SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

A Comparative Analysis of Machine Learning Models for Prediction of Autism Spectrum Disorder Using Screening Data

Ming Yue Yeap[1], Stephanie Chua[1,*], Arif Bramantoro[2]

[1] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia
[2] School of Computing and Informatics, Universiti Teknologi Brunei, Jalan Tungku Link, Gadong BE1410, Brunei Darussalam

ARTICLE INFO

Article history:
Received 3 November 2023
Received in revised form 18 September 2024
Accepted 4 October 2024
Available online 18 November 2024

ABSTRACT

Autism spectrum disorder (ASD) is a neurological and developmental disorder that affects how people interact with others, communicate, learn, and behave. ASD prediction is difficult because the diagnostic factors may not be based solely on observation. In this research paper, an in-depth comparative analysis of various machine learning models applied to the task of classifying autism traits was presented. Our study aimed to assess the performance of these models within the context of identifying individuals with autism based on relevant features and data. The machine learning models investigated in this study encompassed Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Naive Bayes (NB), and Neural Network (NN). The models were evaluated using six essential classification metrics: accuracy, precision, recall, specificity, F1 score, and AUC score. On the training dataset, our results reveal nuanced performance characteristics. SVM and RF excel in precision and recall, showing promise for accurate autism trait classification. KNN exhibits remarkable specificity, suggesting its potential for minimizing false positives. LR and NB demonstrate balanced performance across multiple metrics, while NN exhibits high precision and recall, albeit with higher computational demands. It was concluded that SVM was the best classification model for autism trait classification.

Keywords:

Autism spectrum disorder; Classification; Machine learning; Prediction

1. Introduction

Autism spectrum disorder (ASD) is a disability in development caused by differences in the brain [1]. People with ASD usually have problems with limited or repetitive behaviours or interests, as well as communication skills and social engagement. Although the symptoms are easy to identify, a diagnosis of autism requires skilled medical professionals to supervise behavioural assessments that are measured according to the incidence of numerous symptoms that interfere with a person's capacity to talk, play, and create communication relationships. Depending on how serious the symptoms are, ASD can range from mild to severe [2].

[*] *Corresponding author.*
E-mail address: chlstephanie@unimas.my

https://doi.org/10.37934/araset.53.1.175185

Parents who are concerned that their child may be autistic can bring them to medical practitioners. Medical practitioners will conduct screening tests on toddlers and children to diagnose if they have ASD. Many times, diagnosis cannot be determined in one visit, and it involves multiple visits to the clinic for some time, sometimes up to a few years to finally get a definitive diagnosis [3]. There are also teenagers and young adults who were not diagnosed with ASD from young and did not receive early intervention. The problem with ASD is that it is quite hard to diagnose as every child may progress through life at a different developmental speed. Sometimes, parents may also be unaware of certain ASD traits that they may think are normal in their child. Therefore, this research investigates using the machine learning approach to learn models for the classification of ASD based on past data available.

Machine learning algorithms have been used to discover hidden patterns in medical datasets to better understand diseases [4]. Similarly, machine learning algorithms were being used on datasets related to ASD to identify valuable hidden patterns and create a predictive model for diagnosing its risk [5]. The main purpose of this research is to use machine learning algorithms to learn models for the classification of autism spectrum disorder (ASD) using screening data. In this research, a significant feature set will be determined using a correlation matrix. A feature selection algorithm will also be applied to the feature set to determine the best set of features for learning the classification model. The feature set will then be used in six machine learning algorithms to learn models for the classification of ASD into 'ASD trait' and 'No ASD trait' [6]. A comparative analysis will be conducted to determine the best classification model.

The scope of the study covers the prediction of ASD in adults who are 18 years and older. Six prediction models are built to be compared. The machine learning algorithms used for building the models are Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, and Neural Network.

The study aims to address the gap in the literature on the use of machine learning models for predicting autism spectrum disorder (ASD) using screening data. The previous studies have explored the use of machine learning techniques for ASD diagnosis, most of these studies have used clinical data, which may not be readily available in all settings. In contrast, we propose the use of screening data, which is more widely available and can be used for early detection and diagnosis of ASD. This research holds significant implications for the field of ASD diagnosis. The outcomes of this study have the potential to greatly benefit medical practitioners in their future analyses of ASD. By shedding new light on the diagnosis of ASD, this research will provide valuable insights that can enhance the understanding and treatment of individuals on the autism spectrum. The findings from this study will be instrumental in supporting mental health organizations' efforts to increase awareness and knowledge surrounding various concerns related to ASD. Furthermore, the comprehensive analysis presented in this research will yield valuable information for future studies aimed at developing improved computational methods of diagnosing ASD.

2. Related Works

Many researchers have worked on autism spectrum disorder (ASD) prediction using the machine learning approach. In particular, they adopted popular machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours, Naïve Bayes, and Neural Networks.

Erkan and Thanh [7] used K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forests (RF) in the research. The research was carried out by using three datasets which are AQ-10-Adult, AQ-10-Adolescence, and AQ-10-Child of the UCI database. The models were evaluated in five experiments. The SVM method achieved 100% accuracy in all experiments with complete data. Next, the Random Forest algorithms achieved 100% accuracy in all experiments with complete data. They set the number of nearest neighbours to 3 and used the Euclidean distance for the KNN model. The KNN method achieved an accuracy of 94% to 96% in all experiments of complete data. Based on the data obtained, the RF and SVM algorithms achieved high classification scores as measured by accuracy, sensitivity, F-measure, and Area Under Curve (AUC). In their tests, they discovered that the RF approach performs better than the SVM and kNN methods for the classification of ASD data.

Jalaja *et al.,* [8] used Multilayer Perceptron, Decision Tree J48 Classifier, Naïve Bayes, and Bayesian Networks to analyse the experimental result. The dataset required for this research work was gathered from multiple questionnaires, and interviews conducted by experts from observations on ASD-affected children. The dataset was split into 90%-10%. Multilayer Perceptron algorithm had the highest classifier accuracy (95.11%), Naive Bayes Classifier (93.37%), and Bayesian Classifier (93.97%) for 10% of the training data.

Deepa and Jeen Marseline [9] worked on ASD prediction by comparing the Naive Bayes, Decision Table, and Support Vector Machine algorithms. The training data set was taken from the UCI repository. They use ten behavioural and individual characteristics that have been proven to be effective in detecting ASD cases. The algorithms used were from the Weka Tool for a comparative study. This study aimed to predict early detection and indicate some statistical value. Naïve Bayes and Support Vector Machine acquired the best results compared to the Decision Table based on the result analysis.

Next, Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Convolutional Neural Networks (CNN), K-Nearest Neighbours (KNN) and Artificial Neural Network (ANN) were used in the research of Raj and Masood [10]. The proposed algorithms were evaluated by three different ASD datasets. There were 292 instances and 21 attributes in the first dataset related to ASD screening in children. There were 21 attributes and 704 instances in total for the adult dataset, which was the second dataset. There were 104 instances and 21 attributes in the third dataset, which was focused on ASD screening in adolescent individuals. Results from the application of various machine learning techniques and the handling of missing values strongly suggest that CNN-based prediction models perform better on all these datasets, with higher accuracies of 99.53%, 98.30%, and 96.88% for Autistic Spectrum Disorder screening in data for adult, children, and adolescents respectively.

Vakadkar *et al.,* [11] used Support Vector Machines (SVM), Random Forest Classifier (RFC), Naïve Bayes (NB), Logistic Regression (LR) and K- Nearest Neighbours (KNN) to predict ASD. Their dataset had 1054 instances with 18 attributes (including the class variable). Logistic Regression was observed to give the highest accuracy of 97.15%.

Thabtah *et al.,* [12] used Logistic Regression to make predictions in their research. Their datasets were based on the AQ-10 adult and AQ-10 adolescent screening methods respectively. Each dataset consisted of over 20 variables, ten of which were associated with the screenings plus the individual's features such as age, gender, ethnicity, etc. They used information gain (IG) and Chi-square testing (CHI) to investigate the features in the ASD adult dataset. Logistic Regression was able to generate classifiers with approximately 87% sensitivity, accuracy, and specificity.

Alteneiji *et al.,* [13] employed various machine learning models, including Support Vector Machine (SVM), XgBoost, AdaBoost, CV Boosting, Neural Network, Random Forest, Naïve Bayes, and Random Forest-GBM, in their effort to predict Autism Spectrum Disorder (ASD). They utilized databases specific to distinct age groups: infants, children, and adolescents. The datasets were divided into ten behavioural questions per age group. They applied feature selection algorithms that primarily assessed the relationship between ASD test results and individual variables in the database. Two filter-based techniques, namely Chi-Squared and mutual information, were utilized. The evaluation methods employed in this study were based on the confusion matrix results of each machine learning model. The models' performance was evaluated by computing error rates, accuracy, sensitivity, and specificity. The Neural Networks model outperformed other models across all datasets, yielding superior results.

In their study, Dewi and Imah [14] employed the K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest algorithms. The Random Forest algorithm, utilizing the full set of features, demonstrated the highest accuracy in classifying Autism Spectrum Disorder (ASD) among children, achieving a perfect accuracy score of 1. When considering both specificity and sensitivity values, the Random Forest algorithm with the complete feature set emerged as the superior choice for classifying ASD in children and adolescents, outperforming other algorithms.

Saihi and Alshraideh [3] developed their predictive models for Autism Spectrum Disorder (ASD) utilizing Decision Tree C4.5, Random Forest, and Neural Network algorithms. The classification models were constructed employing a 5-fold cross-validation technique. During this process, 70% of the dataset (739 observations) were employed for training purposes, while the remaining 30% (315 observations) were designated for testing the accuracy of different classifiers. Notably, the Neural Networks model outperformed the other two models, achieving an impressive accuracy rate of 99%.

Amrutha and Sumana [15] conducted research employing the Naïve Bayes (NB), K-Nearest Neighbours (KNN), and Decision Tree (DT) algorithms. They gathered their datasets from kaggle.com, specifically focusing on children with ASD aged between 1 and 5 years. Notably, among the algorithms utilized in their study, the Decision Tree achieved the highest performance with an accuracy of 100%. In a similar work, Zheng *et al.,* [16] achieved notable results, attaining a 97% average precision and recall rate F1 score for their Logistic Regression model.

Abdulrazzaq *et al.,* [17] conducted research primarily focused on data pre-processing tasks, including addressing missing data gaps, converting categorical data into numerical format, and performing data normalization. Subsequently, the features underwent clustering through k-means and x-means clustering techniques. The researchers then employed artificial neural networks and a robust linguistic neuro-fuzzy classifier for classification purposes. The results indicated that, in terms of estimation accuracy, the classification methods outperformed the clustering methods in the context of ASD data for children.

Jebapriya *et al.,* [18] employed machine learning to identify a set of conditions that collectively prove predictive of ASD, offering valuable insights to physicians and facilitating early detection. The objective was to pinpoint predictive conditions, enabling physicians to conduct comprehensive formal ASD screenings. Utilizing complex network parameters, discriminant analysis, and support vector classifiers, they were able to achieve a maximum accuracy of 94.7% with four features and a second-order polynomial kernel in SVM. The study sought to delineate the autism spectrum distribution using supervised learning methods, while future work would explore deep learning techniques for automated recognition of social, motor, and communication behaviours across subjects.

ASD prediction through the machine learning approach had been explored by various researchers, utilizing various algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours, Naïve Bayes, and Neural Networks. There was no definite conclusion on which algorithm worked best. Overall, the Neural Networks and Random Forest algorithms appear to offer promising results, but the most suitable algorithm still depends on the specific experimental context and datasets used.

3. Methodology

The methodology used in this research is discussed in this section. It follows closely the processes in the Knowledge Discovery in Database (KDD) methodology [19]: data selection, data preprocessing, data mining and model evaluation. The performances of six machine learning algorithms: Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours, Naïve Bayes and Neural Networks were compared. The best model will be identified after the comparative analysis of the predictive models.

3.1 Data Selection

In the data selection process, a publicly available dataset on Autism was obtained from Kaggle [20]. This dataset was used to train the machine learning models. Additional data were collected to augment the dataset obtained from Kaggle. The data collected were based on the Autism Spectrum Quotient 10 items (AQ-10) (Adult) from the ASD Tests App [6]. The data were collected from 83 adults. This collected data would be used as real-life unseen data for testing the models learned. The summary of the features used for data collection is shown in Table 1.

Table 1
Features collected, their descriptions and mapping to the AQ-10 questionnaire

Feature	Type	Description
Age	Number	Adults (year), that is, 17 years+
Gender	String	Male or female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician and so on
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening method type	Integer (0, 1, 2, 3)	The type of screening methods chosen based on age category (0=Toddler, 1=Child, 2=Adolescent, 3=Adult). In this case, only adult data have been used
A1	Binary (0, 1)	The answer code of: I often notice small sounds when others do not
A2	Binary (0, 1)	The answer code of: I usually concentrate more on the whole picture rather than the small details
A3	Binary (0, 1)	The answer code of: I find it easy to do more than one thing at once
A4	Binary (0, 1)	The answer code of: If there is an interruption, I can switch back to what I was doing very quickly
A5	Binary (0, 1)	The answer code of: I find it easy to “read between the lines” when someone is talking to me
A6	Binary (0, 1)	The answer code of: I know how to tell if someone listening to me is getting bored
A7	Binary (0, 1)	The answer code of: When I’m reading a story, I find it difficult to work out the character’s intentions
A8	Binary (0, 1)	The answer code of: I like to collect information about categories of things (e.g. types of cars, types of bird, types of train and types of plant)
A9	Binary (0, 1)	The answer code of: I find it easy to work out what someone is thinking of feeling just by looking at their face
A10	Binary (0, 1)	The answer code of: I find it difficult to work out people’s intentions

ASD score	Integer	The final score was obtained based on the scoring function of on AQ-10-Adult. This was computed in an automated manner
Class label	Boolean	The decision of the screening is based on the scoring score of the AQ-10-Adult method. Possible values “0” (No ASD traits) or “1” (ASD traits)

These features are the same features contained in the dataset obtained from Kaggle. Table 2 shows the statistics of the datasets used in this research.

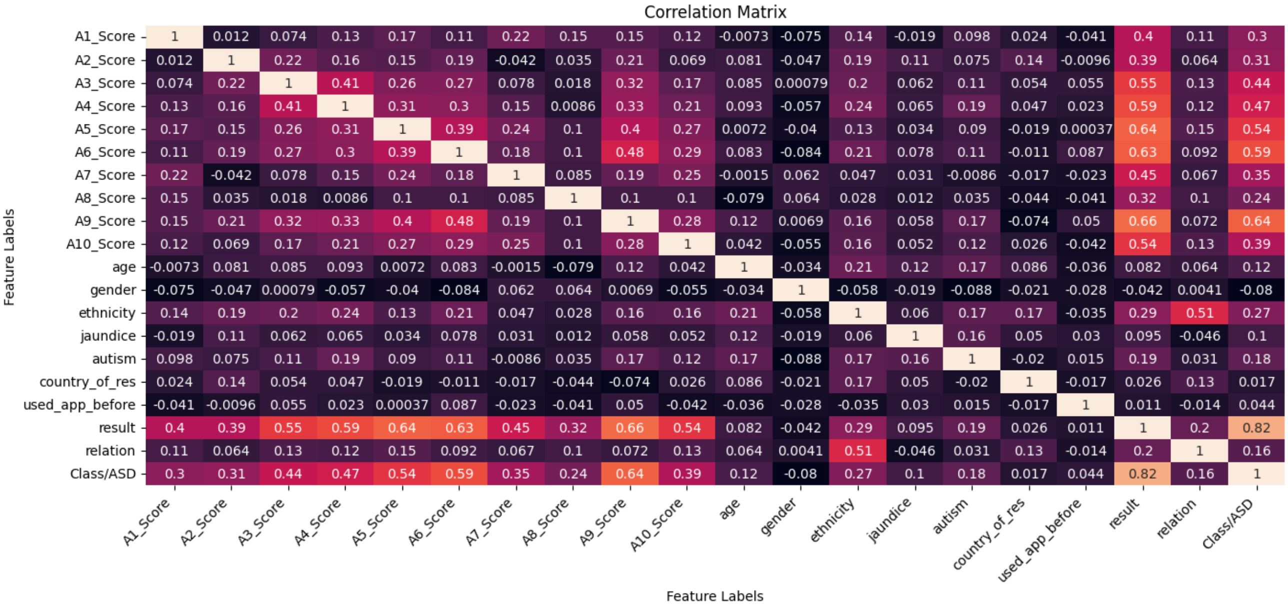
Table 2
Statistics of the Datasets Used

Dataset	Count	Total
Kaggle (for training)	189 ASD traits 515 No ASD traits	704
Collected Data (for testing)	20 ASD traits 63 No ASD traits	83

3.2 Data Preprocessing

In data preprocessing, feature selection was performed to determine suitable features that can be used for predicting ASD traits. A correlation matrix is used to determine a significant feature set, followed by feature selection to enhance model performance by removing less useful features. The feature importance approach, implemented with the ExtraTreeClassifier in scikit-learn, aided in choosing important features. Besides that, information gain and mutual information were utilized to identify the most informative features. The selected features were then used to build a model for predicting the target variables: "ASD traits" and "No ASD traits".

Figure 1 shows the correlation matrix for the features in the dataset while Table 3 shows the comparison of results for the feature selection methods used. From the correlation analysis and the feature selection results analysis, 15 features were selected out of 19 to be used for machine learning. The selected features are ethnicity, jaundice, autism, relation, country_of_res, and the A1 to A10 scores.



	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	age	gender	ethnicity	jaundice	autism	country_of_res	used_app_before	result	relation	Class/ASD
A1_Score	1	0.012	0.074	0.13	0.17	0.11	0.22	0.15	0.15	0.12	-0.0073	-0.075	0.14	-0.019	0.098	0.024	-0.041	0.4	0.11	0.3
A2_Score	0.012	1	0.22	0.16	0.15	0.19	-0.042	0.035	0.21	0.069	0.081	-0.047	0.19	0.11	0.075	0.14	-0.0096	0.39	0.064	0.31
A3_Score	0.074	0.22	1	0.41	0.26	0.27	0.078	0.018	0.32	0.17	0.085	0.00079	0.2	0.062	0.11	0.054	0.055	0.55	0.13	0.44
A4_Score	0.13	0.16	0.41	1	0.31	0.3	0.15	0.0086	0.33	0.21	0.093	-0.057	0.24	0.065	0.19	0.047	0.023	0.59	0.12	0.47
A5_Score	0.17	0.15	0.26	0.31	1	0.39	0.24	0.1	0.4	0.27	0.0072	-0.04	0.13	0.034	0.09	-0.019	0.00037	0.64	0.15	0.54
A6_Score	0.11	0.19	0.27	0.3	0.39	1	0.18	0.1	0.48	0.29	0.083	-0.084	0.21	0.078	0.11	-0.011	0.087	0.63	0.092	0.59
A7_Score	0.22	-0.042	0.078	0.15	0.24	0.18	1	0.085	0.19	0.25	-0.0015	0.062	0.047	0.031	-0.0086	-0.017	-0.023	0.45	0.067	0.35
A8_Score	0.15	0.035	0.018	0.0086	0.1	0.1	0.085	1	0.1	0.1	-0.079	0.064	0.028	0.012	0.035	-0.044	-0.041	0.32	0.1	0.24
A9_Score	0.15	0.21	0.32	0.33	0.4	0.48	0.19	0.1	1	0.28	0.12	0.0069	0.16	0.058	0.17	-0.074	0.05	0.66	0.072	0.64
A10_Score	0.12	0.069	0.17	0.21	0.27	0.29	0.25	0.1	0.28	1	0.042	-0.055	0.16	0.052	0.12	0.026	-0.042	0.54	0.13	0.39
age	-0.0073	0.081	0.085	0.093	0.0072	0.083	-0.0015	-0.079	0.12	0.042	1	-0.034	0.21	0.12	0.17	0.086	-0.036	0.082	0.064	0.12
gender	-0.075	-0.047	0.00079	-0.057	-0.04	-0.084	0.062	0.064	0.0069	-0.055	-0.034	1	-0.058	-0.019	-0.088	-0.021	-0.028	-0.042	0.0041	-0.08
ethnicity	0.14	0.19	0.2	0.24	0.13	0.21	0.047	0.028	0.16	0.16	0.21	-0.058	1	0.06	0.17	0.17	-0.035	0.29	0.51	0.27
jaundice	-0.019	0.11	0.062	0.065	0.034	0.078	0.031	0.012	0.058	0.052	0.12	-0.019	0.06	1	0.16	0.05	0.03	0.095	-0.046	0.1
autism	0.098	0.075	0.11	0.19	0.09	0.11	-0.0086	0.035	0.17	0.12	0.17	-0.088	0.17	0.16	1	-0.02	0.015	0.19	0.031	0.18
country_of_res	0.024	0.14	0.054	0.047	-0.019	-0.011	-0.017	-0.044	-0.074	0.026	0.086	-0.021	0.17	0.05	-0.02	1	-0.017	0.026	0.13	0.017
used_app_before	-0.041	-0.0096	0.055	0.023	0.00037	0.087	-0.023	-0.041	0.05	-0.042	-0.036	-0.028	-0.035	0.03	0.015	-0.017	1	0.011	-0.014	0.044
result	0.4	0.39	0.55	0.59	0.64	0.63	0.45	0.32	0.66	0.54	0.082	-0.042	0.29	0.095	0.19	0.026	0.011	1	0.2	0.82
relation	0.11	0.064	0.13	0.12	0.15	0.092	0.067	0.1	0.072	0.13	0.064	0.0041	0.51	-0.046	0.031	0.13	-0.014	0.2	1	0.16
Class/ASD	0.3	0.31	0.44	0.47	0.54	0.59	0.35	0.24	0.64	0.39	0.12	-0.08	0.27	0.1	0.18	0.017	0.044	0.82	0.16	1

Fig. 1. Correlation matrix for the features in the dataset

Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) [21] was used to address the class imbalance issue in the Kaggle dataset that will be used for training the models.

Table 3
Comparison of results for the feature selection methods used

Rank	Feature Importance	Information Gain	Mutual Information
1	A9__Score	result	result
2	result	A9__Score	A9__Score
3	A6__Score	A6__Score	A5__Score
4	A4__Score	A5__Score	A6__Score
5	A5__Score	country_of_res	A4__Score
6	A7__Score	A4__Score	A10__Score
7	A1__Score	A3__Score	A3__Score
8	A10__Score	A10__Score	A2__Score
9	A3__Score	ethnicity	A7__Score
10	age	A7__Score	A1__Score
11	country_of_res	A1__Score	autism
12	A8__Score	age	ethnicity
13	A2__Score	A2__Score	country_of_res
14	ethnicity	A8__Score	jaundice
15	jaundice	relation	A8__Score
16	gender	autism	relation
17	relation	jaundice	used_app_before
18	autism	gender	age
19	used_app_before	used_app_before	gender

SMOTE worked by oversampling the minority class (“ASD traits” instances), resulting in a balanced dataset with an equal number of instances for both "ASD traits" and "No ASD traits" classifications, thereby improving modelling and analysis accuracy. Table 4 shows the statistics for the Kaggle dataset before and after SMOTE.

Table 4
Statistics of the Kaggle dataset before and after SMOTE

Kaggle Dataset	Count	Total
Before SMOTE	189 ASD traits 515 No ASD traits	704
After SMOTE	515 ASD traits 515 No ASD traits	1030

3.3 Data Mining

The Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Neural Networks (NN) algorithms were used to learn the model using the Python Scikit-Learn module. The models were built with 10-fold cross-validation using GridSearchCV in every fold to obtain the best parameters for each fold, using the LogisticRegression for LR, RandomForestRegressor for RF, SVC for SVM, KNeighborsClassifier for KNN, GaussianNB for NB, and MLPClassifier for NN. The paths of the machine learning algorithms in Scikit-Learn are shown in Table 5.

Table 5
Path of machine learning algorithms in Scikit-Learn

Classifiers	Paths
LR	sklearn.linear_model.LogisticRegression
RF	sklearn.ensemble.RandomForestClassifier
SVM	sklearn.svm.SVC
KNN	sklearn.neighbors.KNeighborsClassifier
NB	sklearn.naive_bayes.GaussianNB
NN	sklearn.neural_network.MLPClassifier

3.4 Model Evaluation

The six machine learning models learned were evaluated using evaluation metrics like accuracy, precision, recall, specificity, F_1 score, and area under the curve (AUC) score. These metrics were derived from the confusion matrix. The best model will be determined from the comparative analysis. Table 6 shows the confusion matrix with a description of ASD traits.

Table 6
Confusion matrix with description for ASD traits

		Predicted	
		Positive (ASD traits)	Negative (No ASD traits)
Actual	Positive (ASD traits)	ASD traits	Wrongly predicted “No ASD traits”
	Negative (No ASD traits)	Wrongly predicted “ASD traits”	No ASD traits

4. Results and Discussion

Table 7 presents a summary of the training set results. The LR model provided a solid overall performance with good balance between precision and recall. It achieved high accuracy, indicating that it correctly classified the majority of instances. The AUC score suggested that the model's ROC curve was well above random chance. However, it might not be the best choice if very high precision or recall was prioritized, as other models perform better in these specific aspects. The RF model excelled in terms of precision and recall, both at 97.09%, which means it was excellent at correctly classifying positive instances while minimizing false positives and false negatives. However, it was essential to be cautious about potential overfitting, as the model might have memorized the training data. The SVM model demonstrated outstanding precision and recall, making it highly reliable for tasks where minimizing both false positives and false negatives was crucial. The high AUC score indicated excellent overall performance. The SVM model, however, might be computationally intensive and might require careful tuning of hyperparameters. The KNN model achieved near-perfect recall and specificity, making it great for identifying true positives and avoiding false negatives. However, it is important to note that KNN is sensitive to the choice of k and can be computationally expensive, especially with large datasets. The NB model provided a balanced performance with reasonable accuracy, precision, and recall. It was computationally efficient and worked well with text data and high-dimensional datasets but might not be the best choice for complex data with strong dependencies. The NN model achieved high precision and recall, making it suitable for tasks requiring a good balance between these metrics. However, it is computationally expensive and may require significant data preprocessing and tuning to perform at its best.

Table 7
Summary of training set results

Models	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	AUC score (%)
LR	93.01	93.60	93.40	93.42	93.59	93.59
RF	93.50	97.09	97.09	97.09	97.09	97.09
SVM	94.85	99.70	99.61	99.61	99.71	99.71
KNN	91.94	99.90	100.00	100.00	99.90	99.90
NB	91.75	92.02	93.59	93.40	92.14	92.14
NN	93.88	98.84	98.45	98.46	98.83	94.85

Table 8 shows the summary of the test set results. The LR model achieved perfect scores across all metrics, indicating that it performed flawlessly on the test set. Such performance suggested that the LR model could generalize well on unseen data. The RF model maintained high accuracy and specificity but experienced a significant drop in recall compared to the training set. This suggested that the RF model might not generalize well to unseen data, potentially due to overfitting. The precision remained perfect, which could be a sign of imbalanced classes. Similar to the LR model, the SVM model performed perfectly on all metrics for the test set. Again, this showed that the SVM model was able to generalize well on unseen data. The KNN model exhibited a perfect performance on all metrics, mirroring the training set results. The NB model stood out as the model with significantly lower accuracy and precision on the test set, compared to the training set. While it maintained perfect recall, its specificity was lower. This suggested that the NB model might not be well-suited for this specific test data, potentially due to differences in data distribution. The NN model performed flawlessly on all metrics for the test set, similar to its performance on the training set.

The performance discrepancies between the training and test sets highlighted the importance of evaluating models on diverse datasets. Notably, certain models, including LR, SVM, KNN, and NN, demonstrated perfect performance on the test set, indicating their ability to generalize and make accurate predictions on unseen data. Conversely, the NB model struggled with certain instances, resulting in lower precision and accuracy. These findings offered valuable insights into the strengths and limitations of the models in predicting ASD using screening data, emphasizing the need for appropriate model selection based on dataset characteristics and desired performance metrics.

Table 8
Summary of test set results

Models	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	AUC score (%)
LR	100.00	100.00	100.00	100.00	100.00	100.00
RF	97.59	100.00	90.00	100.00	94.74	95.00
SVM	100.00	100.00	100.00	100.00	100.00	100.00
KNN	100.00	100.00	100.00	100.00	100.00	100.00
NB	78.31	52.63	100.00	71.43	68.97	85.71
NN	100.00	100.00	100.00	100.00	100.00	100.00

4. Conclusion

In conclusion, the comparative study of machine learning models for the prediction of ASD using screening data has shed light on the effectiveness of different approaches. The findings highlighted the superior performance of the SVM model, which consistently achieved near-perfect accuracy, recall, specificity, precision, F1 score, and AUC score. To advance the field of ASD prediction, future work should focus on addressing these limitations by incorporating larger and diverse datasets, exploring advanced feature selection techniques, investigating alternative modelling approaches, and considering additional factors that contribute to ASD prediction. By further developing and refining the predictive models, it is possible to enhance the early detection and intervention of autism spectrum disorder, leading to improved outcomes for individuals on the autism spectrum.

Acknowledgment

The authors would like to acknowledge the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS) for the research opportunity and financial assistance.

References

[1] Centers for Disease Control and Prevention. "Autism Spectrum Disorder (ASD)."

[2] Hodges, Holly, Casey Fealko, and Neelkamal Soares. "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation." *Translational pediatrics* 9, no. Suppl 1 (2020): S55. https://doi.org/10.21037/tp.2019.09.09

[3] Saihi, Afef, and Hussam Alshraideh. "Development of an autism screening classification model for toddlers." *arXiv preprint arXiv:2110.01410* (2021). https://doi.org/10.5121/csit.2021.111508

[4] Abdulkadium, Ahmed Mahdi, Raid Abd Alreda Shekan, and Haitham Ali Hussain. "Application of Data Mining and Knowledge Discovery in Medical Databases." *Webology* 19, no. 1 (2022): 4912-4924. https://doi.org/10.14704/WEB/V19I1/WEB19329

[5] Victoire, T. Amalraj, A. Ramalingam, A. Naresh, K. M. Nasimudeen, and MS JAYA Kumar. "An Efficient Approach to Detect Autism in Child Using Machine Learning and Deep Learning." *Journal of Theoretical and Applied Information Technology* 99, no. 20 (2021): 4759-4769.

[6] Thabtah, Fadi. "An accessible and efficient autism screening method for behavioural data and predictive analyses." *Health informatics journal* 25, no. 4 (2019): 1739-1755. https://doi.org/10.1177/1460458218796636

[7] Erkan, Uğur, and Dang NH Thanh. "Autism spectrum disorder detection with machine learning methods." *Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews* 15, no. 4 (2019): 297-308. https://doi.org/10.2174/2666082215666191111121115

[8] Jayalakshmi, V. Jalaja, V. Geetha, and R. Vivek. "Classification of autism spectrum disorder data using machine learning techniques." *International Journal of Engineering and Advanced Technology (IJEAT) ISSN* 8, no. 6 (2019): 2249-8958. https://doi.org/10.35940/ijeat.F1114.0886S19

[9] Deepa, B., and KS Jeen Marseline. "Exploration of autism spectrum disorder using classification algorithms." *Procedia Computer Science* 165 (2019): 143-150. https://doi.org/10.1016/j.procs.2020.01.098

[10] Raj, Suman, and Sarfaraz Masood. "Analysis and detection of autism spectrum disorder using machine learning techniques." *Procedia Computer Science* 167 (2020): 994-1004. https://doi.org/10.1016/j.procs.2020.03.399

[11] Vakadkar, Kaushik, Diya Purkayastha, and Deepa Krishnan. "Detection of autism spectrum disorder in children using machine learning techniques." *SN computer science* 2 (2021): 1-9. https://doi.org/10.1007/s42979-021-00776-5

[12] Thabtah, Fadi, Neda Abdelhamid, and David Peebles. "A machine learning autism classification based on logistic regression analysis." *Health information science and systems* 7, no. 1 (2019): 12. https://doi.org/10.1007/s13755-019-0073-5

[13] Alteneiji, Maitha Rashid, Layla Mohammed Alqaydi, and Muhammad Usman Tariq. "Autism spectrum disorder diagnosis using optimal machine learning methods." *International Journal of Advanced Computer Science and Applications* 11, no. 9 (2020). https://doi.org/10.14569/IJACSA.2020.0110929

[14] Dewi, Erina S., and Elly M. Imah. "Comparison of machine learning algorithms for autism spectrum disorder classification." In *International joint conference on science and engineering (IJCSE 2020)*, pp. 152-159. Atlantis Press, 2020. https://doi.org/10.2991/aer.k.201124.028

[15] Amrutha, S. M., and K. R. Sumana. "Autism Spectrum Disorder Detection Using Machine Learning Techniques." *International Research Journal of Engineering and Technology (IRJET)* 8, no. 8 (2021): 1252-1254.

[16] Zheng, Yuanrui, Tingyan Deng, and Yaozheng Wang. "Autism classification based on logistic regression model." In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp. 579-582. IEEE, 2021. https://doi.org/10.1109/ICBAIE52039.2021.9389914

[17] Abdulrazzaq, Ammar Akram, Sana Sulaiman Hamid, Asaad T. Al-Douri, A. A. Mohamad, and Abdelrahman Mohamed Ibrahim. "Early detection of autism spectrum disorders (asd) with the help of data mining tools." *BioMed Research International* 2022 (2022). https://doi.org/10.1155/2022/1201129

[18] Jebapriya, S., David Shibin, Jaspher W. Kathrine, and Naveen Sundar. "Support vector machine for classification of autism spectrum disorder based on abnormal structure of corpus callosum." *International Journal of Advanced Computer Science and Applications* 10, no. 9 (2019). https://doi.org/10.14569/IJACSA.2019.0100965

[19] Soundappan, S. Jagadeesh, Rajendran Sugumar, Krzysztof J. Cios, Witold Pedrycz and Roman W. Swiniarski. “Survey on Knowledge Discovery in Database and Challenges in KDD.” (2017).

[20] Nabi, Faizun. "Autism Screening." *Kaggle: Your Machine Learning and Data Science Community*. (2022).

[21] Zhang, Aimin, Hualong Yu, Shanlin Zhou, Zhangjun Huan, and Xibei Yang. "Instance weighted SMOTE by indirectly exploring the data distribution." *Knowledge-Based Systems* 249 (2022): 108919. https://doi.org/10.1016/j.knosys.2022.108919