Proceedings of the 2nd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7048/35/20232015

Prediction of Autism Spectrum Disorder: Comparison and Tuning of Machine Learning Models

Zixuan Yang[1,a,*]

[1]Hangzhou Xuejun High School, Hangzhou, China
a. EricYang0728@163.com
**corresponding author*

Abstract: The early diagnosis in Autism Spectrum Disorder (ASD) is crucial for timely interventions to address the patients' attentional and social challenges. The currently study aims to use machine learning algorithms to accurately predict ASD outcomes. Dataset from a Kaggle competition was used to perform the prediction analysis. Five supervised machine learning algorithms were employed: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine Classifier (SVC), Random Forest (RF), and Decision Trees (DT). The models were fine-tuned using a range of possible hyperparameters and evaluated using ROC AUC scores. The best-performing model, Random Forest, achieved a training ROC AUC of 0.93. The model's performance in predicting the unseen test set resulted in a ROC AUC score of 0.8623. The outcome demonstrates the potentials of machine learning models in early prediction of ASD symptoms, which provides support for autistic individuals to enhance their quality of life and education.

Keywords: Autism Spectrum Disorder, supervised machine learning, classification, Random Forest

1. Introduction

Autism spectrum disorder (ASD), commonly referred to as autism, is a neurodevelopmental condition characterized by a diverse range of challenges, including difficulties in social interaction and communication as well as the presence of repetitive and stereotyped interests and behaviors [1]. Children diagnosed with autism often exhibit distinctive cognitive and attentional patterns, which can present substantial obstacles in their daily life and during school education. The attentional issues in autistic children frequently manifest as excessive attention to details, susceptibility to distractions, and challenges in transitioning between tasks. They tend to be drawn towards non-critical information while overlooking the broader context.

Cognitive factors and attention are closely related to the attention problems of children with autism. Sensory processing, attention allocation, and working memory are some of the cognitive factors that contribute to attention problems in autistic children. Sensory processing refers to the way the brain receives, interprets, and organizes sensory information from the environment, which can lead to effective interaction with the surroundings. Efficient sensory processing is crucial for adequate perception, cognition, and behavior in everyday life [2]. Attention allocation refers to the process of selectively focusing on specific aspects of information while ignoring others, which can lead to improved task performance. A study conducted by Posner demonstrated that successful attention allocation helps individuals filter out irrelevant stimuli, enabling them to concentrate on pertinent information [3]. Working memory refers to the temporary storage and manipulation of information necessary for complex cognitive tasks, which can lead to enhanced problem-solving and decision-making abilities. For example, Baddeley established the significance of working memory in various cognitive functions, such as language comprehension, learning, and reasoning [4].

The absence or deprivation of certain cognitive factors in autistic children can hinder their ability to process intricate information and adapt to evolving circumstances. Thus, early diagnosis for autistic individuals is of utmost importance. By improving diagnostic capabilities for early autism detection, the affected individuals could receive proper interventions to address attentional and social challenges more effectively at an early state.

This article will focus on the development of machine learning models to accurately predict one's ASD outcome. That is, the current project will investigate whether machine learning models can accurately predict if an individual should be categorized as autistic by using patient demographic and questionnaire information. An empirical analysis will be performed on different popular classification supervised machine learning algorithms that were commonly used, such as Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine Classifier (SVC), Random Forest (RF), Decision Trees (DT). The hope is that thorough research will improve autistic prediction model, potentially offering more holistic assistance to children with autism, thus enhancing their daily experiences and educational endeavors.

2. Dataset

2.1. Description of dataset

The dataset used in the current paper is from an autistic prediction competition on Kaggle (https://www.kaggle.com/competitions/autismdiagnosis). The training set includes 800 individuals with the following features: ID (patient ID), A1_Score to A10_Score (score based on Autism Spectrum Quotient (AQ) 10 item screening tool), age (patient age in years), gender (patient gender), ethnicity (patient ethnicity), jaundice (whether the patient had jaundice at the time of birth), autism (whether an immediate family member has been diagnosed with autism), contry_of_res (patient country of residence), used_app_before (whether the patient has undergone a screening test before), result (score for AQ1-10 screening test), relation (relation of patient who completed the test), Class/ASD (classified result as 0 [non-autistic] or 1 [autistic]).

The AQ total score ranges from -6.14 to 15.85, with a mean of 8.54 and a standard deviation of 4.81. The individual AQ items score only contains 0 (disagree with the item) and 1 (agree with the item). Patient age ranges from 2.72 years to 89.46 years, with a mean of 28.45 years and a standard deviation of 16.31 years. There is a total of 530 (66.25%) male and 270 (33.75%) female. Out of the 800 patients, 615 (76.87%) did not have jaundice while 185 (23.13%) had jaundice at time of birth. The ethnicity of 203 people is unknown. For the rest of the 597 individuals, 257 (43.05%) are White-European, 97 (16.25%) are Middle Eastern, 101 (16.92%) are Asians, 47 (7.87%) are Black, 32 (5.36%) are Pasifika, and 63 (10.55%) are Other ethnicity. The sample are collected from individuals residing in 56 distinct countries, with the highest number of people residing in the United States (134), followed by India (108), New Zealand (78), United Kingdom (67), and Jordan (55) as the fifth.

The target variable is "Class/ASD". Of the 800 participants, 639 (79.88%) were not diagnosed with autism, while 161 (20.12%) were diagnosed with autism, which was unbalanced.

2.2. Statistical Analysis

To test whether the two continuous variables were significantly different for the individuals with autism or not, one-way ANOVA was applied. The results showed that age (F(1, 798) = 9.49, p=0.002) and AQ result score (F(1, 798) = 112.79, p<0.001) all have p-values smaller than 0.05, which means both of the continuous variables are different across autistic/non-autistic individuals. Thus, they were both included in the prediction model.

To test whether the five categorical features were significantly different across individuals with or without autism, chi-square test of independence was applied. The result of the four categorical features were all significantly different: ethnicity ($\chi2(10) = 112.49$, $p < 0.001$), jaundice at birth ($\chi2(1) = 14.59$, $p < 0.001$), immediate family member diagnosed with autism ($\chi2(1) = 100.82$, $p < 0.001$), and country of residence ($\chi2(55) = 205.75$, $p < 0.001$). The results show that gender ($\chi2(210) = 0$, p =0.98) was not significant, meaning whether an individual is male or female did not affect if they were autistic. Nevertheless, gender was retained as one of the predictor variables, as it is a frequently utilized factor in the models to address individual variations based on sex.

3. Prediction Task

3.1. Description of Task

As described above in the introduction, the current paper examined how accurately machine learning models could predict one's diagnosis of autism using demographic information and questionnaires. A categorical prediction task was conducted using machine learning. The five supervised machine learning models were: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine Classifier (SVC), Random Forest (RF), Decision Trees (DT). Details of the tuning hyperparameters are described below.

3.2. Supervised Machine Learning Algorithms

All five supervised machine learning models were implemented using python's sklearn packages. The specific parameter spaces are described as follows.

LR. penalty = 'l2'; regularization parameter C: 10^(-4) to 10^2; solvers: 'newton-cg', 'lbfgs', 'sag', and 'saga'.

KNN. p = 2; n_neighbors: 2,4,7,13,24,44,81,150,277,512; leaf_size: 2,4,6,8,10,12,14,16,18.

SVC. kernel: 'linear', 'rbf', and 'poly' degree 2 and 3; regularization parameter C: 10^(-3) to 10^3.

RF. n_estimators: 10,30,50; max_features: 10,20,50, None; max_depth: 3,5,7,15,20.

DT. max_depth: 2,3,5,7,10, None; min_samples_split: 2,3,5,7,10,20,35,50,70; min_samples_leaf: 2,3,5,7,10,15,20,35,50,70.

3.3. Performance Metrics

The only performance metric that was used in the current paper was the ROC AUC score. The reason for choosing the ROC AUC score was that the predicted variable "Class/ASD" was unbalanced, thus simply using accuracy score might bias the result towards the heavily weighted category (in this case, the non-autistic individuals). An ROC AUC score of 1 indicates that it can perfectly categorize the labels.

3.4. Feature Engineering & Cross Validation

The data will be split into 80% training set and 20% validation set to explore and fine tune the best model. All numeric variables (including AQ items scores) will be standardized and then applied with different polynomial degrees (degree=[1,2,3,4]). All categorical variables will be one-hot encoded.

A grid search for the hyperparameters of each machine learning models will be conducted to fine tune the model to have better performance. A 5-fold cross-validation will be used to assess the validity of the models. The cross-validation process will be performed 3 times to account for randomness in the splitting process. The average train and validation ROC AUC score will be calculated.

4. Result

The overall training and validation mean ROC AUC scores across 3 trials for each algorithm were listed below in Table 1. The boldfaced score is the algorithm with the best performance on train or validation set.

Table 1: Overall training & validation mean ROC AUC scores for each algorithm across averaged trials.

Table 1	Training Score	Validation Score
Logistic Regression (LR)	0.93	0.90
K-Nearest Neighbor (KNN)	0.91	0.90
Support Vector (SVC)	0.92	0.90
Random Forest (RF)	0.91	0.93
Decision Tree (DT)	0.90	0.85

From table 1, overall LR performed better than all the other algorithms during training (ROC AUC = 0.93), while RF performed better than all the other algorithms during validation (ROC AUC = 0.93). The fine-tuned hyperparameters of the best model (i.e., the trial that has the best validation ROC AUC score) of each algorithm was detailed below in Table 2.

Table 2: Fine-tuned hyperparameters of each model.

Model	Hyperparameters
LR	'degree':1; 'C': 0.001; 'penalty': 'l2'; 'solver': 'saga'
KNN	'degree':2; 'leaf_size': 2; 'n_neighbors': 24, 'p': 2
SVC	'degree':1; 'C': 0.001; 'gamma': 0.01; 'kernel': 'rbf'
RF	'degree':1; 'max_features': 10; 'n_estimators': 50; 'max_depth': 3
DT	'degree':4; 'max_depth': 5; 'min_samples_leaf': 10; 'min_samples_split': 35

Thus, the best model chosen to fit on the whole training set was a Random Forest (RF), with hyperparameters max_features = 10, n_estimators = 50, and max_depth = 3. A polynomial feature of degree = 1 was applied to the standardized numerical columns, and all categorical columns were one-hot encoded.

The final model rendered a training ROC AUC score of 0.93. Using the model to predict the provided test set, this renders a test ROC AUC score of 0.8623 (see https://www.kaggle.com/competitions/autismdiagnosis/leaderboard).

5. Discussion

The above statistical analysis, model selection, and model comparison indicate that machine learning algorithms can accurately predict an autism diagnosis using demographic information and questionnaires. If this model is useful in predicting whether an individual should be diagnosed with autism, then the answer shifted towards better treatment to tackle social and learning problems associated with autistic symptoms.

Sensory integration therapy is a type of therapy that aims to improve the ability of children with autism to process sensory information. This therapy is based on the idea that children with autism have difficulty processing sensory information, which can lead to attention problems. The goal of this therapy is to help the child integrate and organize sensory inputs more effectively, resulting in improved attention and overall functioning. For example, one study by Pfeiffer and colleagues found that sensory integration therapy led to significant improvements in behavior, sensory processing, and motor skills compared to the control group [5]. Another study by Fazlioglu & Baran found that sensory integration therapy helped reduce stereotypic behavior and increased purposeful play among children with autism [6].

Additionally, behavioral intervention is a widely used approach to improve attention and focus in children with autism. Research has shown that behavioral interventions can be effective for children with autism. For example, a study by Foxx & Azrin found that children with autism who received behavioral interventions demonstrated significant improvements in attention, communication, and social interaction skills [7]. Another study by Koegel and colleagues found that children receiving a comprehensive behavior intervention program made significant gains in cognitive, language, adaptive behaviors, and social skills [8]. Designing an effective reward system is an important part of behavioral intervention for children with autism. Rewards should be tailored to the specific needs of each child and should be designed to reinforce positive behaviors [9]. For example, a child who enjoys playing with toys could be rewarded with extra playtime or a new toy for successfully completing tasks requiring attention and focus.

Digital technology and media have also been widely used to improve the attention span of children with autism. According to a study by Ntalindwa [10], digital technology can provide engaging and interactive learning experiences that cater to the unique needs of children with autism. Using applications, games, and virtual reality tools can help children with autism practice attention in a controlled environment. Digital technology can also improve play skills, decrease challenging behaviors, provide video models, and help with speech in autistic children. For instance, a study by Ploog and colleagues found that the use of computer-assisted instruction resulted in better engagement and attention to tasks, as well as improvements in social interaction skills among children with autism [11].

Considering the wide range of therapies available for addressing social and attentional deficits in autistic patients, the classification model aims to provide valuable insights into the early diagnosis and intervention of autism.

6. Conclusion

The present study explored the application of machine learning algorithms for the accurate prediction of ASD outcomes using demographic information and questionnaires, emphasizing the importance of early diagnosis for autistic individuals to address attentional and social challenges. The simple-tuned best-performing model could accurately classify ASD outcome at 86%. This result signifies the potential of machine learning models in early ASD symptom prediction, offering significant support for individuals with autism to enhance their quality of life and educational experiences.

References

[1] American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.).

[2] Engel, A., & Keller, P. E. (2011). The perception of musical spontaneity in improvised and imitated jazz performances. Frontiers in psychology, 2, 83.

[3] Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. Journal of experimental psychology: General, 109(2), 160.

[4] Baddeley, A. (1992). Working memory. Science, 255(5044), 556-559.

[5] Pfeiffer, B. A., Koenig, K., Kinnealey, M., Sheppard, M., & Henderson, L. (2011). Effectiveness of sensory integration interventions in children with autism spectrum disorders: A pilot study. The American journal of occupational therapy, 65(1), 76-85.

[6] *Fazlioğlu, Y., & Baran, G. (2008). A sensory integration therapy program on sensory problems for children with autism. Perceptual and motor skills, 106(2), 415-422.*

[7] *Foxx, R. M., & Azrin, N. H. (1973). THE ELIMINATION OF AUTISTIC SELF-STIMULATORY BEHAVIOR BY OVERCORRECTION 1. Journal of Applied Behavior Analysis, 6(1), 1-14.*

[8] *Koegel, R. L., Koegel, L. K., & Carter, C. M. (1999). Pivotal teaching interactions for children with autism. School Psychology Review, 28(4), 576-594.*

[9] *Kasari, C., Freeman, S., & Paparella, T. (2006). Joint attention and symbolic play in young children with autism: A randomized controlled intervention study. Journal of child psychology and psychiatry, 47(6), 611-620.*

[10] *Ntalindwa, T., Nduwingoma, M., Uworwabayeho, A., Nyirahabimana, P., Karangwa, E., Rashid Soron, T., ... & Hansson, H. (2022). Adapting the use of digital content to improve the learning of numeracy among children with autism spectrum disorder in Rwanda: Thematic content analysis study. JMIR Serious Games, 10(2), e28276.*

[11] *Ploog, B. O., Scharf, A., Nelson, D., & Brooks, P. J. (2013). Use of computer-assisted technologies (CAT) to enhance social, communicative, and language development in children with autism spectrum disorders. Journal of autism and developmental disorders, 43, 301-322.*

Appendix

Table 3: Training & validation raw ROC AUC scores for each algorithm across trials.

Table 3	Trials #	ADULT	
		Train Score	Validation Score
LR	1	0.93	0.89
	2	0.92	0.91
	3	0.93	0.88
KNN	1	0.91	0.89
	2	0.91	0.91
	3	0.92	0.89
SVC	1	0.92	0.91
	2	0.92	0.90
	3	0.93	0.88
RF	1	0.92	0.90
	2	0.91	0.94
	3	0.91	0.94
DT	1	0.91	0.83
	2	0.90	0.85
	3	0.90	0.87