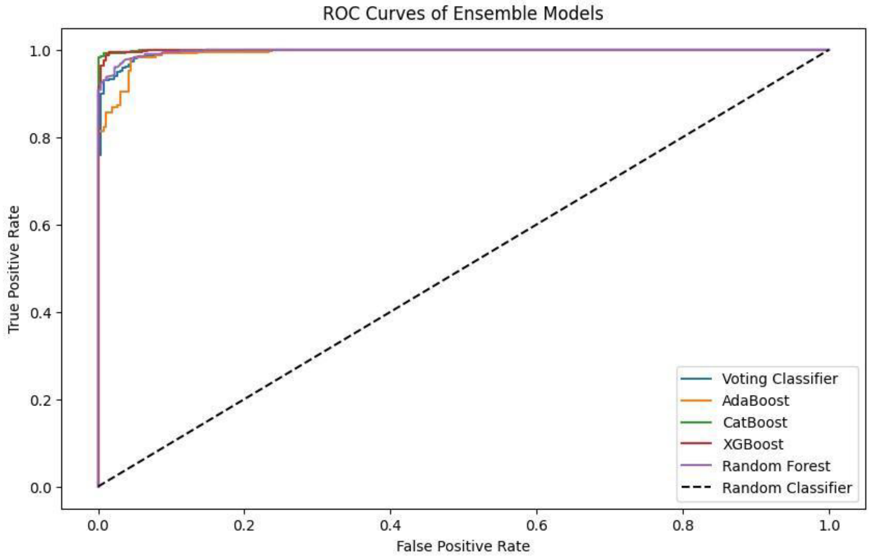

Fig. 13. ROC Curves of Ensemble Models

5 Conclusion:

This study demonstrates the potential of ensemble techniques, in improving the early diagnosis of Autism Spectrum Disorder in pediatric populations. By analyzing a dataset of behavioral, demographic, and clinical parameters, baseline models such as Logistic Regression, Decision Tree, and Naïve Bayes achieved promising results, with accuracies up to 91.41%. Further enhancement through a Voting Classifier increased accuracy to 95.95%, while advanced ensemble models like CatBoost and XGBoost achieved superior performance, attaining an accuracy of 98.95% with high precision and recall. These outcomes highlight the efficacy of ensemble methods in addressing limitations of traditional diagnostic approaches, including subjectivity, time consumption, and variability. In particular, CatBoost and XGBoost proved to be robust in handling complex datasets, making them ideal for real-world applications in ASD diagnosis. The results of this research pave the way for scalable and automated ASD diagnostic tools that can support early intervention during critical developmental years. Future work could explore integrating these models into clinical workflows, validating them on larger and more diverse datasets, and incorporating multimodal data sources such as speech, eye-tracking, or genetic markers to further improve diagnostic precision.

In conclusion, the adoption of ensemble techniques offers a transformative approach to ASD diagnostics, promising a more accurate, efficient, and accessible solution for early intervention and improved outcomes in children with ASD.

References:

1. Simeoli R, Rega A, Cerasuolo M, Nappo R, Marocco D. Using Machine Learning for Motion Analysis to Early Detect Autism Spectrum Disorder: A Systematic Review. Rev J Autism Dev Disord [Internet]. 2024 Mar 2 [cited 2024 Nov 16]; Available from: https://doi.org/10.1007/s40489-024-00435-4
2. Vakadkar K, Purkayastha D, Krishnan D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN COMPUT SCI. 2021 Jul 22;2(5):386.
3. Prasad R, Musa F, Ahmad HM, Upadhyay SK, Sharma BK. Autism Spectrum Disorder Prediction Using Machine Learning and Design Science. International Journal of Experimental Research and Review. 2024 May 30;39(Spl Volume):213–28.
4. Shahamiri SR, Thabtah F. Autism AI: a New Autism Screening System Based on Artificial Intelligence. Cognitive Computation. 2020 Jul 1;12.
5. Rasul RA, Saha P, Bala D, Karim SMRU, Abdullah MdI, Saha B. An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder. Healthcare Analytics. 2024 Jun 1;5:100293.
6. Thapa R, Garikipati A, Ciobanu M, Singh N, Browning E, DeCurzio J, et al. Machine Learning Differentiation of Autism Spectrum Sub-Classifications. J Autism Dev Disord. 2024 Nov 1;54(11):4216–31.
7. Rubio-Martín S, García-Ordás MT, Bayón-Gutiérrez M, Prieto-Fernández N, Benítez-Andrades JA. Enhancing ASD detection accuracy: a combined approach of machine learning and deep learning models with natural language processing. Health Inf Sci Syst. 2024 Mar 6;12(1):20.
8. amala B, Mahanaga Pooja KS, Varsha S, Sivapriya K. ML Based Approach to Detect Autism Spectrum Disorder (ASD). In: 2021 4th International Conference on Computing and Communications Technologies (ICCCT) [Internet]. 2021 [cited 2024 Nov 16]. p. 313–8. Available from: https://ieeexplore.ieee.org/document/9711826
9. Afarin Bargrizan [Internet]. [cited 2024 Nov 16]. Available from: https://www.kaggle.com/code/afarinbargrizan/children-asd
10. Jenabi E, Bashirian S, Khazaei S. Association between neonatal jaundice and autism spectrum disorders among children: a meta-analysis. Clin Exp Pediatr. 2019 Nov 7;63(1):8–13.

11. Association of Autism Spectrum Disorders With Neonatal Hyperbilirubinemia [Internet]. [cited 2024 Nov 27]. Available from: https://journals.sagepub.com/doi/epub/10.1177/2333794X15596518
12. Hyde KK, Novack MN, LaHaye N, Parlett-Pelleriti C, Anden R, Dixon DR, et al. Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. Rev J Autism Dev Disord. 2019 Jun 1;6(2):128–46.
13. BalaKrishna N, Mukesh Krishnan MB, Reddy SM, Irfan SK, Sumaiya S. AUTISM Spectrum Disorder Detection Using Machine Learning. In: 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) [Internet]. 2023 [cited 2024 Nov 27]. p. 1645–50. Available from: https://ieeexplore.ieee.org/document/10183095
14. Tikaram, Ingle MB, Rakshe CT, Agastinose Ronickom JF. Advancing ASD Diagnostic Classification with Features of Continuous Wavelet Transform of fMRIand Machine Learning Algorithms. In: 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA) [Internet]. 2023 [cited 2024 Nov 27]. p. 571–7. Available from: https://ieeexplore.ieee.org/document/10346926
15. Zhang Y, Ni M, Zhang C, Liang S, Fang S, Li R, et al. Research and Application of AdaBoost Algorithm Based on SVM. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) [Internet]. 2019 [cited 2024 Nov 29]. p. 662–6. Available from: https://ieeexplore.ieee.org/document/8785556
16. Yulistiani R, Kurniadi FI. Catboost in Fetal Health Classification. In: 2024 International Conference on ICT for Smart Society (ICISS) [Internet]. 2024 [cited 2024 Nov 29]. p. 1–4. Available from: https://ieeexplore.ieee.org/document/10751585
17. Hasan M, Ahamad MdM, Aktar S, Moni MA. Early Stage Autism Spectrum Disorder Detection of Adults and Toddlers Using Machine Learning Models. In: 2021 5th International Conference on Electrical Information and Communication Technology (EICT) [Internet]. 2021 [cited 2024 Nov 29]. p. 1–6. Available from: https://ieeexplore.ieee.org/document/9733664
18. Gupta V, Manoj G, Bhattacharya A, Singh Sengar S, Mishra R, Kar B, et al. A Framework to Diagnose Autism Spectrum Disorder Using Morphological Connectivity of sMRI and XGBoost. In: Studies in health technology and informatics. 2023.