Proceedings of the International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS-2024)
IEEE Xplore Part Number: CFP242B2-ART: ISBN: 979-8-3315-1809-7

Employing Artificial Intelligence Methods for the Diagnosis of Autism Spectrum Disorder in Children

Sridevi R
Department of Computer Science.
CHRIST (Deemed to be University
Bangalore,India
sridevi.r@christuniversity.in

Helen k joy
Department of Computer Science.
CHRIST (Deemed to be University)
Bangalore,India
helenk.joy@christuniversity.in

Karthikeyan K J
Department of Computer Science
CHRIST (Deemed to be University
Bangalore, India
karthikeyan.kj@msam.christuniversity.in

Gopika S S
Department of Computer Science,
CHRIST(Deemed to be University),
Bangalore, India.
gopika.ss@msam.christuniversity.in

Neha Seirah Biju
Department of Computer Science,
CHRIST(Deemed to be University),
Bangalore, India.
neha.biju@msam.christuniversity.in

Shriniha PA
Department of Computer Science,
PSG Institute of Technology and Applied Research,
Coimbatore, India.
22 b146@psgitech.ac.in

2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS) | 979-8-3315-1809-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICNIS64247.2024.10823377

***Abstract*—Accurate and timely diagnosis of disorder known as autism spectrum disorder (ASD) is not an easy task due to the complicated neurodevelopmental condition's high clinical presentation variation. In order to improve the diagnostic procedure for ASD in pediatric patients, machine learning (ML) techniques have come to light as potential approaches. The previous surveys about the practice of ML algorithms for diagnosing ASD in children has been thoroughly reviewed and summarized. The supervised and unsupervised learning, feature selection, and ensemble methods used in ASD research are among the many ML techniques that is methodically examined. The necessity of large-scale, diverse datasets, cross-validation methods, and interpretability are emphasized over the advantages, disadvantages, and potential future directions of ML-based ASD diagnostic models. This study attempts to offer insights for researchers, clinicians, and other stakeholders in the field of ASD diagnosis by critically analyzing the current status of ML in ASD Diagnosis.**

***Index Terms*—Autism detection, machine learning algorithms, pediatric autism, supervised learning, unsupervised learning.**

I. Introduction

A spectrum of neurodevelopmental problems popularly called as autism spectrum disorder (ASD) is typified with enduring complications with society interaction, interpersonal communication, and tedious activities. About 18 million individuals in India have been diagnosed with autism, according to the prevalence rate. Approximately 1% to 1.5% of children between the ages of two and nine have an ASD diagnosis. The effects of ASD are profoundly detrimental to impacted people, their families, and society at large. For children with ASD to achieve better results and reach their full developmental potential, early diagnosis and intervention are essential. Nevertheless, because ASD has a wide range of clinical presentations and no conclusive molecular indicators, identifying autism can be challenging. Clinical observation and subjective evaluations play a major role in traditional diagnostic methods, which can cause unpredictability and even delays in diagnosis.

Optimizing ASD diagnosis in juvenile populations with machine learning (ML) approaches has garnered increasing attention in recent years. Machine learning algorithms present the possibility of examining extensive datasets and detecting patterns that might not be deceptive using conventional techniques. ML models seek to provide more objective and reliable assessments of ASD risk and severity by combining several data sources, such as behavioral observations, genetic profiles, neuroimaging data, and clinical histories. Notwithstanding the ability of machine learning (ML) in the ASD Diagnosis, a number of obstacles still need to be addressed, such as the requirement for thorough validation studies, openness in the interpretation of models, and ethical considerations. By integrating the body of data, debating methodological strategies, and outlining potential future paths for both clinical and research applications, this review needs to critically assess the state of ML techniques for ASD diagnosis in pediatric patients. This review helps to the creation of useful and approachable diagnostic instruments for kids with ASD by deepening the understanding of machine learning's potential in ASD diagnosis.

II. Literature Review

In recent years, researchers have made significant strides in harnessing advanced computational techniques to revolutionize

the diagnosis of autism spectrum disorder (ASD). One notable avenue of exploration has been the application of deep learning methodologies, particularly convolutional neural networks (CNNs), to analyze electroencephalography (EEG) signals with the goal of identifying distinctive patterns associated with ASD. This approach, as pursued by Smith et al. holds promise for enhancing diagnostic accuracy by uncovering subtle neurophysiological variations between individuals with ASD and neurotypical individuals [1]. Concurrently, machine learning algorithms have been employed to analyze structural magnetic resonance imaging (MRI) data for early detection of ASD in young children, as demonstrated in the work of Johnson et al. [2]. By identifying MRI-based biomarkers, as elucidated by Johnson et al., these predictive models offer the potential for timely intervention and support, critical for optimizing long-term outcomes. Furthermore, researchers such as Garcia et al. have explored the integration of diverse data modalities, including genetic, neuroimaging, and behavioral data, to enhance the precision of ASD diagnosis [3]. By amalgamating multifaceted information sources, this approach seeks to develop more robust diagnostic models capable of comprehensively capturing the complexity of ASD. In tandem, advancements in computer vision algorithms, as investigated by Patel et al., have facilitated the development of automated facial analysis systems for ASD screening in young children [4]. These systems leverage facial expressions and features to detect potential indicators of ASD, offering a non-invasive and efficient screening mechanism for early identification. Additionally, research efforts led by scholars such as Kim et al. have emphasized the importance of feature selection techniques in refining the interpretability and efficacy of ASD diagnostic models. By carefully selecting informative features from heterogeneous datasets, these methodologies aim to cultivate clinically relevant diagnostic tools endowed with heightened discriminative capabilities [5]. These interdisciplinary endeavors, spearheaded by esteemed researchers such as Smith, Johnson, Garcia, Patel, and Kim, underscore the transformative potential of leveraging technological innovations to advance ASD diagnosis, ultimately fostering improved outcomes for individuals with ASD.

III. Working Model

Fig. 1 illustrates the general flow and operation of the system. Preprocessing the dataset first helps to eliminate outliers and missing values while also reducing noise and encoding categorical features. Feature engineering is also used to select the most favourable features from all the features in the data collection. The dimensionality of the data gets lowered to improve training speed and efficiency. Utilizing classification techniques like Support Vector Machine, Random Forest Classifiers, Extreme Gradient Boosting(XGB) and Logistic Regression, the output label (ASD or no ASD) is predicted after the data set has undergone preprocessing. The accuracy of each classifier is compared and noted. F1 score and precision-recall values are two more metrics that have been generated to improve the evaluation of each classifier. The training accuracy of the classifier will exceed its test accuracy classification if it is determined to be the best model.

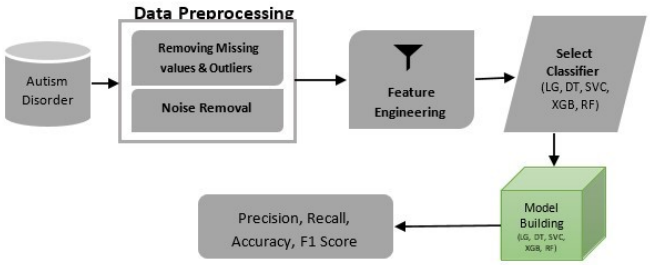

Fig. 1. Working Model of Autism detection using machine learning algorithms

if it operates effectively. This model can then be utilized for additional training and classification if it is the best model.

IV. Methodology

A. Data Preprocessing

The assembled dataset [6] that is utilized in this review includes binary, continuous, and category attributes. The collection contained 28 attributes and 1985 occurrences at first. Preprocessing the data was necessary because the dataset included a few non-contributing and category attributes. The changes made to a data collection prior to feeding it into the model are referred to as preprocessing. In order to improve its suitability for training and analysis, raw or noisy data must be cleaned. The NA values for ”Depression,” ”Qchat-10-Score,” and ”Social/Behavioral Disorder” were eliminated. In order to handle the category data, label encoding is being used. In order to render the labels machine-readable, label encoding transforms them into numerical form. The value allocated to repeated labels remains the same as it was previously. The inefficiency of Label Encoding occurs when there are more than two classes.

B. Classification Algorithms

1) Support Vector Machines: The goal of SVM is to identify the ideal hyperplane for classifying data points into distinct groups. Support vectors are the data points that are closest to the hyperplane and they are used to calculate the decision boundary. Maximizing the margin between the support vectors and the hyperplane is the SVM optimization goal, and it is accomplished by resolving the optimization problem [7]:

$$minw,b1/2w2 \quad (1)$$

subject to:

$$yi(wTxi + b)1fori = 1,...,n \quad (2)$$

Here, w represents the vector representing weight, b is the term representing bias, xi are the feature vectors taken as an input, yi are the class labels (+1 or -1), and n is the number of training samples.

2) Random Forest Classifier: During training, Random Forest constructs several decision trees and then combines their predictions to produce the final classification [8]. Random data samples and a selection of features are used in the construction of each decision tree. The final forecast is the average of the various projections for each tree. Random Forest introduces randomization into the tree-building process, hence reducing overfitting and improving generalization performance. A feature's importance in a Random Forest can be ascertained by measuring how much it lowers impurity across all decision trees.

3) Extreme Gradient Boosting (XGBoost): An enhanced version of the ensemble learning method known as gradient boosting is called XGBoost [9]. It builds a sequence of decision trees one after the other, correcting the errors in the previous trees. XGBoost minimizes a regularized objective function, which incorporates a loss function and a regularization term, to prevent overfitting.

4) Decision Tree Classifier: A flexible supervised learning method used for both regression and classification applications is the decision tree classifier [10]. To improve the homogeneity of the resultant subsets, it divides the feature space iteratively according to the most informative features and matching thresholds. Until certain stopping conditions are satisfied, like attaining a maximum tree depth or a minimum impurity level, this recursive process continues. Because decision trees imitate human decision-making processes, they are well-known for being interpretable and useful for understanding underlying patterns in data. Nevertheless, they may have trouble capturing complex decision boundaries and may suffer from overfitting noisy data. Multiple trees are combined in ensemble methods such as Random Forest and Gradient Boosting to overcome these difficulties. Decision tree classifiers, in spite of their simplicity, provide efficacy and transparency across different domains.

5) Logistic Regression: When attempting to forecast the likelihood of a binary outcome based on one or more predictor variables, a statistical technique called logistic regression [11] is employed. The logistic function, commonly known as the sigmoid function, is used in logistic regression to represent the probability that an observation belongs to a particular class, in contrast to linear regression, which predicts continuous outcomes. The output of a linear combination of predictor variables is mapped by this function to a value between 0 and 1, which denotes the likelihood of the positive class. Using methods like maximum likelihood estimation and gradient descent optimization, logistic regression calculates the model's parameters, including an intercept term and coefficients for the predictor variables. Logistic regression is not just for regression, despite its name and used in many classification problems in terms of its efficiency and interpretability. it also give insights into the correlation between the predictor variables and the outcome probability, which makes it as a valuable tool in health sector, finance and marketing sectors.

C. Hyperparameter optimization

Hyperparameters are preset configuration options that are selected prior to the commencement of the training process; they are not learned directly from the data. Precise parameter tuning necessitates systematically probing the hyperparameter space using methods such as random search, grid search, or more advanced strategies like Bayesian optimization. The goal of this investigation is to find the combination of hyperparameters that maximizes a selected evaluation measure on a validation dataset, like accuracy or F1 score. Reducing the chance of overfitting to the training set, this tuning must be done on a different validation set or via cross-validation. When choosing hyperparameters, one should also take into account the computational resources that are available and the trade-offs between the model complexity and the performance. Optimizing parameters effectively is crucial to improving a model's capacity to generalize and perform well on unseen data.

D. Results and Disussion

1) Evaluation Metrics: During the machine learning model development, the meticulous examination of evaluation metrics and performance assessment stands as a pivotal aspect, offering invaluable insights into the capacity of the model to generalize and execute proficiently on novel data samples. These metrics serve to quantify the model's efficacy by scrutinizing its predictions against actual outcomes derived from independent validation or test datasets. Within classification tasks, an array of standard evaluation metrics come into play, including precision, recall, accuracy, F1 score, and the area under the receiver operating characteristic curve as in Fig 2. Accuracy denotes the ratio of correctly classified instances, while precision delineates the fraction of true positives amid all positive predictions. Sensitivity, or recall, gauges the proportion of true positives accurately identified by the model. F1 score harmonizes precision and recall, furnishing a unified metric that accounts for both false positives and false negatives. The trade-off between the true positive rate and false positive rate is evaluated over a range of probability thresholds using ROC-AUC. On the other hand, evaluation measures like mean squared error (MSE), mean absolute error (MAE), or R-squared (coefficient of determination) are required for regression assignments. The judicious selection of evaluation metrics hinges upon the particular problem domain, objectives, and constraints at hand. Table 1 entails a comprehensive comparative analysis of model performance against baseline models, robustness evaluations across diverse datasets or temporal domains, and sensitivity analyses to gauge the influence of hyperparameters or feature selection techniques. Effective deliberation of

evaluation metrics and performance assessment is indispensable for iteratively refining model accuracy, dependability, and applicability in real-world scenarios.

Table 1. Comparison of the evaluation metrics of different classifiers

varied models have varied predictive performance strengths and shortcomings, as shown by a comparative analysis. It is critical to evaluate the models' performance using a variety of measures, including precision, recall, Accuracy and F1 score, and to take into account the trade-offs between training and validation accuracy. A clear conclusion regarding the best model to predict ASD can be made by evaluating each model's relative performance, taking accuracy and generalization to unknown data into account.

2) Dataset Analysis: In machine learning model development, conducting a thorough analysis of the dataset is a critical preliminary step, providing foundational insights into the inherent characteristics and patterns within the data. This

Classifier	Precision	Recall	Accuracy	F1 Score
Logistic Regression	90%	84.11%	85.97%	86.96%
Decision Tree	93.33%	78.50%	84.93%	85.27%
SVC	69.76%	66.82%	65.45%	68.26%
XGBoost	98.58%	97.66%	97.92%	98.12%
Random Forest	94.63%	65.88%	78.96%	77.68%

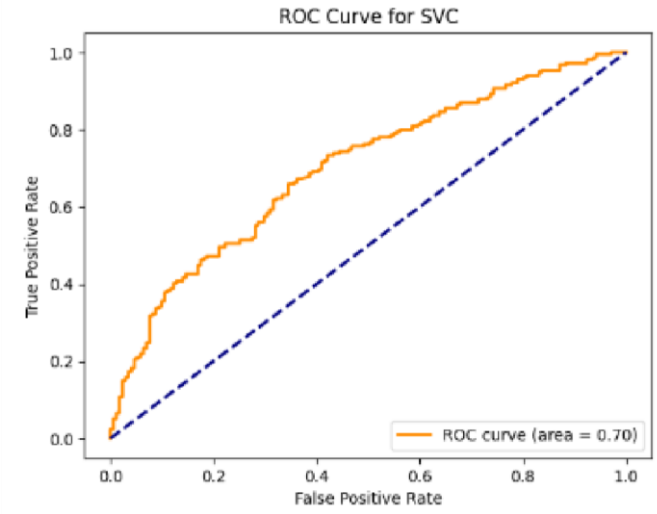

Fig. 2. ROC Curve for SVC

initial phase involves a meticulous examination of the dataset's structural attributes, distributional tendencies, and overall quality, with the primary objective of deriving valuable insights to guide subsequent modeling efforts. Key tasks in dataset analysis include evaluating descriptive statistics to uncover nuanced feature distributions, implementing strategies to address missing or erroneous data points through careful imputation or prudent exclusion, and assessing class distributions to identify potential imbalances. Furthermore, the employment of sophisticated techniques such as data visualization and dimensionality reduction facilitate the discernment of salient features and the comprehension of inter-variable relationships. Through the comprehensive undertaking of dataset analysis, researchers and practitioners are poised to extract indispensable insights vital coefficients that are closer to 1 imply strong positive correlations, whereas those closer to -1 suggest strong negative correlations. There is little to no linear relationship between the variables when the coefficient is close to 0. Using color gradients—warmer colors for positive correlations and cooler colors for negative correlations—the heatmap effectively depicts the direction and strength of these associations. Based on the underlying correlations found in the dataset, this analysis technique helps researchers to make well-informed decisions on feature selection, model development, and hypothesis formulation.

3) Comparison of Classification Models: The model evaluation's findings offer insightful information on how well machine learning methods predict ASD. The interpretation of these results should consider the models, potential overfitting, and their capacity to generalize to new data. It essential to for crafting robust and efficacious machine learning models, finely attuned to the intricacies and idiosyncrasies of the dataset under examination.

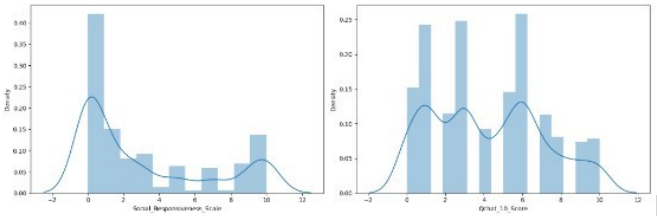

Fig. 3. Analysis of the density spread of social responsiveness scale and Qchat_10_score

The data appears as in Fig.3, looks roughly symmetrical when analyzing the density spread of social responsiveness scale and Qchat10*score.*

An effective tool for examining correlations within the dataset is a correlation matrix heatmap, which is particularly useful for delving into the subtleties of the supplied data. With the use of visualization technique represented in Fig 4, which shows the correlations between the actual and predicted values of dataset's many variables, patterns, dependencies, and possible multicollinearity problems can be found. Each cell in the heatmap represents the correlation coefficient between two variables, and its values range from -1 to 1. Correlation analyse the implications of each model's performance in a clinical context to understand how these models could potentially contribute to early detection and treatment for the affected kids. From the model evaluation in Fig.5, it is found that the logistic regression model demonstrates good precision and recall, showing its ability to correctly classify positive cases while minimizing false positives. The decision tree classifier exhibits precision, recall, accuracy, and F1 score, suggesting robust performance in predicting autism spectrum disorder. The SVC model shows lower precision, recall, accuracy, and F1 score compared to the other models, indicating potential

limitations in its predictive performance for ASD. This holistic approach to model evaluation and performance analysis ensures that the predictive models for ASD are rigor

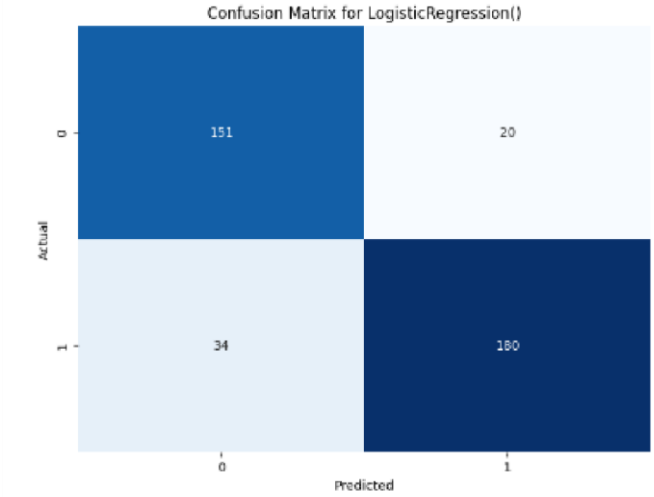

Fig. 4. Confusion matrix for Logistic Regression

assessed and their implications for clinical application are carefully considered.

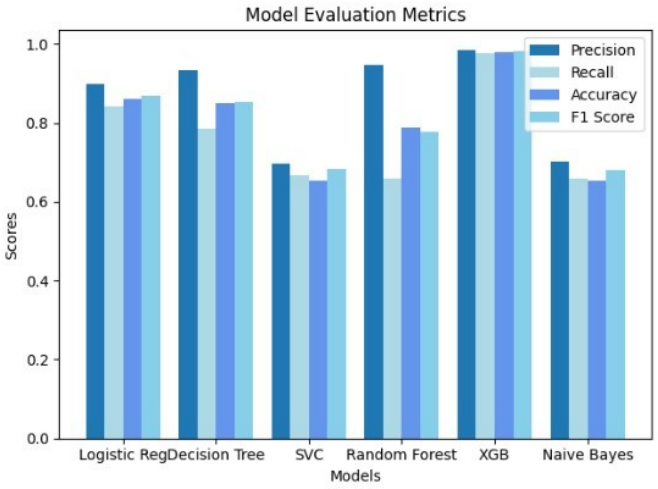

Fig. 5. Comparison of different classifiers upon various metrics

CONCLUSION

The review demonstrates how machine learning techniques can be used to predict autism spectrum disorder (ASD). The most reliable model found was the XGBoost, which consistently outperformed support vector machine (SVC) and logistic regression models in terms of various evaluation metrics. Its superior performance suggests its significance in clinical applications, particularly in aiding early detection and treating for individuals with ASD. Further research is warranted to refine and expand predictive models, potentially incorporating additional features or exploring ensemble methods. However, ethical considerations, patient privacy, and responsible technology use in healthcare must be carefully addressed before practical implementation of ASD predictive models. Rigorous validation and clinical trials are essential to ensure the safety, reliability, and ethical deployment of such models in realworld healthcare settings. Overall, the study underscores the promise of machine learning in ASD prediction and emphasizes the importance of continued research efforts to advance early detection and intervention for individuals on the autism spectrum

REFERENCES

[1] Laje, G., Morse, R., Richter, W., Ball, J., Pao, M., Smith, A. C.:Autism spectrum features in Smith–Magenis syndrome. American journal of medical genetics part C: Seminars in medical genetics, vol. 154, no. 4 , pp. 456-462 (2010)

[2] Stahl, D., Pickles, A., Elsabbagh, M., Johnson, M. H., BASIS Team.:Novel machine learning methods for ERP analysis: a validation from research on infants at risk for autism.Developmental neuropsychology, vol. 37, no. 3, pp. 274-298 (2012)

[3] Alcaniz Raya, Mariano, et al.: Machine learning and virtual reality on body movements' behaviors to classify children with autism spectrum disorder,Journal of clinical medicine, cilt 9, no. 5, p. 1260 (2020)

[4] Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., Wall, D. P.: Mobile detection of autism through machine learning on home video: A development and prospective validation study,PLoS medicine, cilt 15, no. 11, p. e1002705 (2018)

[5] Moon, S. J., Hwang, J., Kana, R., Torous, J., Kim, J. W.: Accuracy of machine learning algorithms for the diagnosis of autism spectrum disorder: Systematic review and meta-analysis of brain magnetic resonance imaging studies, JMIR mental health, vol. 6, no. 12, p. e14108 (2019)

[6] Dataset: https://www.kaggle.com/datasets/uppulurimadhuri/dataset Kaggle, India (2023)

[7] Pisner, Derek A., and David M. Schnyer.: Support vector machine, Academic Press, pp. 101-121 (2020)

[8] Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., Razak, R. A.:A review of machine learning methods of feature selection and classification for autism spectrum disorder,Brain sciences, cilt 10, no. 12, p. 949 (2020)

[9] Malviya, Meenakshi, and J. Chandra,:A systematic review on prognosis of autism using machine learning techniques,ECS Transactions 107, cilt 1, no. 1, p. 11623 (2022)

[10] Cavus, N., Lawan, A. A., Ibrahim, Z., Dahiru, A., Tahir, S., Abdulrazak, U. I., Hussaini, A.: A systematic literature review on the application of machine-learning models in behavioral assessment of autism spectrum disorder,Journal of Personalized Medicine, cilt 11, no. 4, p. 299 (2021)

[11] Zope, Vidya, Tanvi Shetty, Maitraiyi Dandekar, Anmol Devnani, and Puneet Meghrajani.:ML based approaches for detection and development of autism spectrum disorder: A review,IEEE (2022)