IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 6, March 2024

An Early-Stage Autism Spectrum Detection System

Prof. D. C. Pardeshi[1], Parth S. Mishra[2], Shripad T. Kulkarni[3]
Professor, Department of AI & ML[1]
Students, Department of AI & ML[2,3]
AISSMS Polytechnic, Pune, India

Abstract: *The Early-Stage Autism Detection System presents a breakthrough approach to identifying Autism Spectrum Disorder (ASD) in its initial stages, particularly focusing on early childhood diagnosis. Leveraging machine learning (ML) techniques such as Random Forest and Support Vector Machines, the system meticulously analyses behavioural patterns and social interactions to pinpoint potential indicators of ASD, even in toddlers. It adeptly tackles challenges like imbalanced class distributions by employing random oversampling and adopts feature scaling and selection methods to heighten prediction accuracy. Through extensive experimentation on diverse ASD datasets, the system discerns crucial features pivotal for precise diagnosis. Its implementation promises timely intervention and improved outcomes by enabling the early detection and support of individuals with ASD from the outset of development. This system represents a paradigm shift in ASD diagnosis, offering a more efficient and effective means of identifying and assisting individuals with ASD at the earliest possible stage, thereby potentially mitigating the impact of the disorder and enhancing quality of life.*

Keywords: Autism spectrum disorder, machine learning, classification, feature scaling, feature selection technique.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) poses significant challenges in early identification and intervention due to its varied symptoms and impacts on social interactions and behaviors. Despite lacking a sustainable solution, early intervention and proper medical care can significantly improve a child's development, focusing on enhancing communication skills and behavioural patterns.Traditional methods of ASD diagnosis rely on behavioural science and are often complex and difficult. Typically, ASD is diagnosed around the age of two, although it can be identified later based on severity. Various treatment strategies exist, but they aren't universally applied until ASD development becomes severe.In recent years, researchers have turned to machine learning (ML) approaches to aid in ASD diagnosis.

These studies have utilized a range of ML algorithms, including rule-based ML techniques, Random Forest (RF), Support Vector Machines (SVM), Decision Trees (DT), and others. By analyzing ASD attributes, researchers have developed predictive models for different age groups, from toddlers to adults.The integration of cognitive computing has enabled the identification of significant features for ASD diagnosis.

Furthermore, studies have examined the predictive performance of ML models such as Deep Neural Networks (DNN), ensemble ML approaches, and bio-inspired algorithms. These efforts have resulted in the development of smartphone applications and programming interfaces for ASD diagnosis across all age groups.

To enhance prediction accuracy, researchers have addressed imbalanced class distributions in ASD datasets and investigated various feature scaling methods. Additionally, feature selection techniques have been employed to identify the most important risk factors for ASD prediction.One notable study has compared different ML models and feature selection methods to identify the most effective approaches for ASD detection. By conducting extensive experiments on standard ASD datasets, researchers have aimed to improve the accuracy and reliability of ASD diagnosis.

In summary, the application of ML techniques in ASD diagnosis shows promise for early-stage detection and intervention. By leveraging diverse datasets and advanced algorithms, researchers aim to enhance our understanding of

DOI: 10.48175/IJARSCT-16823

ASD and provide better support for affected individuals across the lifespan. Ongoing research in this area continues to refine ML models and improve their effectiveness in ASD diagnosis and management

II. MATERIALS & METHODS.

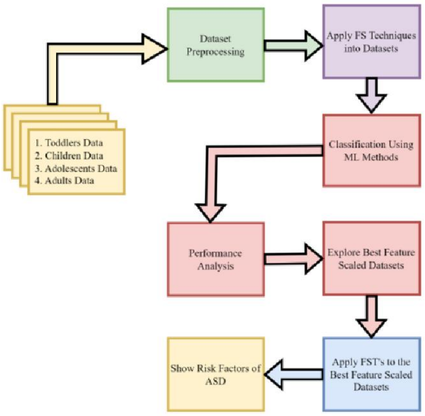
A. DATASET DESCRIPTION

We collect the four ASD datasets (Toddlers, Adolescents, Children, and Adults) from the publicly available repositories: Kaggle and UCI ML [36], [37], [38], [39]. The authors in [13] created the ASD Tests smartphone app for Toddlers, Children, Adolescents, and Adults ASD screening using QCHAT-10 and AQ-10. The application computes a score of 0 to 10 for every individual, with which the final score is 6 out of 10 which indicates an individual has positive ASD.

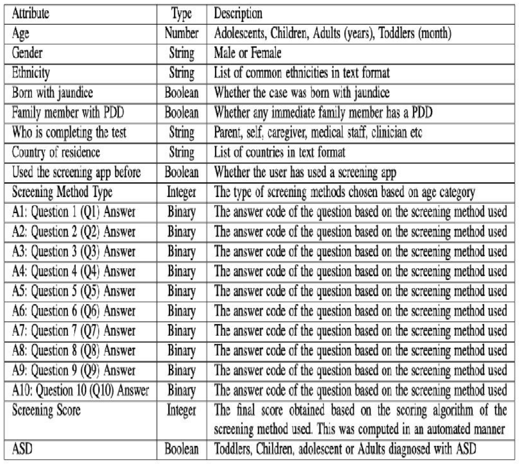
In addition, ASD data is obtained from the ASD Tests app while open-source databases are developed in order to facilitate research in this area. The detailed description of the Toddlers, Children, Adolescents, and Adults ASD datasets are given in Table 1 and Table 2.

B. METHOD OVERVIEW

This research aims to create an effective prediction model using different types of ML methods to detect autism in people of different ages. First of all, the datasets are collected, and then the preprocessing is accomplished via the missing values imputation, feature encoding, and oversampling.


1. Toddlers Data
2. Children Data
3. Adolescents Data
4. Adults Data
Dataset Preprocessing
Apply FS Techniques into Datasets
Classification Using ML Methods
Performance Analysis
Explore Best Feature Scaled Datasets
Apply FST's to the Best Feature Scaled Datasets
Show Risk Factors of ASD

The Mean Value Imputation (MVI) method is used to impute the missing values of the dataset. Then, the categorical feature values are converted to their equivalent numerical values using the One Hot Encoding (OHE) technique.



Attribute	Type	Description
Age	Number	Adolescents, Children, Adults (years), Toddlers (month)
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean	Whether the case was born with jaundice
Family member with PDD	Boolean	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician etc
Country of residence	String	List of countries in text format
Used the screening app before	Boolean	Whether the user has used a screening app
Screening Method Type	Integer	The type of screening methods chosen based on age category
A1: Question 1 (Q1) Answer	Binary	The answer code of the question based on the screening method used
A2: Question 2 (Q2) Answer	Binary	The answer code of the question based on the screening method used
A3: Question 3 (Q3) Answer	Binary	The answer code of the question based on the screening method used
A4: Question 4 (Q4) Answer	Binary	The answer code of the question based on the screening method used
A5: Question 5 (Q5) Answer	Binary	The answer code of the question based on the screening method used
A6: Question 6 (Q6) Answer	Binary	The answer code of the question based on the screening method used
A7: Question 7 (Q7) Answer	Binary	The answer code of the question based on the screening method used
A8: Question 8 (Q8) Answer	Binary	The answer code of the question based on the screening method used
A9: Question 9 (Q9) Answer	Binary	The answer code of the question based on the screening method used
A10: Question 10 (Q10) Answer	Binary	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
ASD	Boolean	Toddlers, Children, adolescent or Adults diagnosed with ASD

TABLE 1. Feature description of the ASD datasets.

DOI: 10.48175/IJARSCT-16823

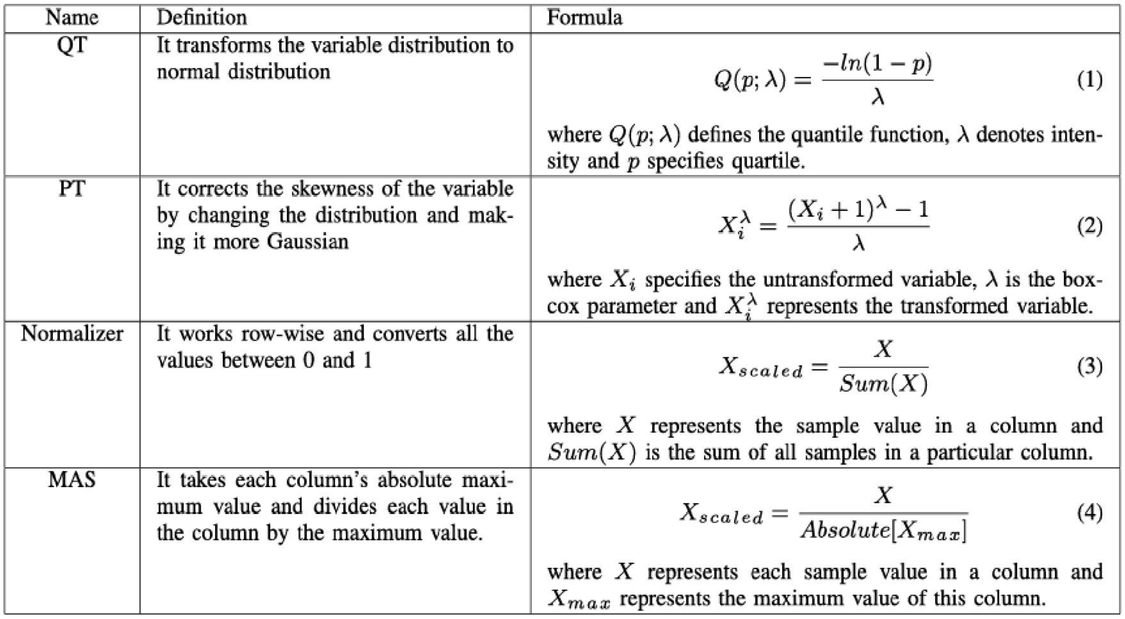
Table 1 shows that all four datasets used in this work have an imbalanced class distribution problem. As such, a Random Over Sampler strategy is used to alleviate this issue. After completing the initial preprocessing, the datasets feature values are scaled using four different FS techniques i.e., QT, PT, Normalizer, and MAS (see their detailed operations in Table 3). The feature-scaled datasets are then classified using eight different ML classification techniques i.e., AB, RF, DT, KNN, GNB, LR, SVM, and LDA. Comparing the classification outcomes of the classifiers on different feature-scaled ASD datasets, the best-performing classification methods, and the best FS techniques for each ASD dataset are identified. After those analyses, the ASD risk factors are calculated, and the most important attributes are ranked according to their importance values using four different FSTs i.e., IGAE, GRAE, RFAE, and CAE. To this end, Fig. 1 represents the proposed research pipeline to analyse the ASD datasets and calculate the risk factors that are most responsible for ASD detection.

C. MACHINE LEARNING METHODS:

1) ADA BOOST (AB)

AB is a tree-based ensemble classifier that incorporates many weak classifiers to reduce misclassification errors

TABLE 2. Detailed description of the different FS methods



Name	Definition	Formula
QT	It transforms the variable distribution to normal distribution	$$Q(p;\lambda) = \frac{-ln(1-p)}{\lambda} \quad (1)$$ where $Q(p;\lambda)$ defines the quantile function, λ denotes intensity and p specifies quartile.
PT	It corrects the skewness of the variable by changing the distribution and making it more Gaussian	$$X_i^{\lambda} = \frac{(X_i+1)^{\lambda}-1}{\lambda} \quad (2)$$ where X_i specifies the untransformed variable, λ is the box-cox parameter and X_i^{λ} represents the transformed variable.
Normalizer	It works row-wise and converts all the values between 0 and 1	$$X_{scaled} = \frac{X}{Sum(X)} \quad (3)$$ where X represents the sample value in a column and $Sum(X)$ is the sum of all samples in a particular column.
MAS	It takes each column's absolute maximum value and divides each value in the column by the maximum value.	$$X_{scaled} = \frac{X}{Absolute[X_{max}]} \quad (4)$$ where X represents each sample value in a column and X_{max} represents the maximum value of this column.

TABLE 3. Detailed description of the different FST methods.

Name	Definition	Formula
IGAE	It measures the value of information gain for each attribute concerning the output variable.	$$IG(C,A) = H(C) - H(C\|A) \quad (5)$$ where H is the information entropy, IG is the information gain, C represents the class and A represents the attribute.
GRAE	It computes the gain ratio value for each attribute with respect to the class variable.	$$GR(C,A) = \frac{H(C) - H(C\|A)}{H(A)} \quad (6)$$ where GR is the gain ratio.
RFAE	It calculates a feature's worthiness by repeatedly sampling an instance and considering the given attribute's value for the closest instance of the same and different classes.	$$R_x = P(DiffX\|DiffClass) - P(DiffX\|SameClass) \quad (7)$$ where $P(DiffX\|DiffClass)$ represents the conditional probability value for the different class and $P(DiffX\|SameClass)$ specifies the conditional probability for the same class.
CAE	It evaluates the worth of a feature by calculating Pearson's correlation value for the class variable.	$$\rho(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \quad (8)$$ where $COV(X,Y)$ represents the covariance, and σ_X and σ_Y denote the standard deviation of X and Y respectively. $\rho(X,Y)$ represents the correlation between X and Y.

2) RANDOM FOREST (RF)

RF is a decision tree-based ensemble classification method and follows the split and conquer technique in the input dataset to create multiple decision-making trees (known as the forest)[42]. It works in two phases. At first, it creates a forest by combining the 'N' number of decision trees and in the second phase, it makes predictions for each tree generated in the first phase. The working process of the RF algorithm is illustrated below:

Select random samples from the training dataset.

Construct decision trees for each training sample.

Select the value of N to define the number of decision trees.

Repeat Steps 1 and 2.

For each test sample, find the predictions of each decision tree, and assign the test sample a class value based on majority voting.

3) DECISION TREE (DT)

DT follows a top-down approach to build a predictive model for class values using training data-inducing decision-making rules[43].This research utilized the information gain method to select the best attribute. Assuming Pi, the probability such that $xi \in D$, exists to a class Ci ,and is predicted by$|Ci\ D|/|D|$. To classify instances in the dataset D, the required information is needed, and the following equation calculates it:

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

Where $Info(D)$ is the average amount of information needed to identify Ci of an instance, $xi\epsilon D$ and the objective of DT is to divide repeatedly, D , into sub data sets D1, D2, ……., Dn.

Finally, the following equation calculates the information gain value:

$$Gain(A) = Info(D) - Info_A(D)$$

4) GAUSSIAN NAIVE BAYES (GNB)

GNB algorithm follows a normal distribution and is used for classification when all the data values of a dataset are numeric [43]. To compute the probability values of any instance with respect to the class value mean and standard deviation are calculated for each attribute of the dataset. Consequently, for testing, when any instance comes, it utilizes the mean and standard deviation values to calculate the probability of the test instance. The necessary equations are given below:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\delta = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$f(x) = \frac{1}{\sqrt{2\pi}} * \frac{1}{\delta} * e^{-(x-\mu)2}$$

Where μ indicates the mean, δ represents standard deviation, xi denotes all samples in a particular column, n indicates the total number of samples and fx presents the conditional probability of class value.

5) K-NEAREST NEIGHBORS (KNN)

KNN classifies the test data by utilizing the training data directly by calculating the K value, indicating the number of KNN[43]. For each instance, it computes the distance between all the training instances and sorts the distance. Furthermore, a majority voting technique is employed to assign the final class label to the test data. This research applies Euclidean distance to calculate the distances among instances. The following equation represents the Euclidean distance calculation:

$$D_e = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

Where De indicates the Euclidean distance, X_i denotes the testing sample values, Yi specifies the training sample values and n represents the total number of sample values.

6) SUPPORT VECTOR MACHINE (SVM)

SVM is used to classify both linear and non-linear data and mostly works well for high-dimensional data with non-linear mapping. It explores the decision boundary or optimal hyperplane to separate one class from another. This study used Radial Basis Function (RBF) as a kernel function and SVM automatically defines centers, weights, and thresholds and reduces an upper bound of expected test error [29],[44]. The following equation represents the RBF function:

$$K(x, x') = exp(-\frac{(||x - x'||)^2}{2\delta^2})$$

III. RESULTS AND REVIEWS

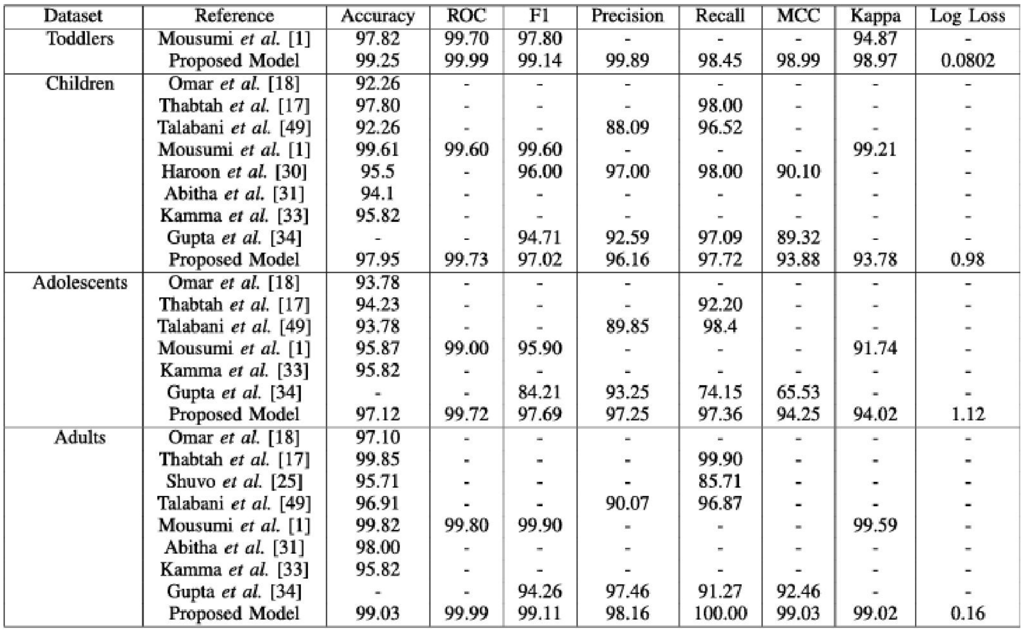
We analysed four different ASD datasets to build prediction models for different stages of people. In order to do this, we applied various FS methods to those ASD datasets and classified them utilizing eight different simple but effective ML classifiers and also determined how the FS methods affect the classification performance. Furthermore, we also employed four different FSTs to compute the importance of the features which are more responsible for ASD prediction. Inspecting the experimental findings, the best performing classifiers model predicted ASD with AB (99.25%), AB (97.95%), LDA (97.12%), LDA (99.03%) accuracy; AB, LR (99.99%), GNB

IV. DISCUSSION AND EXTENDED COMPARISON

In the previous section, we analysed four different ASD datasets to build prediction models for different stages of people. In order to do this, we applied various FS methods to those ASD datasets and classified them utilizing eight different simple but effective ML classifiers and also determined how the FS methods affect the classification performance. Furthermore, we also employed four different FSTs to compute the importance of the features which are

ISSN 2581-9429 IJARSCT

more responsible for ASDprediction. Inspecting the experimental findings, the best performing classifiers model predicted ASD with AB(99.25%), AB (97.95%), LDA (97.12%), LDA (99.03%) accuracy; AB, LR (99.99%), GNB (99.73%), AB, LDA (99.72%), LDA(99.99%)ROC;AB(99.14%),AB(97.02%), AB (97.69%), LDA (99.11%) F1-score; AB (99.95%), LR (96.16%), DT (97.25%), SVM (98.16%) precision; AB 15054 (98.45%), LR (97.72%), AB (97.36%), RF, DT, KNN, LR (100%) recall; LR, LDA (99.31%), AB (93.88%), LDA (94.25%), LR, LDA (99.03%) MCC; LR, LDA(99.31%),AB(93.78%),LDA(94.02%),LR,LDA(99.02%)kappa;AB (0.0802%), AB (0.98%), LDA (1.12%), LR, LDA (0.16%) log loss for Toddlers, Children, Adolescents, Adults datasets respectively. After analysing the experimental outcomes of different classifiers on feature-scaled ASD datasets, it is found that AB for Toddlers and Children, and LDA for Adolescents and Adults outperformed the other ML classifiers in terms of classification performance. Besides, the experimental outcomes implied that the normalizer FS method for Toddlers, normalizer FS method for Children, QT FS method for Adolescents, and QT FS method for Adults showed better performance. Additionally, we calculated the feature importance using the IGAE,GRAE,RFAE,andCAEFSTmethods on the normalizer-scaled Toddlers, normalizer-scaled Children, QT-scaled Adolescents, and QT-scaled Adults to enumeratethe risk factors for ASD prediction



Dataset	Reference	Accuracy	ROC	F1	Precision	Recall	MCC	Kappa	Log Loss
Toddlers	Mousumi *et al.* [1]	97.82	99.70	97.80	-	-	-	94.87	-
	Proposed Model	99.25	99.99	99.14	99.89	98.45	98.99	98.97	0.0802
Children	Omar *et al.* [18]	92.26	-	-	-	-	-	-	-
	Thabtah *et al.* [17]	97.80	-	-	-	98.00	-	-	-
	Talabani *et al.* [49]	92.26	-	-	88.09	96.52	-	-	-
	Mousumi *et al.* [1]	99.61	99.60	99.60	-	-	-	99.21	-
	Haroon *et al.* [30]	95.5	-	96.00	97.00	98.00	90.10	-	-
	Abitha *et al.* [31]	94.1	-	-	-	-	-	-	-
	Kamma *et al.* [33]	95.82	-	-	-	-	-	-	-
	Gupta *et al.* [34]	-	-	94.71	92.59	97.09	89.32	-	-
	Proposed Model	97.95	99.73	97.02	96.16	97.72	93.88	93.78	0.98
Adolescents	Omar *et al.* [18]	93.78	-	-	-	-	-	-	-
	Thabtah *et al.* [17]	94.23	-	-	-	92.20	-	-	-
	Talabani *et al.* [49]	93.78	-	-	89.85	98.4	-	-	-
	Mousumi *et al.* [1]	95.87	99.00	95.90	-	-	-	91.74	-
	Kamma *et al.* [33]	95.82	-	-	-	-	-	-	-
	Gupta *et al.* [34]	-	-	84.21	93.25	74.15	65.53	-	-
	Proposed Model	97.12	99.72	97.69	97.25	97.36	94.25	94.02	1.12
Adults	Omar *et al.* [18]	97.10	-	-	-	-	-	-	-
	Thabtah *et al.* [17]	99.85	-	-	-	99.90	-	-	-
	Shuvo *et al.* [25]	95.71	-	-	-	85.71	-	-	-
	Talabani *et al.* [49]	96.91	-	-	90.07	96.87	-	-	-
	Mousumi *et al.* [1]	99.82	99.80	99.90	-	-	-	99.59	-
	Abitha *et al.* [31]	98.00	-	-	-	-	-	-	-
	Kamma *et al.* [33]	95.82	-	-	-	-	-	-	-
	Gupta *et al.* [34]	-	-	94.26	97.46	91.27	92.46	-	-
	Proposed Model	99.03	99.99	99.11	98.16	100.00	99.03	99.02	0.16

TABLE 4.Comparison with other works.

V. CONCLUSION

In this study, we developed a machine-learning framework to detect AutismSpectrum Disorder (ASD) across different age groups (Toddlers, Children, Adolescents, and Adults) using predictive models based on various ML techniques. Following initial data processing, ASD datasets underwent scaling using four feature scaling techniques (QT, PT, Normalizer, MAS) and classification using eight ML classifiers (AB, RF, DT, KNN, GNB, LR, SVM, LDA). We evaluated classification performance using accuracy, ROC, F1-Score, precision, recall, MCC, kappa score, and Log loss metrics to determine optimal FS and classification methods. Our ML-based prediction models offer accurate ASD identification, aiding physicians in diagnosis. We also analyzed feature importance using IGAE, GRAE, RFAE, and CAE, identifying key ASD prediction features. While data limitations hindered a fully generalized model, future work will gather more diverse ASD data to enhancedetection across all ages and neuro-developmental disorders.

REFERENCES

[1]. M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, Efficient machine learning models for early stage detection of autism spectrum disorder, Algorithms, vol. 15, no. 5, p. 166, May 2022.

[2]. D. Pietrucci, A. Teofani, M. Milanesi, B. Fosso, L. Putignani, F. Messina, G. Pesole, A. Desideri, and G. Chillemi, Machine learning data analysis highlights the role of parasutterella and alloprevotella in autism spectrum disorders, Biomedicines, vol. 10, no. 8, p. 2028, Aug. 2022.

[3]. R. Sreedasyam, A. Rao, N. Sachidanandan, N. Sampath, and S. K. Vasudevan, Aarya A kinesthetic companion for children with autism spectrum disorder, J. Intell. Fuzzy Syst., vol. 32, no. 4, pp. 29712976, Mar. 2017.

[4]. J. Amudha and H. Nandakumar, A fuzzy based eye gaze point estimation approach to study the task behavior in autism spec trum disorder, J. Intell. Fuzzy Syst., vol. 35, no. 2, pp. 14591469, Aug. 2018.

[5]. H. Chahkandi Nejad, O.Khayat, and J.Razjouyan, Software development of anintelligent spirography test system for neurological disorder detection and quantification, J. Intell. Fuzzy Syst., vol. 28, no. 5, pp. 21492157, Jun. 2015.

[6]. F. Z. Subah, K.Deb, P. K. Dhar, and T. Koshiba, A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI, Appl. Sci., vol. 11, no. 8, p. 3636, Apr. 2021.

[7]. [7] 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 systematic review, Electronics, vol. 10, no. 23, p. 2982, Nov. 2021.

[8]. I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. H. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani, Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques, Electronics, vol. 11, no. 4, p. 530, Feb. 2022.

[9]. P. Sukumaran and K. Govardhanan,Towards voice based prediction and analysis of emotions in ASD children, J. Intell. Fuzzy Syst., vol. 41, no. 5, pp. 53175326, 2021

[10]. S. P. Abirami, G. Kousalya, and R. Karthick, Identification and exploration of facial expression in children with ASD in a contact less environment, J. Intell. Fuzzy Syst., vol. 36, no. 3, pp. 20332042, Mar. 2019.

[11]. M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, Detecting autism spectrum disorder using machine learning techniques, Health Inf. Sci. Syst., vol. 9, no. 1, pp. 1 13, Dec. 2021.

[12]. C. Allison, B. Auyeung, and S. Baron-Cohen, Toward brief red flags for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls, J. Amer. Acad. Child Adolescent Psychiatry, vol. 51, no. 2, pp. 202212, 2012.

[13]. F. Thabtah, F. Kamalov, and K. Rajab, A new computational intelligence approach to detect autistic features for autism screening, Int. J. Med. Inform., vol. 117, pp. 112124, Sep. 2018.

[14]. E. Dritsas and M. Trigka, Stroke risk prediction with machine learning techniques, Sensors, vol. 22, no. 13, p. 4670, Jun. 2022

DOI: 10.48175/IJARSCT-16823