**Analysis And Detection Of Autism Spectrum Disorder(ASD) Using Machine Learning**

\*Shobha Rani.M1,a) , Karim Kadapa2,b), Jogi Om Sai Santosh3,c), Alluri Venkata Sai4,d), Dheeraj.M5,e)

1,2,3,4,5 *School of Computer Science and Engineering, Reva University, Bengaluru, India–560064.*

*\*a)mshobha.rani@reva.edu.in,b)kadapakareem1423@gmail.com,c)omsaisantosh789@gmail.com,*

*d)allurivenkatasai02@gmail.com,e)tarak140202@gmail.com*

**Abstract.** The primary aim of this study is to address the challenges associated with the lack of comprehensive medical tests for Autism Spectrum Disorder (ASD) and to promote early intervention in healthcare. This paper introduces a novel method for identifying ASD using advanced machine learning algorithms. The proposed model leverages multiple classifiers, including Logistic Regression, K-Neighbors Classifier, Gaussian Naive Bayes, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest and XGBoost. The approach includes the development of a flexible web interface designed for use by both medical professionals and the general public, allowing for the input of relevant data to predict the probability of ASD. To enhance model performance, advanced techniques such as Grid Search Cross-Validation (CV) and Pipeline methods were employed during the refinement phase. The application of these techniques resulted in a significant improvement in the accuracy of the Random Forest model, increasing from 64% to 96%. This study demonstrates the potential of using machine learning to improve early detection and intervention for ASD in healthcare.

**Keywords.** Autism Spectrum Disorder(ASD), Machine Learning(ML), Early Detection, Web Applications, Accessibility.

# **1. INTRODUCTION**

Autism Spectrum Disorder (ASD) is a developmental disability that affects the brain, causing challenges with social interaction and communication, along with limited interests and repetitive behaviors. Early diagnosis and intervention can significantly improve the outcomes and lives of individuals with ASD and their families. However, current diagnostic methods are often time-consuming and subjective, frequently leading to late diagnosis [1].

The motivation behind this research is to leverage machine learning (ML) to address these challenges. ML offers promising potential for more accurate and efficient ASD diagnosis by analyzing large datasets of clinical and behavioral features. These algorithms can identify patterns associated with ASD and predict an individual’s risk of developing the disorder, potentially allowing for earlier and more objective diagnoses [2].

This paper aims to develop a ML tool for the early intervention of ASD. With a focus on accessibility and efficiency, various ML algorithms, including Logistic Regression, SVM, Decision Tree and Random Forest, will be used and compared for classification tasks. Emphasis will be placed on optimizing hyperparameters for the Random Forest model to improve its performance. Additionally, a web application will be created to allow easy data input and real-time access to ASD risk statistics, prioritizing both accessibility and environmental sustainability.

Evaluation will involve assessing accuracy, utility, and impact, with user feedback integrated for continuous improvement. Ethical considerations will guide data privacy and transparency throughout the development process to build user trust [3]. By providing an accessible, efficient, and ethical tool for early ASD diagnosis, this research aims to improve the lives of those affected by ASD.

The paper introduces the topic in Section-1. Section-2 explores existing research, while Section-3 outlines the proposed methodology. Section-4 presents the experimental findings, and the paper concludes with a summary in Section-5.

# **2. RELATED WORK**

Machine learning techniques for ASD assessment and detection. These papers highlight various approaches to utilizing machine learning algorithms for different aspects of ASD diagnosis and classification. Related work focuses to identify and highlight existing challenges and pinpoint opportunities for innovation in the current approach.

In the paper [5], the authors introduce a machine-learning framework for autism assessment utilizing logistic regression. They utilize data from AQ-10 adult and adolescent assessment methods and employ search methods to identify significant features, achieving high sensitivity, specificity, and classification accuracy. While this addresses the scarcity of autism datasets and enhances research quality, limitations include potential biases in data collection and subjectivity in relying on domain expert code. Additionally, the assumption of a consistent relationship between features and autism may restrict its generalizability.

In the paper [14], the study aims to distinguish speech sounds of children with ASD from typical children, achieving high accuracy. However, the study relies on public datasets, raising concerns about data quality and representativeness. Moreover, the selected feature selection methods may introduce bias or overlook crucial features necessary for precise ASD diagnosis.

In paper [15], the study addresses Autism Spectrum Disorder (ASD), a neurological condition impacting socialization and learning. The paper presents a method utilizing Kernel Extreme Learning Machine (KELM) optimized by the Giza Pyramids Construction (GPC) algorithm for accurate ASD classification, achieving high accuracy. The GPC data processing step enhances the performance of the KELM model, aiding in the automation of ASD diagnosis.

In the paper [2], the study explores utilizing social nonverbal interactions captured in videos with deep neural networks to differentiate children with ASD from typically developing peers, achieving 80.9% accuracy. It observes a correlation between prediction probability and the severity of autistic symptoms. However, the study is limited by a small sample size, context limitations, definitional subject matter, scalability challenges, and a focus on specific aspects rather than providing a comprehensive understanding of ASD.

In the paper [8], the study presents a ML framework for early detection of ASD, utilizing feature measurement and various ML algorithms on standard ASD datasets. While demonstrating efficient methods, the study lacks detailed comparison with previous work and generalization to diverse datasets, which may limit its power. Additionally, notable limitations include biases in feature selection and the need for revalidation in different populations and clinical settings.

In the paper[1],the study introduces a novel machine learning framework for ASD detection, emphasizing the potential of ML to provide more accurate diagnoses compared to traditional methods. The authors use various ML algorithms to analyze behavioral data, achieving significant accuracy improvements. However, the paper primarily focuses on algorithmic development and does not address integration into clinical workflows.

Paper [4] compares various ML models for early ASD detection, highlighting the efficiency of these models in identifying ASD. Despite demonstrating high accuracy, the study lacks detailed comparisons with existing literature and fails to address the diversity of datasets.

In the paper [18], the study employs various ML models for early-stage ASD detection, achieving significant accuracy improvements. The need for generalization to diverse datasets and clinical validation are notable limitations.

In the paper [16], the study uses a light gradient boosting machine to detect ASD from screening test data, demonstrating cost-effective and efficient methods. However, the study's focus on specific test data may limit its applicability in broader clinical settings.

In the paper [17], the authors analyze randomization-based approaches for ASD, highlighting the potential of these methods for accurate classification. However, the study lacks detailed comparisons with existing literature and broader datasets.

Research Gap and Major Contributions

The existing literature demonstrates the potential of ML for ASD diagnosis but highlights several gaps: the need for diverse and high-quality datasets, the integration of ML tools into clinical workflows, and the generalization of models to broader populations. This paper addresses these gaps by developing an ML tool, optimizing hyperparameters for the Random Forest model, and creating a web application for easy data input and real-time ASD risk assessment. The focus on accessibility, environmental sustainability, and continuous improvement through user feedback represents a significant contribution to the field.

The goal here is to develop a robust machine learning-based system for the early detection of autism spectrum disorder (ASD) in toddlers. This system will utilize a dataset consisting of various demographic and behavioral features, including responses to specific questions and clinical observations. The ultimate aim is to create a tool that can accurately classify a person as either having ASD traits or not, based on the provided features.

# **3. PROPOSED METHODOLOGY**

## **3.1 Data Collection**

The Data set consists of 1054 records of which 728 are positive cases indicating the presence of ASD and 326 are negative cases indicating no signs of ASD. The dataset consists of 735 males and 319 females. It serves as the basis for developing and training ML models for early detection of ASD. Identifying and accessing a variety of information such as validated for demographics (age, gender, ethnicity) and data methods (clinical examinations, behavioral observations, physical measurements) of ASD research [20].

## **3.2 Data Preprocessing**

During data preprocessing, various techniques are used to deal with missing values ​​to ensure that they do not affect the analysis. Outliers are removed to maintain data integrity and reduce noise through techniques such as smoothing or filtering. Additionally, the data is converted into a format compatible with machine learning algorithms for further processing. Relevant features are included during model development through making use of filter-out strategies together with correlation evaluation. This guarantees that the maximum informative features are retained, enhancing the performance and effectiveness of the predictive models.

|  |  |
| --- | --- |
| **Features** | **Description** |
| Case-No | Serial Number |
| Q1 | I pick up on subtle sounds that others seem to miss. |
| Q2 | I'm more interested in the overall concept than getting bogged down in specifics. |
| Q3 | I can handle switching between tasks with ease. |
| Q4 | I can swiftly return to my previous task after an interruption |
| Q5 | I'm good at picking up on unspoken cues in conversation. |
| Q6 | I can usually tell if someone's attention is wandering during a conversation. |
| Q7 | Understanding characters' intentions in stories can be challenging for me. |
| Q8 | I enjoy gathering information about different categories of things. |
| Q9 | I'm good at picking up on people's emotions by looking at their faces. |
| Q10 | I struggle to read people's true purposes. |
| Qchat-10-Score | The results come from a questionnaire called the Quantitative Checklist for Autism in Toddlers. |
| Age-Mons | Age of the toddlers in months |
| Sex | The gender of the toddlers is categorized as male or female |
| Ethnicity | Ethnic background of the toddlers |
| Who completed the test | Details about the person who completed the ASD assessment test. |
| Family\_mem\_with\_ASD | Any previous history of ASD in family members. |
| Jaundice | Presence or absence of jaundice at birth |
| Class/ASD Traits | Binary classification indicating the presence or absence of ASD traits in toddlers |

Table-1.Description of features in ASD dataset.

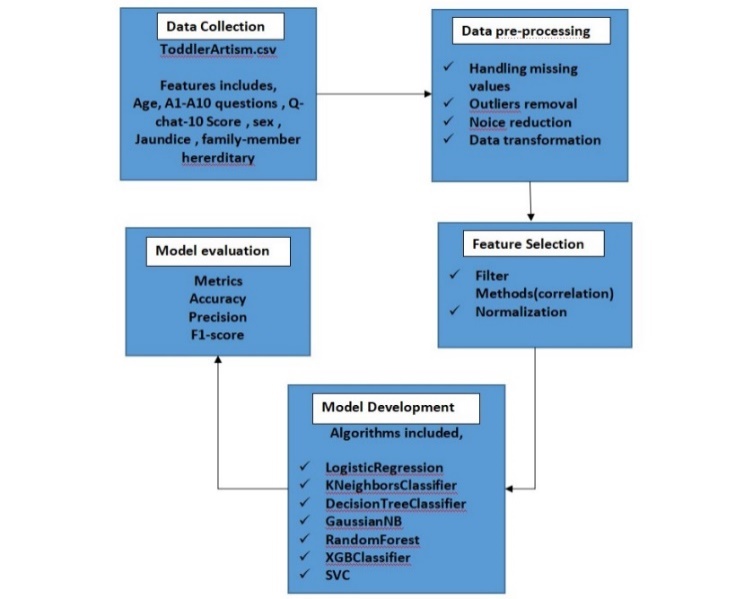


Figure1. Proposed methodology for autism classification.

Figure 1 presents the proposed methodology for toddler autism classification as discussed as follows.Initially, the data is collected from the Kaggle and stored in a CSV file named as Toddler\_Artism.csv. Next, the Data pre-processing cleans the data by handling missing values, removing outliers, and reducing noise. Then feature selection techniques are applied to choose the most relevant attributes for model training. The relevant features are fed into the machine learning classifiers are trained and calculated using metrics such as accuracy, F1-score and precision.

**3.3 Algorithms Used**

A Various set of machine learning algorithms are integrated for developing predictive models, consisting of KNeighbors classifier, SVC Logistic Regression, ,Gaussian NB, , XG Boost, Decision Tree , and Random Forest. Each algorithm gives particular strengths and insights, taking into consideration a comprehensive exploration of the information and the development of a strong predictive model. Each algorithm was trained on the dataset, which had been preprocessed to handle missing values, remove outliers, and transform the data into a suitable format. Feature selection was applied to ensure the inclusion of the most informative attributes.By combining multiple decision trees, Random Forest improves generalization performance and reduces variance compared to individual trees. This can lead to more accurate predictions for ASD risk assessment. The trained models are precisely evaluated using performance metrics including F1-Score, recall, accuracy, and precision. These metrics provide deep insights into the effectiveness of the predictive nature of the models and enable comparisons between algorithms to become aware of the best suitable technique for ASD detection.

The dataset 𝐷 is split into training set *Dtrain*​ and testing set *Dtest*​.

This is mathematically denoted as:

*D*→(*Dtrain*​,*Dtest*​)

The Random Forest aggregates the predictions from all trees. For classification, it uses majority voting:

*F*(*x*)=mode{*f*1​(*x*),*f*2​(*x*),...,*fN*​(*x*)}

where N is the number of trees.

The performance metrics of the Random Forest model reveal significant insights into its efficacy and areas for improvement in ASD detection. The low training accuracy of 49% indicates potential underfitting, where the model fails to capture the underlying patterns of the training data. Conversely, a perfect test accuracy of 100% suggests possible overfitting, where the model may have memorized the test data rather than generalizing from it, highlighting a critical discrepancy that needs addressing for reliability. A precision of 1.00 is vital in a medical context, ensuring no false positives and preventing unnecessary stress for toddlers and their families. Similarly, a recall of 1.00 guarantees that all true positive cases are identified, crucial for early ASD intervention. The F1-score of 1.00 indicates a perfect balance between precision and recall, underscoring the model's robustness in detecting ASD. However, the discrepancy between training and test performance necessitates caution in interpreting these results. Addressing these issues through data augmentation, regularization, and cross-validation will enhance the model's reliability and effectiveness in early ASD detection.

**3.4 Application Development**

A computer screen shot of a computer screen

Description automatically generated

Figure 2. Web-developed page

For web application development, selecting a scalable framework like Flask focuses on user interface design and integration capabilities. The user interface will be intuitive, allowing users to assess easy input of context and providing transparent ASD risk estimates. To conduct the evaluation and impact assessment, of standardized metrics on independent testing data to assess the performance of accuracy, utility, and potential impact of early ASD identification and intervention.

## **4. EXPERIMENTAL RESULTS**

The experiment was conducted on the Anaconda-Jupyter Notebook and extensively employs machine learning techniques to analyze and predict ASD risk accurately. HTML, JavaScript, and CSS are utilized for developing a user-friendly web interface, enhancing accessibility and usability

Figure 3 presents the visual representation of feature name vs feature importance. Where the x-axis tells the importance score, ranging from 0 to 0.2 and y-axis contains features like Family\_mem\_with\_ASD, Jaundice, Ethnicity, Sex, Age\_Mons, and Features labeled A1 to A10. The bar indicates how significant each feature is in predicting outcomes. According to Fig-4, the feature age\_in\_mons has the highest feature importance.

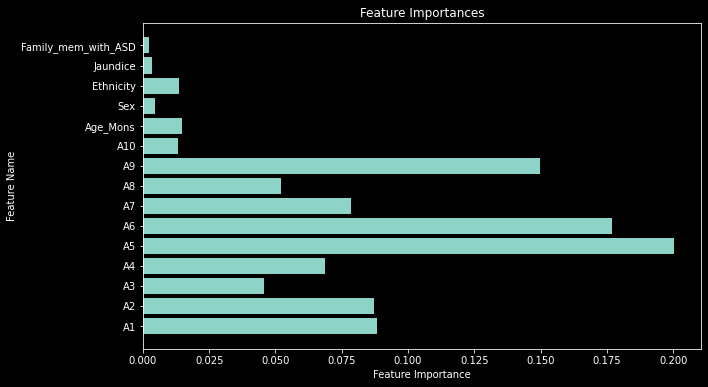


Figure 3. Feature Name vs Feature Importance

Figure 4 discusses the heatmap representing how strongly different variables in the dataset are related to each other. Darker colors indicate stronger relationships, helping to visualize patterns and connections between the dataset's features. Overall, this graph is a visual representation of feature co-occurrence analysis in the text data, helping to identify patterns and relationships between variables. The heatmap provides a quick and effective way to assess the strength and direction of these correlations.

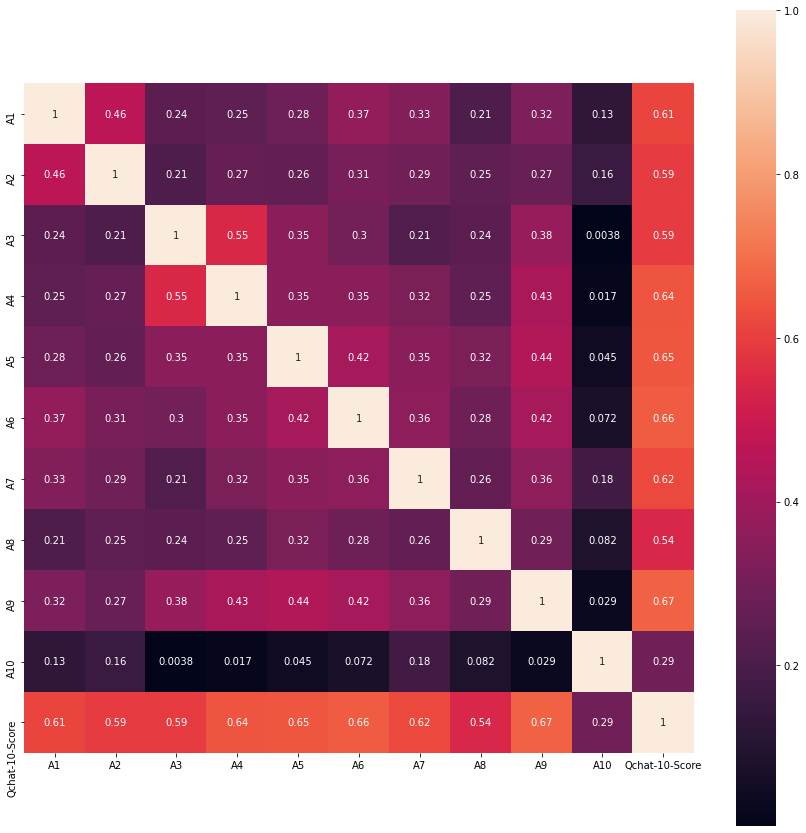


Figure 4.Correlation between features

A confusion matrix is displayed in Figure 5, summarizing the performance of a classification model by categorizing predictions into true positives, true negatives, false positives, and false negatives. The output includes a summary of the test accuracy rounded to two decimal places. In this case, it is 1.0.

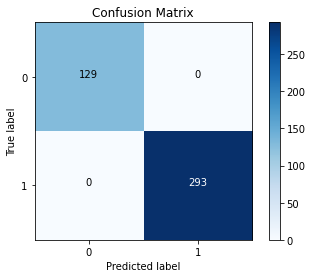


Figure 5. Performance of Support Vector Classifier (SVC)

The Random Forest classifier, trained with a class weight of 'balanced', achieves incredible performance in ASD prediction with an accuracy of 99%. It has well-balanced precision and recall values, along with a minimal number of misclassifications, which underscore its effectiveness in distinguishing between individuals with and without ASD. It achieves high precision and recall for both classes, and the model demonstrates robust generalization to unseen data, making it a reliable tool for ASD risk assessment.

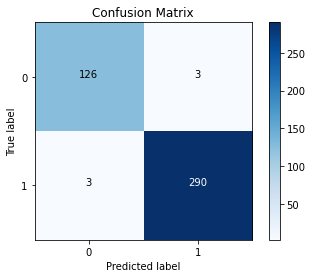


Figure 6. Grid Search CV with Random Forest Classifier, Random Forest Classifier within a Pipeline.

Figure 7 presented the test accuracy scores of various classifiers. Logistic Regression, Linear Discriminant Analysis, and XG Boost achieved high accuracy, indicating their effectiveness in predicting ASD risk. Classifiers like Ada Boosting and Random Forest\_before grid exhibited lower accuracy performance. Random Forest\_after grid performed well after applying GridSearch CV and Pipeline methods enhance the accuracy performance.

|  |  |
| --- | --- |
| **Classifier** | **Test Accuracy** |
| Logistic Regression | 1.0 |
| KNeighborsClassifer | 0.91 |
| DecisionTreeClassifier | 0.91 |
| GaussianNB | 0.94 |
| SVC\_beforegrid | 0.78 |
| RandomForest\_beforegrid | 0.64 |
| XGBClassifier | 0.99 |
| GradientBoosting | 0.64 |
| AdaBoosting | 0.49 |
| SVC\_aftergrid | 1.0 |
| RandomForest\_aftergrid | 0.96 |

**Table-2.** Applied Classifiers concerning test accuracy

# **5. CONCLUSION**

This paper attempted to use machine learning techniques to detect autism spectrum disorder (ASD) across different age groups, including children, teens, and adults in non-clinical cases. Various performance evaluation parameters were employed to assess the models used in diagnosing ASD. The results were compared with recent studies, demonstrating that our approach enhances the prediction model's performance significantly.

By utilizing Random Forest sampling, we were able to detect ASD more accurately. The application of Grid Search CV and pipeline techniques proved to be instrumental in improving the model's accuracy from 64% to 96%. Despite the notable performance metrics such as precision, recall, and F1-score all reaching 1.00, the observed discrepancy between training and test accuracy indicates potential issues of underfitting and overfitting that need to be addressed.

These findings suggest that with further refinement and addressing of the overfitting issue, the model could serve as a robust tool for early detection of ASD. The improvements in accuracy and the balanced performance metrics highlight the potential of machine learning techniques, particularly Random Forest, in enhancing ASD detection and providing timely interventions. Future work should focus on fine-tuning the model, incorporating more diverse datasets, and exploring additional machine learning algorithms to ensure the model's generalizability and reliability across various populations.

# **6. REFERENCES**

1. Sharif, H., & Khan, R. A. (2022). A novel machine learning based framework for detection of autism spectrum disorder (ASD). Applied Artificial Intelligence, 36(1), 2004655.
2. Kojovic, N., Natraj, S., Mohanty, S. P., Maillart, T., & Schaer, M. (2021). Using 2D Video-based Pose Estimation for Automated Prediction of Autism Spectrum Disorders in Preschoolers. medRxiv, 2021-04.
3. Del Coco, M., Leo, M., Carcagnì, P., Fama, F., Spadaro, L., Ruta, L., ... & Distante, C. (2017). Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot. IEEE Transactions on Cognitive and Developmental Systems, 10(4), 993-1004.
4. 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.
5. Thabtah, F., Abdelhamid, N., & Peebles, D. (2019). A machine learning autism classification based on logistic regression analysis. Health information science and systems, 7, 1-11.
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. H. Chahkandi Nejad, O. Khayat, and J. Razjouyan, ‘‘Software development of an intelligent spirography test system for neurological disorder detection and quantification,’’ J. Intell. Fuzzy Syst., vol. 28, no. 5, pp. 2149–2157, Jun. 2015.
8. Hasan, S. M., Uddin, M. P., Al Mamun, M., Sharif, M. I., Ulhaq, A., & Krishnamoorthy, G. (2022). A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders. IEEE Access, 11, 15038-15057.
9. 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.
10. 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. 2971–2976, Mar. 2017.
11. M. F. Misman, A. A. Samah, F. A. Ezudin, H. A. Majid, Z. A. Shah, H. Hashim, and M. F. Harun, ‘‘Classification of adults with autism spectrum disorder using deep neural network,’’ in Proc. 1st Int. Conf. Artif. Intell. Data Sci. (AiDAS), Sep. 2019, pp. 29–34.
12. S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, ‘‘Applications of support vector machine (SVM) learning in cancer genomics,’’ Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41–51, Jan./Feb. 2018.
13. Mohanta, A., & Mittal, V. K. (2022). Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. Computer Speech & Language, 72, 101287.
14. R. Gaspar, A., Oliva, D., Hinojosa, S., Aranguren, I., & Zaldivar, D. (2022). An optimized Kernel Extreme Learning Machine for the classification of the autism spectrum disorder by using gaze tracking images. Applied Soft Computing, 120, 108654.
15. M. Alsuliman and H. H. Al-Baity, ‘‘Efficient diagnosis of autism with optimized machine learning models: An experimental analysis on genetic and personal characteristic datasets,’’ Appl. Sci., vol. 12, no. 8, p. 3812, Apr. 2022.
16. S. P. Kamma, S. Bano, G. L. Niharika, G. S. Chilukuri, and D. Ghanta,‘‘Cost-effective and efficient detection of autism from screening test data using light gradient boosting machine,’’ in Intelligent Sustainable Systems.Singapore: Springer, pp. 777–789, 2022.
17. U. Gupta, D. Gupta, and U. Agarwal, ‘‘Analysis of randomization-based approaches for autism spectrum disorder,’’ in Pattern Recognition and Data Analysis with Applications. Singapore: Springer, pp. 701–713, 2022.
18. T. Akter, M. Shahriare Satu, M. I. Khan, M. H. Ali, S. Uddin, P. Lio, J. M. W. Quinn, and M. A. Moni, ‘‘Machine learning-based models for early-stage detection of autism spectrum disorders,’’ IEEE Access, vol. 7, pp. 166509–166527, 2019.
19. Kaggle. (2022). Autism Spectrum Disorder Detection Dataset for Toddlers. [Online]. Available: <https://www.kaggle.com/fabdelja/autism-screeningfor-toddlers>