Age-based Data Analysis and Prediction of Autism Spectrum Disorder

Sujatha R[1†], Aarthy SL[1†], Jyotir Moy Chatterjee[2††], NZ Jhanjhi[3†††], Azween Abdulla[3†††], Mahadevan Supramaniam[4††††]

[†]School of Information Technology & Engineering, Vellore Institute of Technology, Vellore, India
[††]Department of Information Technology, Lord Buddha Education Foundation, Kathmandu, Nepal
[†††]School of Computer Science & Engineering (SCE), Taylor's University, Malaysia
[††††]Research and Innovation Management Center, SEGi University,

Summary
Autism spectrum disorder (ASD) is a syndrome prevalent in all age groups that causes immense changes in all aspects of an affected person's life, including social skills, communication, and behavioral style. Screening of the same is a challenging task, and classification must be conducted with great care. The dataset considered for this work is the benchmark dataset retrieved from the UC Irvine (UCI) Machine Learning Repository. The case sample considered here includes approximately 1,000 children of various autism spectrum conditions and age groups mapped as a child, adult, or adolescent. The autism or no autism class categorized based on the following attributes assessed include age, gender, ethnicity, born with jaundice, pervasive developmental disorder of any family member, information about a relationship who is undergoing the test, country of residence, screening methods. Autism spectrum quotients (AQs) varied among a number of scenarios for toddlers, adults, adolescents, and children that include positive predictive value for the scaling purpose. AQ questions referred to topics pertaining to attention to detail, attention switching, communication, imagination, and social skills. The diagnostic decision support system with the provided features for the ASD was optimized on the basis of the selected dataset with the help of machine learning algorithms and soft computing techniques. The dataset was classified by using various algorithms, and accuracies in the range of 85%–95% were obtained.

Keywords:
Autism; International Classification of Disease; machine learning; soft computing; training set; testing set; cross validation; specificity; sensitivity; accuracy.

1. Introduction

The autism spectrum disorder (ASD) screening process differs according to age. Two global classification systems for ASD diagnosis, namely, the Diagnostic Statistical Manual (DSM-5), which is provided by the American Psychiatric Association and considers the condition as a single diagnosis by removing subgroups, and the International Classification of Disease (ICD-11), which was created by the World Health Organization. According to the DSM, autism and intellectual disability occur concurrently. By contrast, the ICD provides a detailed guide to distinguish autism prevailing with and without an intellectual disability; it also considers historical data on loss of previous skill in the diagnostic process. The most difficult aspect of diagnosing ASD is that no single pathognomonic feature exists and all symptoms revolve around the modification of an individual's behavioral profile, which varies according to age and severity.

In the prevailing system, classification is carried out using datasets of cases collected from a versatile group. The data depend on the autism diagnostic observation schedule (ADOS) and autism diagnostic interview (ADI), which is conducted in a clinical setting. ADOS sessions are 30–45 minutes long, and the examiner records the provided responses. ADI refers to interviews of suspected autism individuals over 18 with their parents or caregivers in the clinic. The interview is performed in five phases using a questionnaire that probes areas related to communication, social development, play, restricted behavior, and general skills. The individual's responses are evaluated by using scoring algorithms, and three major areas, namely, language and communication, social interaction, restricted and repetitive behavior, are assessed. Cumulative scores exceeding the corresponding cut off values indicate a positive syndrome that must be addressed immediately by proper diagnosis. Determining the most prominent features from a massive dataset is challenging work that must be done by careful analysis. Data processing tasks also present a potential hurdle in managing missing values in attributes. The rest of the process of applying machine learning largely depends on the quality of data taken into consideration. Automation based on the diagnostic perspective must be fine-tuned.

ASD is a neurodevelopment disorder that can occur in adults, adolescents, children, and toddlers. Leo Kanner refers to autism as a prototypical condition with a spectrum of presentations and phenotypes that become more subtle in terms of behavioral features when a change in environment occurs. It is characterized by behavioral abnormalities in communication and reciprocal social interaction, together with patterns of repetitive, restricted, and stereotyped interests and activities. These issues are usually present in early childhood and are likely to increase in intensity in different settings. The heterogeneity of the affected individuals and their genetic complexity has helped researchers identify the causes of ASD. Diagnosis of ASD is a lengthy process and varies from individual to individual. Symptoms also change across one's lifespan. ASD can be difficult to detect in

Manuscript received April 5, 2025
Manuscript revised April 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.4.9

young children, and parent raise the concern after the persistent monitoring of the children which delays the process of early diagnosis.
This paper is organized as follows. Related work is first presented, and the proposed computational intelligence method workflow is described. Results and a discussion are then provided, and the conclusions are summarized.

2. Related Work

Autism Spectrum Disorder (ASD) refers to a neurodevelopmental issue characterized by confinements in social associations, correspondence, and conduct that become progressively regular [1]. The causes of ASD have been connected to hereditary and neurological factors; however, they are fundamentally analyzed by utilizing non-hereditary factors identified with conduct, such as social cooperation, play, creative thinking, monotonous practices, and correspondence, among others [2]. Existing estimates reveal that approximately 1.5% of the population is on the range, and many persons on the range are believed to remain undetected [3]. Accordingly, need of quick analyzing services corresponding with the developing awareness of ASD [4]. Wall et al. proposed numerous data mining techniques in a precise decision-tree algorithm (ADTree) to moderate the count of items present in the ADOS-Revised test. The intention of this work was to hasten ASD diagnosis so that private members, including family, can utilize the necessary services provided. To accomplish this goal, the authors removed instances of non ASD cases and then investigated the classification frameworks produced by the ADTree calculation on an imbalanced dataset. The Waikato Environment for Knowledge Analysis platform was subsequently utilized to evaluate the classification accuracy obtained by using the ADTree algorithm. After examining the results of the ADTree calculation, the authors found that, among the 29 items included in the ADOS-Revised test, only 8 features appear in the classification framework; thus, the group believed that the 29 items could, in fact, be represented by only these 8 items. There is a necessity to reconsider the features includes within ASD diagnostic tool to satisfy a smaller number of items sets while keeping up the sensitivity and validity of the test. [5][6]

ASD prediction-based Machine Learning (ML) requires cautious examination, particularly when managing diagnostic strategies employing techniques in the clinical setting. Limiting the ADOS-Revised test to eight items may result in misleading results because exercises must be directed by the clinician on an experiment before the grouping [6,8]. Duda et al. [7] conducted a realistic investigation associating numerous intelligent algorithms to differentiate between ASD and attention deficit hyperactivity disorder (ADHD). Six methods were differentiated on a dataset with 65 items obtained from the Simons Simplex Collection version 15.41. Information was gathered by utilizing a parent-directed survey symptomatic strategy called the Social Responsiveness Scale. A preprocessing stage was conducted by the author to (1) dispose of occurrences that had at least four missing qualities, (2) balance data collection by using the under-sampling procedure, and (3) diminish information dimensionality by using feature selection strategies. Chu et al. [9] explored several approaches to separate ADHD and obstructive sleep apnea (OSA) by using the data of 217 children who had been diagnosed as having ADHD, OSA, or a mixture of ADHD and OSA as per the Diagnostic and Statistical Manual of Mental Disorders (fourth edition; DSM IV) standards. Information was gathered by utilizing a diverse diagnostic tool, and three ML techniques were used to infer classifiers that could help clinicians and doctors improve diagnostic criteria. Detailed outcomes demonstrated that 17 highlights show significant distinctions among three classes of pervasive developmental disorders (PDDs), especially in the Child Behavior Checklist (CBCL). Moreover, compared with the neural network and CHAID algorithm, the decision tree generated classifiers faster.

Vu Viet Nguyen et al. [36] introduced novice advanced machine learning method, specifically Adaptive Neuro Fuzzy Inference System optimized by Particle Swarm Optimization (PSOANFIS), Artificial Neural Networks optimized by Particle Swarm Optimization (PSOANN) & Best First Decision Trees based Rotation Forest (RFBFDT) for landslide spatial expectation. Le Hoang Son et al. [37] focused on the latest advancement over for machine learning for big data analytic & various systems with regards to current computing for different cultural applications. Chatterjee [38] attempted to give a sensible progressively significant comprehension about the IoT in BD structure nearby its various issues & difficulties & focused on giving possible strategies by ML technique. Abhishek Kumar [39] introduced an advanced technique for phishing identification consolidating feature extraction & categorization of the mails utilizing SVM. SH Kok et al. [40] tried to break down ongoing explores in IDS utilizing ML approach with explicit enthusiasm for dataset, ML calculations & metrics.

Wolfers et al. [10] researched issues identified with PDDs, including small sample sizes, external legitimacy, and ML algorithmic difficulties, without focusing on ASD. Lopez Marcano [11] inspected the appropriateness of various algorithms, for example, neural system and decision-tree strategies (i.e., random forest), to minimize the time required for ASD diagnosis. Maenner et al. [12] examined the random forest algorithm on a dataset obtained from the Georgia Autism and Developmental Disabilities Monitoring Network using expressions and

words acquired in youngsters' formative assessments. The dataset comprised 5,396 assessments for 1,162 offspring, 601 of whom were on the range. The random forest classifiers were assessed on an autonomous test informational collection containing 9,811 assessments of 1,450 youngsters. The outcomes revealed that random forest achieves approximately 89% predictive ability and 84% sensitivity. Thabtah dissected limitations related to testing reads that embraced ML for ASD classification [13] [14] [15].

Table 1: Comparative Analysis of Existing Works

Sl. No.	Year	Advantages	Limitations
1	2020 [31]	A parent-completed rating size of behavior inflexibility (BI) for youngsters with formative inabilities was built by utilizing a multistep procedure.	Studies need to analyze the focalized and unique legitimacy of the Behavioral Inflexibility Scale (BIS) using a multi-method approach.
2	2020 [32]	An unsupervised online learning model was built for ASD grouping.	Models must be prepared by using the dataset, rather than simply employing a pre-trained model.
3	2020 [33]	Utilizing the Stockholm Youth Cohort, authors analyzed anxiety disorder among mentally imbalanced adults (n = 4,049) with and without scholarly inability against a population control (n = 217,645).	More research is necessary to determine the causes of anxiety among individuals with ASD. Future research is expected improve the understanding of the phenomenology of anxiety disorders and enhance methods to estimate and treat anxiety.
4	2019 [34]	Gaussian mixed models and hierarchical clustering were applied to distinguish among social phenotypes of ASD and assess treatment reactions over scholarly phenotypes.	A limitation of the present investigation is the absence of information from institutionalized appraisals.
5	2018 [35]	Ongoing investigations on mental imbalance were examined. This work not only articulated previously mentioned issues but also suggested ways to improve AI use in ASD in terms of conceptualization, execution, and information.	No implementation work was shown.

3. Proposed Method

3.1 Workflow

The data available in the UCI repository were obtained for our work and collected with the help of a mobile application (hereinafter referred to as an app) developed to perform four ASD screening methods, namely, autism spectrum quotient (AQ)-Adult, AQ-Adolescent, AQ-Child, and AQ-Toddler. The dataset available in the UCI repository includes clean data without missing values. Since the dataset has approximately 21 features and 1 class labeled autism and non-autism. The features age, gender, born with jaundice, family member with ASD, questions A1–A10, and ASD score from the application were used to classify the work as autism or non-autism. Principal components were retrieved by using a principal component analysis (PCA)-based algorithm and applied to the minimized dataset. We considered five eigenvectors from the given data and cooccurrence matrices and then fed the system to different classifier algorithms with a cross-fold value of 10. The system was classified by using the different algorithms, and the best algorithm for early diagnosis of ASD was identified by using precision, recall, F1 score, and accuracy values. Figure 1 illustrates the proposed workflow.

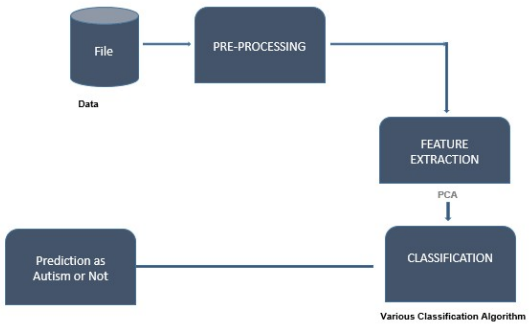

Figure 1: Workflow

3.2 Data Collection & its Description

Four classes include data adolescent, autism data adult, autism child process and toddler. The dataset included the following attributes: age, sex, ethnicity, jaundice, family ASD, residence, previous app use, screening, language, and classes. The screening test was conducted among age groups of 4–11 years, 12–16 years, and 17 years and older. Upon completion of the test by the user (questions A1–A10), a screen appeared so that the user can review and modify his/her responses. The screen serves as a form of quality assurance to enable users to verify their responses before moving to the page wherein the data are finally submitted. The value "0" or "1" is recorded based on the response given by the participants. The features and their data types are illustrated in Table 2.

Table 2: Feature Description

Feature	***Type***	***Description***
Age	Number	Toddler (months), child, adolescent, and adult (year)
Sex	String	Male or female
Ethnicity	String	List of common ethnicities in text format
Jaundice	Boolean (Yes or No)	Whether the case was born with jaundice
Family_ ASD	Boolean (Yes or No)	Whether any immediate family member has an ASD
Who completed the test	String	Parent, self, caregiver, medical staff, clinician, etc.
Residence	String	List of countries in text format
Previous app use	Boolean (Yes or No)	Whether taken screening test
Screening method type (A1–A10)	Binary (0,1)	Question method type
Score	Integer	Values generated based on conditions
Screening type	Integer	Age of the individual
Language	String	Regional language
Class	Boolean (Yes or No)	Class description

3.3 Feature Extraction

The features were extracted by using PCA. Certain rules are associated with feature extraction, as discussed below. The main idea of PCA is to reduce the dimensionality of the dataset variables available in the given input data. PCA uses orthogonal transformation to transfer the set of possible correlated components or variables into a set of linearly uncorrelated variables called principal components. In this work, we used five principal components (PC1–PC5) derived from the set of data inputs after preprocessing. These vectors have been used as feature extraction variables for the rule-based algorithm described in our previous work [16].

The steps involved in PCA are:

1. Normalize the data.
2. Identify the covariance matrix.
3. Calculate eigenvalues and eigen vectors.
4. Choose the principal component and form the feature vector.

3.4 Classification Algorithm

According to the workflow for ASD diagnosis and prediction, the dataset is first framed, after which feature selection is conducted. The severity of autism is calculated by applying machine learning classification algorithms. After a review of their characteristics, the following supervised classification algorithms are applied.

3.4.1 Support Vector Machine (SVM)

The goal of SVM is to find a hyperplane in the N-dimensional field (N number of characteristics). Many possible hyperplanes could be selected to separate two classes of data points. We intend to find an aircraft with the greatest margin, i.e., the maximum distance between two class data points. Maximizing the margin gap offers some consolidation to improve the trustworthiness of future data points. Hyperplanes are decision limits that help categorize data points. Data points on either side of the hyperplane can be assigned to various classes. The hyperplane dimension also depends on the number of characteristics. If the input number is 2, for example, the hyperplane is only one line. A hyperplane is a two-dimensional plane if the number of features to be entered is 3. The number of features approaching 3 is difficult to

imagine. Vectors supporting the hyperplane are similar data points and influence the hyperplane position and orientation. We optimized the margin of the classifier by using these support vectors. Elimination of support vectors would change the hyperplane's location. These concepts were used to build our SVM [17] [18].

3.4.2 K-Nearest Neighbor (KNN)

KNN is a data mining algorithm used for classification. The steps involved in KNN are as follows. (1) Obtain the unclassified data, (2) evaluate the distance from new data to all other already categorized (Euclidian, Manhattan, Minkowski, or weighted) data, (3) calculate k value, (4) Review the list of classes at the minimum distance, counting the number of every appearing class, (5) selection of the class that occurs most often as the right one, and (6) classify actual data with the class obtained in (5). The distance between two points can be easily calculated using several formulas [19] [20]. The formula for the Euclidean distance is as follows:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

3.4.3 Random Forest

A decision tree is an abstract model that can be used a building block for a random forest. This paradigm is interpretable because the classifications are familiar before a decision reached (in an ideal world), how the issue affecting the data is built are the technical details of a decision tree. The decision tree in the Classification and Regression Trees (CART) algorithm is constructed by evaluating the questions (called node splits) contributing to the largest reduction in Gini impurities when answered. This means that the decision tree attempts to create nodes involving a high ratio of datasets (data points) from a single class by locating values in the attributes that split the data cleanly into classes [21] [22].

$$I_G(n) = 1 - \sum_{i=1}^{f}(p_i)^2 \quad (2)$$

3.4.4 Naïve Bayes (NB)

A classifier is a model used to distinguish between objects based on certain characteristics. The NB classifier is a deterministic prediction system model. The cluster is focused on the principle of Bayes. Finding the likelihood of A occurring as B is happening is conducted by using NB [23]. Here B is the proof, and A is the assumption. The predictions/features here are believed to be independent, i.e., one function has no impact on the other. The various kinds of NB Classifiers namely Multinomial, Bernoulli and Gaussian.

3.4.5 Adaptive Boosting (AdaBoost)

AdaBoost is a sub-algorithm used for machine learning developed by Yoav Freund and Robert Schapire, who received the Nobel Prize for their research in 2003. It can be used to enhance performance in combination with several other learning algorithms. The performance of other optimization algorithms is incorporated into a weighted sum representing the boosted classifier's overall results. While AdaBoost is prone to loud outliers and data, it is less vulnerable than most other learning algorithms to overfitting issues in several situations. AdaBoost is frequently known as the satisfactory out-of-the-field classifier. However, the pattern is introduced in every level of the AdaBoost set of rules [24][25].

3.4.6 Stochastic Gradient Descent (SGD)

Stochastic refers to a random probability-related scheme or method. Thus, in stochastic gradient descent (SGD), several samples, rather than the whole set of data for each iteration, are randomly chosen. In the descending gradient, the term batch indicates the maximum set of data from a sample used to measure the gradient of iteration. The goal of the stochastic gradient descent is to find the better way of travelling the error surface so that minimum error value is achieved quickly without resorting to brute force search, therefore it is very costly to perform computationally. SGD solves this problem because, in SGD, only one sample is used for iteration. The batch is spontaneously mixed and chosen to carry out the computation process [26].

$$for\ i\ in\ range\ (m): \theta_j - \alpha\,(\hat{y}^i - y^i)x_j^i \quad (3)$$

3.4.7 CN2 Rule

CN2 rule induction is a classification algorithm that works on rules based on a condition followed by a prediction class [27] on different datasets.

3.4.7.1 Adult

If PC1<=1.034 AND PC1>=1.034 AND PC1<=1.067 AND PC1>=1.067 AND PC1<=1.2963 AND PC1>=1.2972 AND PC1<=1.434 AND PC1>=1.434
ELSE IF PC1<=1.080 AND PC2>=0.260 AND PC2>=3.464
THEN
CLASS=NO

3.4.7.2 Adolescent

IF PC1<=-0.507 AND PC4>=-0.736 AND PC1<=-0.330 AND PC3>=- 0.782 AND PC2>=2.276 AND PC4>=1.297

THEN
CLASS=YES
ELSE IF PC1>=0.507 AND PC1>=0.507 AND PC4>=2.589
THEN
CLASS=NO

3.4.7.3 Child

IF PC1<=-0.3214 AND PC2>=0.4409 AND PC4>=0.8436 AND PC2>=0.7726 AND PC4>=0.8436
THEN
CLASS=YES
ELSE IF PC1>=0.1277 AND PC1<=-0.8413 AND PC1>=-0.8413 AND PC1<=0.563 AND PC1>=-0.5635
THEN
CLASS=NO

3.4.7.4 Toddler

IF PC1<=1.5616 AND PC5>=-1.0834 AND PC3>=2.563 AND PC4>=-0.864 AND PC3>=1.959 AND PC5>=2.781
THEN
CLASS=YES
ELSE IF PC1>=1.7457 AND PC1<=1.5118 AND PC1>=1.5118 AND PC1<=1.6124 AND PC1>=1.6124
THEN
CLASS=N

4. Result & Discussion

4.1 Performance metrics

The performance of the system was calculated based on F1 scores, precision, recall, and accuracy [28] [29] [30].

4.2 F1 Score

F1 score is the weighted normal of precision and recall. This score evaluates false positives and negatives. While it is not as straightforward as exactness, F1 score is normally more helpful than precision. Precision works best if false positives and negatives have a comparative expense. If the expense of false positives and negatives are altogether different, precision and recall may be more informative. In the adult dataset, the F1 score for AdaBoost is 0.993, which is higher than the F1 scores obtained from the other methods. In the adolescent dataset, the F1 score of random forest is 0.972, which is higher than the F1 scores obtained from the other methods. In the child dataset, the F1 score of SGD is 0.996, which is higher than the F1 scores obtained from the other methods. In the toddler dataset, the precision rate of SGD is 0.996, which is higher than the F1 scores obtained from the other methods.

$$F1\ Score = 2 * \frac{recall*precision}{recall+precision} \quad (4)$$

4.3 Precision

Precision is the proportion of effectively anticipated positive perceptions compared with all-out anticipated positive perceptions. In the adult dataset, the precision rate of SGD is 0.997, which is higher than the precision scores obtained from the other methods. In the adolescent dataset, the precision rate of random forest is 0.972, which is higher than the precision scores obtained from the other methods. In the child dataset, the precision rate of SGD is 0.996 for SGD, which is higher than the precision scores obtained from the other methods. In the toddler dataset, the precision rate of SGD is 0.996, which is higher than the precision scores obtained from the other methods.

$$precision = \frac{tp}{tp+fp} \quad (5)$$

4.4 Recall

Recall refers to the proportion of effectively anticipated positive perceptions versus all perceptions in the real class. In the adult dataset, the recall of SGD is 0.997, which is higher than the recall scores obtained from the other methods. In the adolescent dataset, the recall of random forest is 0.972, which is higher than the recall scores obtained from the other methods. In the child dataset, the recall of SGD is 0.996, which is higher than the recall scores obtained from the other methods. In the toddler dataset, the recall of AdaBoost is 0.997, which is higher than the recall scores obtained from the other methods.

$$recall = \frac{tp}{tp+fn} \quad (6)$$

Deploying the autism dataset on various machine learning algorithms provides insights into the type of algorithm yielding optimal results. Figures 2–5 provide a comparative analysis of these methods.

4.5 Classifier Accuracy

In the adult dataset, the highest accuracy (99.7%) was obtained from SGD. In the adolescent dataset, the highest accuracy (97.2%) was obtained from random forest. In the child and toddler datasets, the highest accuracies were obtained from SGD and random forest. Figure 6 shows the cumulative accuracy chart.

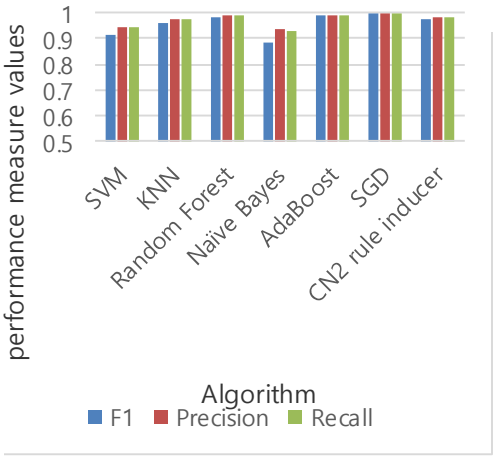

Figure 2: Performance evaluation on the adult ASD dataset

Figure 2 describes the performance values obtained for SVM, KNN, random forest, naïve Bayes, AdaBoost, SGD, and CN2 rule inducer. The precision, recall, and F1 score of each algorithm were obtained for the child ASD dataset, and the random forest algorithm yielded the highest value of 0.98.

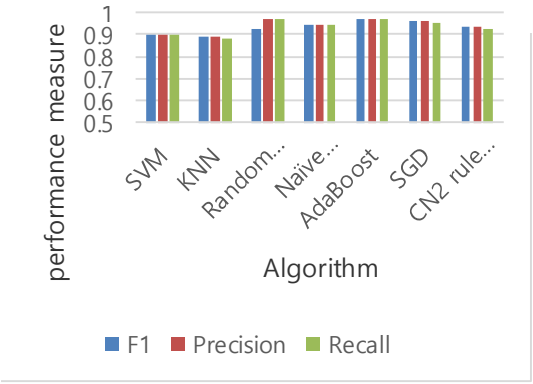

Figure 3: Performance evaluation on the Adoloscent ASD dataset

Figure 3 describes the performance values obtained for SVM, KNN, random forest, naïve Bayes, AdaBoost, SGD, and CN2 rule inducer. The precision, recall, and F1 score of each algorithm were obtained for the adolescent ASD dataset, and the random forest algorithm yielded the highest value of 0.97.

Figure 4 describes the performance values obtained for SVM, KNN, random forest, naïve Bayes, AdaBoost, SGD, and CN2 rule inducer. The precision, recall, and F1 score of each algorithm were obtained for the toddler ASD dataset, and the random forest algorithm yielded the highest value of 0.99.

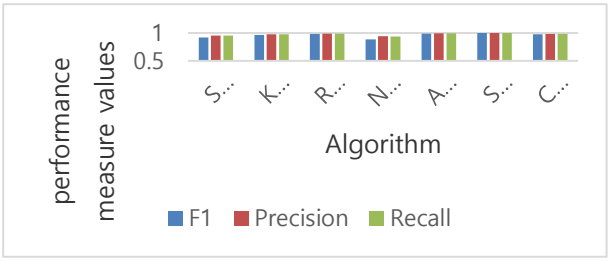

Figure 4: Performance evaluation on the Toddler ASD dataset

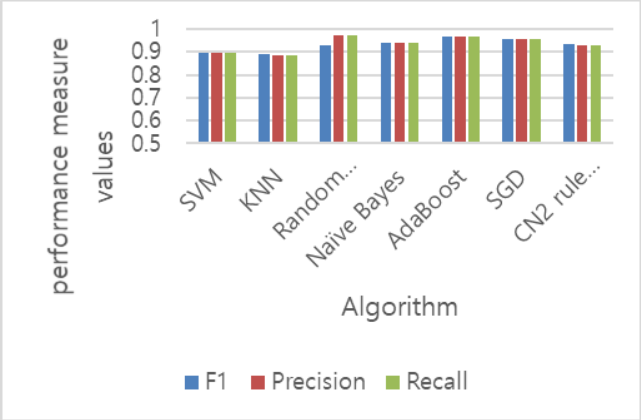

Figure 5: Performance evaluation on the Child ASD dataset

Figure 5 describes the performance values obtained for SVM, KNN, random forest, naïve Bayes, AdaBoost, SGD, and CN2 rule inducer. The precision, recall, and F1 score of each algorithm were obtained for the child ASD dataset, and the SGD algorithm yielded the highest value of 0.99.

4.6 Accuracy Value

Accuracy	*Adult ASD Data*	*Adolescent ASD Data*	*Child ASD Data*	*Toddler ASD Data*
SVM	94.6	89.5	94.1	91.7
KNN	97.4	88.3	95.9	97.8
Random Forest	98.8	97.2	98.1	99.7
Naïve Bayes	93.1	94	95.9	95.8
AdaBoost	99.3	96.8	97.9	99.8
SGD	99.7	95.6	99.6	99.7
CN2 Rule	98.4	92.7	97.5	99.3

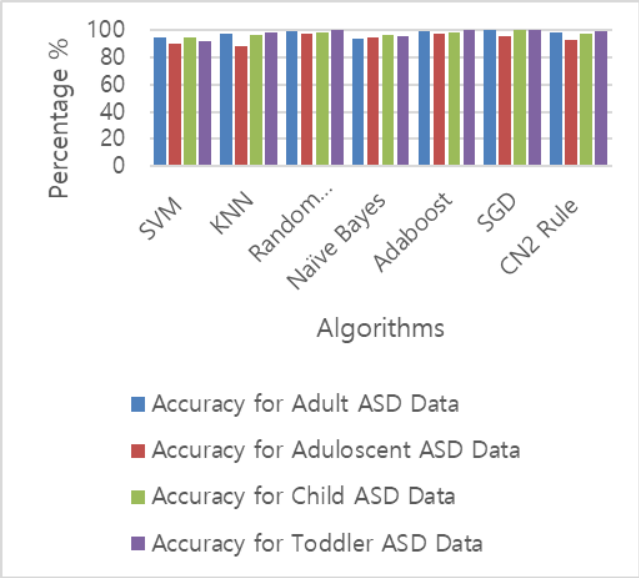

Figure 6: Comparison of various algorithm over Autism Spectrum Disorder

Figure 6 describes the performance values obtained for SVM, KNN, random forest, naïve Bayes, AdaBoost, SGD, and CN2 rule inducer. The best algorithm for the adult dataset is SGD, which yields an accuracy of 99.3%. The best algorithm for the adolescent dataset is random forest, which has an accuracy of 97.2%. The best algorithm for the child dataset is SGD, which yields an accuracy of 99.6%. Finally, the best algorithms for the toddler dataset are random forest and SGD, both of which yield 99.7% accuracy.

5. Conclusion

Awareness of ASD has rapidly increased, and several methods to diagnose and treat the condition as early as possible have been developed. Many researchers worldwide have developed screening and diagnosis methods to detect ASD and assist in its medical diagnosis. In particular, development of machine learning algorithms provides great support for the medical field. A stakeholder of these projects are patients, the caretakers who can provide the best insight about the patients, medical practitioners, psychologist, behavioral science and neuroscience.

In this work, we used several classification algorithms to make the best prediction. The dataset used a supervised classification algorithm, and the model was trained. Performance was evaluated on the basis of precision, recall, F1 score, and accuracy.

The work has shown a predominant result, and the system can be further trained with deep learning algorithms to enhance the early detection of ASD.

References

[1] Ruzich, Emily, Carrie Allison, Paula Smith, Peter Watson, Bonnie Auyeung, Howard Ring, and Simon Baron-Cohen. "Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females." Molecular autism 6, no. 1 (2015): 2.

[2] Ramaswami, Gokul, and Daniel H. Geschwind. "Genetics of autism spectrum disorder." In Handbook of clinical neurology, vol. 147, pp. 321-329. Elsevier, 2018.

[3] Brugha, Traolach S., Sally McManus, John Bankart, Fiona Scott, Susan Purdon, Jane Smith, Paul Bebbington, Rachel Jenkins, and Howard Meltzer. "Epidemiology of autism spectrum disorders in adults in the community in England." Archives of general psychiatry 68, no. 5 (2011): 459-465.

[4] Russell, Ailsa J., Clodagh M. Murphy, Ellie Wilson, Nicola Gillan, Cordelia Brown, Dene M. Robertson, Michael C. Craig et al. "The mental health of individuals referred for assessment of autism spectrum disorder in adulthood: a clinic report." Autism 20, no. 5 (2016): 623-627.

[5] Levy, Sebastien, Marlena Duda, Nick Haber, and Dennis P. Wall. "Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism." Molecular autism 8, no. 1 (2017): 65.

[6] Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal 15 (2017): 104-116.

[7] Duda, M., R. Ma, N. Haber, and D. P. Wall. "Use of machine learning for behavioral distinction of autism and ADHD." Translational psychiatry 6, no. 2 (2016): e732.

[8] Khabbaz, Amir H., Ali A. Pouyan, Mansoor Fateh, and Vahid Abolghasemi. "An adaptive RL Based fuzzy game for autistic children." In 2017 Artificial Intelligence and Signal Processing Conference (AISP), pp. 47-52. IEEE, 2017.

[9] Chu, Kuo-Chung, Hsin-Jou Huang, and Yu-Shu Huang. "Machine learning approach for distinction of ADHD and OSA." In 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp. 1044-1049. IEEE, 2016.

[10] Wolfers, Thomas, Jan K. Buitelaar, Christian F. Beckmann, Barbara Franke, and Andre F. Marquand. "From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics." Neuroscience & Biobehavioral Reviews 57 (2015): 328-349.

[11] Lopez Marcano, Juan L. "Classification of ADHD and non-ADHD using AR models and machine learning algorithms." PhD diss., Virginia Tech, 2016.

[12] Maenner, Matthew J., Marshalyn Yeargin-Allsopp, Kim Van Naarden Braun, Deborah L. Christensen, and Laura A. Schieve. "Development of a machine learning algorithm for the surveillance of autism spectrum disorder." PloS one 11, no. 12 (2016): e0168224.

[13] Thabtah, Fadi, Firuz Kamalov, and Khairan Rajab. "A new computational intelligence approach to detect autistic features for autism screening." International journal of medical informatics 117 (2018): 112-124.
[14] Thabtah, Fadi. "An accessible and efficient autism screening method for behavioural data and predictive analyses." Health informatics journal 25, no. 4 (2019): 1739-1755.
[15] Thabtah, Fadi, and David Peebles. "A new machine learning model based on induction of rules for autism detection." Health informatics journal (2019): 1460458218824711.
[16] Song, Fengxi, Zhongwei Guo, and Dayong Mei. "Feature selection using principal component analysis." In 2010 international conference on system science, engineering design and manufacturing informatization, vol. 1, pp. 27-30. IEEE, 2010.
[17] Gholami, Raoof, and Nikoo Fakhari. "Support vector machine: principles, parameters, and applications." In Handbook of Neural Computation, pp. 515-535. Academic Press, 2017.
[18] Huang, Shujun, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. "Applications of support vector machine (SVM) learning in cancer genomics." Cancer Genomics-Proteomics 15, no. 1 (2018): 41-51.
[19] Dey, Lopamudra, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, and Sweta Tiwari. "Sentiment analysis of review datasets using naive bayes and k-nn classifier." arXiv preprint arXiv:1610.09982 (2016).
[20] Gök, Murat. "An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease." International Journal of Systems Science 46, no. 6 (2015): 1108-1112.
[21] Chen, Wei, Xiaoshen Xie, Jiale Wang, Biswajeet Pradhan, Haoyuan Hong, Dieu Tien Bui, Zhao Duan, and Jianquan Ma. "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility." Catena 151 (2017): 147-160.
[22] Sun, Guanglu, Shaobo Li, Yanzhen Cao, and Fei Lang. "Cervical cancer diagnosis based on random forest." International Journal of Performability Engineering 13, no. 4 (2017): 446-457.
[23] Gao, Chong-zhi, Qiong Cheng, Pei He, Willy Susilo, and Jin Li. "Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack." Information Sciences 444 (2018): 72-88.
[24] Nayak, Deepak Ranjan, Ratnakar Dash, and Banshidhar Majhi. "Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests." Neurocomputing 177 (2016): 188-197.
[25] Wyner, Abraham J., Matthew Olson, Justin Bleich, and David Mease. "Explaining the success of adaboost and random forests as interpolating classifiers." The Journal of Machine Learning Research 18, no. 1 (2017): 1558-1590.
[26] Zou, Difan, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Stochastic gradient descent optimizes over-parameterized deep relu networks." arXiv preprint arXiv:1811.08888 (2018).
[27] Ge, Zhiqiang, Zhihuan Song, Steven X. Ding, and Biao Huang. "Data mining and analytics in the process industry: The role of machine learning." Ieee Access 5 (2017): 20590-20616.
[28] Ye, Yanfang, Tao Li, Donald Adjeroh, and S. Sitharama Iyengar. "A survey on malware detection using data mining techniques." ACM Computing Surveys (CSUR) 50, no. 3 (2017): 1-40.
[29] Rao, K. Sreenivasa, N. Swapna, and P. Praveen Kumar. "Educational data mining for student placement prediction using machine learning algorithms." Int. J. Eng. Technol. Sci 7, no. 1.2 (2018): 43-46.
[30] Chauhan, Ritu, and Harleen Kaur. "Predictive analytics and data mining: a framework for optimizing decisions with R tool." In Business Intelligence: Concepts, Methodologies, Tools, and Applications, pp. 359-374. IGI Global, 2016.
[31] Lecavalier, Luc, James Bodfish, Clare Harrop, Allison Whitten, Desiree Jones, Jill Pritchett, Richard Faldowski, and Brian Boyd. "Development of the Behavioral Inflexibility Scale for Children with Autism Spectrum Disorder and Other Developmental Disabilities." Autism Research (2020).
[32] Liang, Shuaibing, Chu Kiong Loo, and Aznul Qalid Md Sabri. "Autism Spectrum Disorder Classification in Videos: A Hybrid of Temporal Coherency Deep Networks and Self-organizing Dual Memory Approach." In Information Science and Applications, pp. 421-430. Springer, Singapore, 2020.
[33] Nimmo-Smith, Victoria, Hein Heuvelman, Christina Dalman, Michael Lundberg, Selma Idring, Peter Carpenter, Cecilia Magnusson, and Dheeraj Rai. "Anxiety Disorders in Adults with Autism Spectrum Disorder: A Population-Based Study." Journal of Autism and Developmental Disorders 50, no. 1 (2020): 308-318.
[34] Stevens, Elizabeth, Dennis R. Dixon, Marlena N. Novack, Doreen Granpeesheh, Tristram Smith, and Erik Linstead. "Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning." International journal of medical informatics 129 (2019): 29-36.
[35] Thabtah, Fadi. "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward." Informatics for Health and Social Care 44, no. 3 (2019): 278-297.
[36] Nguyen, Vu Viet, Binh Thai Pham, Ba Thao Vu, Indra Prakash, Sudan Jha, Himan Shahabi, Ataollah Shirzadi et al. "Hybrid machine learning approaches for landslide susceptibility modeling." Forests 10, no. 2 (2019): 157.
[37] Tripathy, Hrudaya Kumar, Biswa Ranjan Acharya, Raghvendra Kumar, and Jyotir Moy Chatterjee. "Machine learning on big data: A developmental approach on societal applications." In Big Data Processing Using Spark in Cloud, pp. 143-165. Springer, Singapore, 2019.
[38] Chatterjee, Jyotir. "IoT with Big Data Framework using Machine Learning Approach." International Journal of Machine Learning and Networked Collaborative Engineering 2, no. 02 (2018): 75-85.
[39] Kumar, Abhishek, Jyotir Moy Chatterjee, and Vicente García Díaz. "A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing." International Journal of Electrical & Computer Engineering (2088-8708) 10 (2020).

[40] Kok, S. H., A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam. "A Review of Intrusion Detection System using Machine Learning Approach." Int. J. Eng. Res. Technol 12, no. 1 (2019): 9-16.

R. Sujatha completed the Ph.D. degree in Vellore Institute of Technology, in 2017 in the area of data mining. She received her M.E. degree in computer science from Anna University in 2009 with university ninth rank and done Master of Financial Management from Pondicherry University in 2005. She received her B.E. degree in computer science from Madras University, in 2001. She has 15 years of teaching experience and has been serving as an associate professor in School of Information Technology and Engineering in Vellore Institute of Technology, Vellore. Her areas of research interest include Data mining, Machine learning, Image processing and Management of Information systems.

S.L. Aarthy completed the Ph.D. degree in Vellore Institute of Technology, in 2018 in the area of medical image processing. She received her M.E. degree in computer science from Anna University in 2010. She received her B.E. degree in computer science from Anna University, in 2007. Has 10 years of teaching experience and has been Assistant Professor (Senior) in School of Information Technology and Engineering in Vellore Institute of Technology, Vellore. Her research area includes Image processing, soft computing and data mining. She has published a good number of journal papers in her research filed. She is life member of CSI and IEEE. She is also part of various school activity committees.

Jyotir Moy Chatterjee is currently working as an Assistant Professor of IT department at Lord Buddha Education Foundation (Asia Pacific University of Technology & Innovation), Kathmandu, Nepal. Prior to this he has worked as an Assistant Professor at CSE department at GD Rungta College of Engineering & Technology (CSVTU), Bhilai, India. He has completed M. Tech from Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha and B. Tech in Computer Science & Engineering from Dr. MGR Educational & Research Institute, Chennai. He has more than 30 international publications, 2 authored books, 2 edited books & 10 book chapters into his account. His research interests include the Cloud Computing, Big Data, Privacy Preservation, Data Mining, Internet of Things, Machine Learning, Blockchain Technology. He is member of various professional societies and international conferences.

Noor Zaman received the Ph.D. degree in IT from UTP, Malaysia. He has great international exposure in academia, research, administration, and academic quality accreditation. He was with ILMA University, King Faisal University (KFU) for a decade, and currently with Taylor's University, Malaysia. He has 19 years of teaching & administrative experience. He has an intensive background of academic quality accreditation in higher education besides scientific research activities, he had worked a decade for academic accreditation, and earned ABET accreditation twice for three programs at CCSIT, King Faisal University, Saudi Arabia. He also worked for National Commission for Academic Accreditation and Assessment (NCAAA), Education Evaluation Commission Higher Education Sector (EECHES) formerly NCAAA Saudi Arabia, for institutional level accreditation. He also worked for National Computing Education Accreditation Council (NCEAC).

Dr. Noor Zaman has awarded as top reviewer 1% globally by WoS/ISI (Publons) recently for the year 2019. He has edited/authored more than 11 research books with international reputed publishers, earned several research grants, and a great number of indexed research articles on his credit. He has supervised several postgraduate students including master's and Ph.D. Dr. Jhanjhi is an Associate Editor of IEEE ACCESS, Keynote speaker for several IEEE international conferences globally, External examiner/evaluator for Ph.D. and masters for several universities, Guest editor of several reputed journals, member of the editorial board of several research journals, and active TPC member of reputed conferences around the globe.

Azween Abdullah is a professional development alumni of Stanford University and MIT and his work experience includes thirty years as an academic in institutions of higher learning and as director of research and academic affairs at two institutions of higher learning, vice-president for educational consultancy services, 15 years in commercial companies as Software Engineer, Systems Analyst and as a computer software developer and IT/MIS consultancy and training.

Mahadevan Supramaniam serves as Director, Research and Innovation Management Centre & Institute of Graduate Studies of SEGi University. Mahadevan's expertise lies in R&D development and policies, Enterprise Resource Planning, Computer Science & security system, Business Process Management and Integrated Technologies for industries. He has shared most of his experiences on his expertise area through public talks and has written many papers and books which has been published all over the world. Mahadevan holds a DBA from the Twintech International University College of Technology Malaysia and a Master's of Software Engineering degree from University Malaya.