Autism Spectrum Disorder Prediction Using Machine Learning Algorithms

Chapter *in* Advances in Intelligent Systems and Computing · January 2020

DOI: 10.1007/978-3-030-37218-7_56

Poonkodi Palanisamy
SNS College of Technology
1 PUBLICATION **9** CITATIONS
SEE PROFILE

Summia Parveen
Sri Eshwar College of Engineering
1 PUBLICATION **9** CITATIONS
SEE PROFILE

Monisha Monisha
Vellore Institute of Technology University
1 PUBLICATION **9** CITATIONS
SEE PROFILE

Autism Spectrum Disorder Prediction Using Machine Learning Algorithms

Shanthi Selvaraj(✉), Poonkodi Palanisamy, Summia Parveen, and Monisha

Department of Computer Science and Engineering, SNS College of Technology,
Saravanampatti, Coimbatore 641035, Tamilnadu, India
psshanthiselvaraj@gmail.com, poonkodi.cse@gmail.com,
monish.ragu@gmail.com, summiaparveen@yahoo.in

Abstract. The objective of the research is to foresee Autism Spectrum Disorder (ASD) in toddlers with the help of machine learning algorithms. Of late, machine learning algorithms play vital role to improve diagnostic timing and accuracy. This research work precisely compares and highlights the effectiveness of the feature selection algorithms viz. Chi Square, Recursive Feature Elimination (RFE), Correlation Feature Selection (CFS) Subset Evaluation, Information Gain, Bagged Tree Feature Selector and k Nearest Neighbor (kNN), and to improve the efficiency of Random Tree classification algorithm while modelling ASD prediction in toddlers. Analysis results uncover that the Random Tree dependent on highlights chosen by Extra Tree calculation beat the individual methodologies. The outcomes have been assessed utilizing the execution estimates, for example, Accuracy, Recall and Precision. We present the results and identify the attributes that contributed most in differentiating ASD in toddlers as per machine learning model used in this study.

Keywords: Machine learning · Random Tree · Accuracy · Autism Spectrum Disorder · Feature selection · Classification

1 Introduction

Communication and behaviour is affected by a neuro developmental disorder termed as Autism Spectrum Disorder (ASD) [1]. Primary identification can significantly decrease ASD but, waiting time for an ASD diagnosis is long [2]. The rise in the amount of ASD affected patients across the globe reveals a serious requirement for implementation of easily executable and efficient ASD prediction models [3–6]. Currently, very inadequate autism datasets are accessible and several of them are genetic in nature. Hence, a novel dataset associated with autism screening of toddlers that contained behavioural features (Q-Chat-10) along with the characteristics of other individual's that is effective in detection of ASD cases from behaviour control is collected [3–6].

Clinical information mining [7] is the utilization of information mining methods to clinical information. The classification in healthcare is a dynamic area that provides a deeper insight and clearer understanding of causal factors of diseases.

S. Smys et al. (Eds.): ICCVBIC 2019, AISC 1108, pp. 496–503, 2020.
https://doi.org/10.1007/978-3-030-37218-7_56

In machine learning classification [8, 9] refers to a factual and numerical strategy for predicting a given information into one of a given number of class values. Feature Selection [8] is a mathematical function that picks a subset of significant features from the existing features set in order to improve the classifier performance and time and space complexity.

In this paper, we have used the ASD Toddlers Dataset to evaluate the performance of six feature relevance algorithms viz. Chi Square, Recursive Feature Elimination (RFE), Correlation Feature Selection (CFS) Subset Evaluation, Information Gain, Bagged Tree Feature Selector and k Nearest Neighbor (kNN), and Random Tree classification algorithm. Existing research in the field of machine learning and ASD research is succinctly presented in the following paragraphs.

Recently machine learning methods are widely used to improve the diagnosis process of ASD [10–12]. [13] used different machine learning algorithms like Tree models (random forest, decision tree) and Neural Network and Tree models in order to reduce the ASD diagnostic process time duration [14]. Piloted an experimental review for comparison of six major machine learning algorithms viz. Categorical Lasso, Support Vector Machine, Random Forest, Decision Tree, Logistic Regression, and Linear Discriminant Analysis, in order to choose the best fitting model for ASD and ADHD.

[15] studied the common issues associated to psychiatric disorders inclusive of external validity, small sample sizes, and machine learning algorithmic trials without a strong emphasis on autism. The authors suggested that one of these classifiers, used individually or in combination with the previously generated behavioural classifiers clearly focused on discriminating autism from non-autism category that is encompassing, could behave as a valuable triage tool and pre-clinical screening to evaluate the risk of autism and ADHD [14].

[16] used FURIA (Fuzzy Unordered Rule Induction Algorithm) on ASD dataset for forecasting autistic symptoms of children aged between 4-11 Years. The results exposed that FURIA fuzzy rules were able to identify ASD behaviours with 91.35% accuracy of classification and 91.40% sensitivity rate; these results were better than other Rule Induction and Greedy techniques.

[17] and [18] applied Naïve Bayes, SOM, K-Means, etc. on a set of 100 instances ASD dataset collected using CARS diagnostic tool [19]. Used Random Forest algorithm to predict ASD among 8-year old children in multiple US sites. The results revealed that the machine learning approach predicted ASD with 86.5% accuracy. The above appraisals have concentrated on the comprehensive domain of mining clinical data, the trials encountered, and the results inferred from analysis of clinical data.

2 Methodology

The proposed methodology for ASD prediction in toddlers is portrayed in Fig. 1. In this research, ASD patient data collected using Quantitative Checklist for Autism in Toddlers (Q-CHAT) data provided by ASD Tests app have been used for investigation [3–6]. The dataset used to train the classifier and evaluate its performance were downloaded from Kaggle [20]. The dataset contains 1054 samples with 17 features inclusive of class variables. According to WHO report, around 1% of the population has ASD, but this study samples get around 69% Toddlers of the data with positive ASD. It is because the test parameters have only qualitative properties of ASDs.

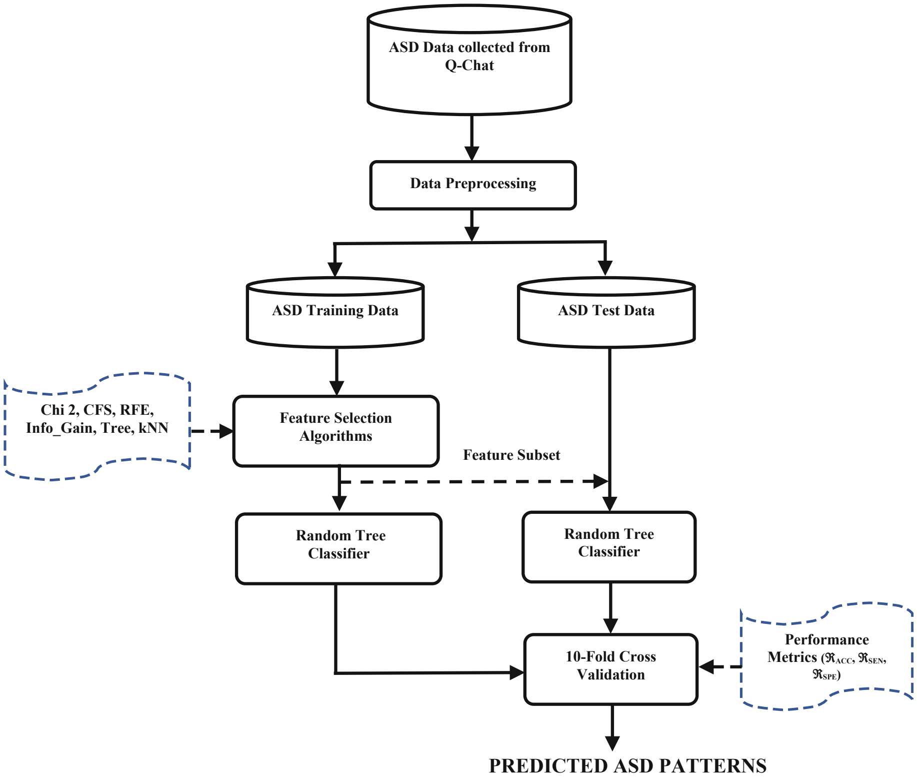
ASD Data collected from Q-Chat
Data Preprocessing
ASD Training Data
ASD Test Data
Chi 2, CFS, RFE, Info_Gain, Tree, kNN
Feature Selection Algorithms
Feature Subset
Random Tree Classifier
Random Tree Classifier
10-Fold Cross Validation
Performance Metrics ($\Re_{ACC}$, $\Re_{SEN}$, $\Re_{SPE}$)
PREDICTED ASD PATTERNS

Fig. 1. Proposed prediction model for ASD data

In order to identify the most significant features for classification, six feature selection algorithms viz, Chi Square Co-efficient [21, 22], Information Gain [23–26] based on ranked features, Recursive Feature Elimination (RFE) [27, 28], CFS [23–26, 29] using Best First search method, Tree Classifier [30] and kNN [31] were used. We utilized the feature subsets returned by the feature selection algorithms to estimate the performance of the Random Tree Classifier [23–26], in ASD prediction.

An in-depth evaluation revealed the relevance of the features and the predictive power of the classifier in ASD prediction. The confusion matrix [8] for ASD Prediction is given in Table 1.

For a binary classification problem (ASD_Positive, ASD_Negative), the following performance parameters [4, 8] were utilized for the classifier performance evaluation [8, 23–25]: Accuracy ($\Re_{ACC}$), Sensitivity ($\Re_{SEN}$) and Specificity ($\Re_{SPE}$). The measures are represented as follows:

Table 1. Confusion matrix of ASD class variables

Actual class	Predicted class	
	ASD_ Positive	ASD_ Negative
ASD_Positive	True Positive (TP)	False Negative (FN)
ASD_Negative	False Positive (FP)	True Negative (TN)

$$\Re_{\mathrm{ACC}} = Accuracy(\%) = \frac{|TP + TN|}{|TP + TN + FP + FN|} \quad (1)$$

$$\Re_{\mathrm{SEN}} = Sensitivity\,(\%) = \frac{|TP|}{|TP\,+\,FN|} \quad (2)$$

$$\Re_{\mathrm{SPE}} = Specificity\,(\%) = \frac{|TN|}{|TN\,+\,FP|} \quad (3)$$

The detailed description of obtained results is discussed in the subsequent sections.

3 Results and Discussion

The proposed experimentations were carried out on Python Scikit machine learning tool [32] and executed in a PC with Intel Core i5 processor with 1.8 GHz speed and 6 GB of RAM. The objective of this research work is to improve the performance of the classification algorithm using effective feature selection algorithms for ASD prediction in toddlers. The results are discussed in the subsequent sub sections.

3.1 ASD Toddlers Dataset

The analysis was carried out using ASD Toddlers dataset obtained from Kaggle [Kaggle]. The details of the datasets used for the study are described in Sect. 3.

3.2 Classification Accuracy Without Feature Selection

Random was used to predict ASD positive in toddler's dataset. It used the entire feature set of the dataset. The accuracy of the Random Tree algorithms without feature selection are listed in Table 2.

With all the predictive features, Random Tree classifier predicts ASD in Toddlers with 95.1% & 94.3% accuracies using training and testing datasets respectively.

3.3 Significant Features Selected by the Feature Selection Algorithms

The efficiency of six feature selection algorithms were investigated to identify the most relevant features for ASD prediction in toddlers. All the algorithms used default parameters. The Chi-Square algorithm selected the features using the Chi Square

Table 2. Performance of classification algorithm with all features

Classifier	Predictive performance measures in %					
	Training set (790 samples)			Test set (264 samples)		
Random tree	Accuracy ($\Re_{ACC}$)	Sensitivity ($\Re_{SEN}$)	Specificity ($\Re_{SPE}$)	Accuracy ($\Re_{ACC}$)	Sensitivity ($\Re_{SEN}$)	Specificity ($\Re_{SPE}$)
	95.1	89.3	97.6	94.3	93.9	94.5

Significance score and the bagged tree algorithm fits a number of randomized decision trees to rank the features based on their significance [30]. Table 3 depicts the features selected by the feature selection algorithms on the toddler ASD dataset. The symbol '*' in the table indicates that the feature is selected as a significant feature by the corresponding feature selection algorithm. The number specified in the () denotes the number of features selected by the respective feature selection algorithm.

Table 3. Significant features selected by feature selection algorithms

Feature #	Feature name	Chi Square (14)	RFE (15)	CFS (14)	Info_Gain (12)	Bagged Tree (10)	kNN (14)
f_1	A1	*	*	*	*	*	*
f_2	A2	*	*	*	*	*	*
f_3	A3	*	*	*	*		*
f_4	A4	*	*	*	*	*	*
f_5	A5	*	*	*	*	*	*
f_6	A6	*	*	*	*	*	*
f_7	A7	*	*	*	*	*	*
f_8	A8	*	*	*	*	*	*
f_9	A9	*	*	*	*	*	*
f_{10}	A10	*	*	*	*	*	*
f_{11}	Age_Mons			*	*		
f_{12}	Sex	*	*	*			*
f_{13}	Ethnicity		*	*	*		
f_{14}	Jaundice	*	*	*		*	*
f_{15}	Family_Member_With_ASD	*	*				*
f_{16}	By_Whom	*	*				*

From Table 3 it is observed that features f_1 to f_{10} and f_{14} are the most Significant features compared to all other features and plays vital role in predicting ASD in toddlers. RFE selected the maximum of 15 features and bagged tree selected the minimum of 10 features as significant features among 16 features. The optimal selection of these feature selection algorithms is substantiated by the Random Tree classification algorithm with 10 folds cross validation. The performance metrics of the

Random Tree classifier with the features selected by the feature selection algorithms are listed in Table 4.

Table 4. Performance of classification algorithm with significant features selected by feature selection algorithms

Feature selection algorithms	Predictive performance measures in %					
	Training set (790 samples)			Test set (264 samples)		
	Accuracy (ℜACC)	Sensitivity (ℜSEN)	Specificity (ℜSPE)	Accuracy (ℜACC)	Sensitivity (ℜSEN)	Specificity (ℜSPE)
Chi Square	94.9	90.2	97.1	94.3	93.9	94.5
RFE	95.2	90.2	97.4	94.3	93.9	94.5
CFS	93.5	88.1	96	94.3	95.1	94
Info_Gain	95.1	89.8	97.4	93.9	92.7	94.5
Bagged Tree	95.7	90.2	98.2	97	93.9	98.4
kNN	98	98	98	96.2	96.3	96.2

The results in Table 4 reveals that Random Tree classifier gave better accuracy of 98% with training set and 97% with the testing set using kNN and Bagged Tree feature selection algorithms respectively. kNN algorithm selected the k-value based on the error rates over the iterations. Based on the error rates 15 was selected as the k-value with low error rates.

The classification accuracy of the Random Tree classifier with the significant features selected by the feature selection algorithms improved (2.9% for training set and 2.7% for test set) compared to the model without feature selection. Careful evaluation of the questionnaire and Jaundice history are the significant features which decide ASD positive in Toddlers.

4 Conclusion

In this paper, we predicted ASD in toddlers using Random Tree classification algorithm and feature ranking algorithms (Chi Square, RFE, CFS, Information Gain, Bagged Tree and kNN). The results revealed that, Random Tree classifier using kNN feature selection algorithm gives high accuracy of 98% for training data set and Random Tree classifier through the features obtained by bagged tree algorithm gives high accuracy of 97% for test data set. The model improved the accuracy from 95.1% to 98% for the training set and from 94.3% to 97% for the test set. We conclude that Random Tree classifier using the relevant features selected by kNN and bagged tree algorithms outperforms other feature selection algorithms while modelling ASD in toddlers. From the rules generated by the Random Tree classifier, we observe that careful observation of kids and properly answering the ASD questionnaire will help to diagnose ASD and treat the toddlers in advance. Thus, these issues need to be concentrated to decrease the ASD positive rate in toddlers.

Compliance with Ethical Standards

✓ All authors declare that there is no conflict of interest.
✓ No humans/animals involved in this research work.
✓ We have used our own data.

References

1. National Institute of Mental Health. https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml
2. Penner, M., Anagnostou, E., Ungar, W.J.: Practice patterns and determinants of wait time for autism spectrum disorder diagnosis in Canada. Mol. Autism **9**, 16 (2018)
3. Thabtah, F., Kamalov, F., Rajab, K.: A new computational intelligence approach to detect autistic features for autism screening. Int. J. Med. Inform. **117**, 112–124 (2018)
4. Tabtah, F.: Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment. In: Proceedings of the 1st International Conference on Medical and Health Informatics, pp. 1–6, Taichung City. ACM (2017)
5. Thabtah, F.: ASDTests a mobile app for ASD, screening (2017). www.asdtests.com
6. Thabtah, F.: Machine learning in autistic spectrum disorder behavioural research: a review. Inform. Health Soc. Care J. **44**(3), 278–297 (2017)
7. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinf. **19**(6), 1236–1246 (2018)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Academic Press, Cambridge (2012)
9. Breiman, L.: Classification and Regression Trees. Routledge, New York (1984)
10. Wall, D.P., Dally, R., Luyster, R., Jung, J.Y., Deluca, T.F.: Use of artificial intelligence to shorten the behavioural diagnosis of autism. PLoS One **7**(8), e43855 (2012)
11. Wall, D.P., Kosmiscki, J., Deluca, T.F., Harstad, L., Fusaro, V.A.: Use of machine learning to shorten observation-based screening and diagnosis of Autism. Transl. Psychiatry. **2**(2), e100 (2012)
12. Lopez Marcano, J.L.: Classification of ADHD and non-ADHD Using AR Models and Machine Learning Algorithms. (Doctoral dissertation), Virginia Tech (2016)
13. Duda, M., Ma, R., Haber, N., Wall, D.P.: Use of machine learning for behavioral distinction of autism and ADHD. Transl. Psychiatry. **9**(6), 732 (2016)
14. Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F.: From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. **57**, 328–349 (2015)
15. Al-diabat, M.: Fuzzy data mining for autism classification of children. Int. J. Adv. Comput. Sci. Appl. **9**(7), 11–17 (2018)
16. Pratap, A., Kanimozhiselvi, C.: Soft computing models for the predictive grading of childhood Autism—a comparative study. Int. J. Soft Comput. Eng. (IJSCE) **4**(3), 64–67 (2014). ISSN 2231–2307
17. Pratap, A., Kanimozhiselvi, C.S., Vijayakumar, R., Pramod, K.V.: Predictive assessment of autism using unsupervised machine learning models. Int. J. Adv. Intell. Para. **6**(2), 113–121 (2014). https://doi.org/10.1504/IJAIP.2014.062174
18. Maenner, M.J., Yeargin-Allsopp, M., Van Naarden Braun, K., Christensen, D.L., Schieve, L. A.: Development of a machine learning algorithm for the surveillance of autism spectrum disorder. PLoS One **11**(12), e0168224 (2016)
19. Autism Screening. https://www.kaggle.com/fabdelja/autism-screening-for-toddlers

20. Ramani, G., Selvaraj, S.: A novel approach to analyze a combination of I x J categorical data for estimating road accident risk. Asian J. Inf. Technol. **15**(12), 2005–2015 (2016)
21. Liu, H., Setiono, R.: Chi2: feature selection and discretization of numeric attribute. In: Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, 5–8 November, p. 388 (1995)
22. Ramani, G., Selvaraj, S.: A pragmatic approach for refined feature selection for the prediction of road accident severity. Stud. Inf. Control **23**(1), 41–52 (2014)
23. Shanthi, S., Geetha Ramani, R.: Classification of vehicle collision patterns in road accidents using data mining algorithms. Int. J. Comput. Appl. **35**(12), 30–37 (2011)
24. Shanthi, S., Geetha Ramani, R.: Classification of seating position specific patterns in road traffic accident data through data mining techniques. In: Proceedings of Second International Conference on Computer Applications, vol. 5, pp. 98–104 (2012)
25. Latkowski, T., Stanislaw, O.: Data mining for feature selection in gene expression autism data. Expert Syst. Appl. **42**, 864–872 (2015)
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
27. Hoeft, F., Walter, E., Lightbody, A.A., Hazlett, H.C., Chang, C., Piven, J., Reiss, A.L.: Neuroanatomical differences in toddler boys with fragile X syndrome and idiopathic autism. Arch. Gen. Psychiatry **68**(3), 295–305 (2011)
28. Price, T., Wee, C.Y., Gao, W., Shen, D.: Multiple-network classification of childhood autism using functional connectivity dynamics. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. LNCS, vol 8675 (2014)
29. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
30. Altay, O. Ulas, M.: Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In: 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–4, Antalya (2018)
31. Machine Learning in Python. https://scikit-learn.org/stable/
32. Kotsiantis, S.B.: Supervised machine learning: a review of classification techniques. Informatica **31**, 249–268 (2007)