Predicting Autism Spectrum Disorder (ASD) for Toddlers and Children Using Data Mining Techniques

Rasool Azeem Musa, Mehdi Ebady Manaa and Ghassan Abdul-Majeed
University of Babylon, College of IT, Department of Information Networks

altaerasool@itnet.uobabylon.edu.iq
meh_mana@yahoo.com
it.mehdi.ebady@itnet.uobabylon.edu.iq
ghassan@uob.edu.iq

Abstract. Autism Spectrum Disorder (ASD) is a contemporary disease that has recently spread among toddlers and children. Many researchers have been interested to determine the features of the autism. However, this kind of studies is costly in term of the gathering information from several sources. In this paper, we introduced and applied a novel and early prediction techniques based on the using of data mining and machine learning tools. It is difficult to determine the features of any autism ages. In this paper, we used data mining prediction techniques which play an integral role to predict the symptoms of autism for any age group. The data of this study is AQ-10 dataset which are involved for toddlers and children. The results present a superior performance for ASD classification. Random forest, Decision Tree, Support Vector Machine, and Naive Bayes the accuracy of 1.0 with the features selected by correlation technique.

Keywords. ASD, data mining, random forest, J48, SVM, Naïve Bayes, AQ-10 dataset

1. Introduction

Autism Spectrum Disorder (ASD) is a condition of Growth trouble in the nervous system. The symptoms of this condition vary from mild to severe. Diagnosis is made at any stage of life, and symptoms appear fully in the first two years and the condition develops with age. Among the challenges facing ASD is difficulty concentrating, education, movement, and mental health problems such as anxiety, depression, and others. According to the World Health Organization, there is one child with ASD, compared to 160 children. Early diagnosis can greatly reduce ASD and help patients and healthcare sectors prescribe treatment and reduce costs. Data mining is the semi-automated method of evaluating broad collections of data sets, correlations, and dependencies [2].

Automated healthcare evaluation is the main target of the health services in recent days. Basically, health system is one of important and vital system in the people life. Automated analysis support medical staff and patients in standardization of data and provide easy and flexible model in collection data for researchers. Early diagnosis of autism disease is one of major challenge because consultants could require up to 6 months to recognize the main features of the autism disease. To avoid these long procedures and the traditional methods used to diagnose the features of people with autism, data mining techniques were used to accelerate this process, which in turn led to a major development in this field to improve the condition of a person with autism [1].
Supervised data mining techniques have been involved in the disorder detection. Primarily, feature selection techniques [3]. Healthcare professionals need easy and effective ASD screening methods in the terms of time and accuracy. The main objective of this paper is to propose and apply a model for the early predicting of the autism using data mining techniques that are a powerful and efficient tool to achieve this purpose.

2. Related Works
In this paper, related works of the ASD prediction are introduced in this section. In the study [4], multiple machine learning algorithms (Random Forest RF, Decision Tree J48, Naïve Bayes, Support Vector Machine SVM) was implemented to Identify ASD in children the researcher in this research [4] proposed using multiple machine learning algorithms The RF is obtained 0.861 accuracy and 0.854 precision and 0.851 recall. The naïve Bayes is obtained 0.865 accuracy and 0.866 precision and 0.865 recall. The SVM is obtained 0.833 accuracy and 0.835 precision and 0.833 recall. The J48 is obtained 0.871 accuracy and 0.871 precision and 0.871 recall. In our proposed work, these results have been improved using the same algorithms, but with an improved model builder that gives high prediction accuracy. The deep learning algorithms was applied [5] to identify autism spectrum disorder (ASD) patients from large brain imaging datasets (ABIDE) the accuracy of this system is 70% in the identification of ASD. We propose in our paper to use the Qchat dataset with machine learning algorithms which gave a better outcomes [6].

The researcher used machine learning techniques to predict autism and used Random Forest (RF - Iterative Dichotomiser3) and Random Forest (RF - Classification and Regression Trees) to obtain 96.52, 98.6, and 97.07 sensitivity and 88.52%, 84.6, and 97.11 specificity and accuracy in the child was 92.26%, in adolescent was 93.78%, and in adult reached 97.10%. The prediction results in this related work were improved in our work by using other machine learning algorithms that gave higher prediction accuracy. The Random Tree classifier has been utilized [7] to predict autism spectrum disorder (ASD) in Toddlers and the accuracy in this research was 95.1% to training datasets and 94.3% to testing datasets to obtain 89.3 and 93.9 sensitivity and obtain 97.6 and 94.5 specificity in training and testing datasets respectively. The Random Forest algorithm gave better prediction accuracy in our work by using the strong pre-processing step. Multiple algorithms was performed to prophesy ASD to toddlers, children, adolescents, and adults [8]. The researcher implemented classifiers (Adaboost, Flexible Discriminant Analysis (FDA, Decision Tree (C5.0), Boosted Generalized Linear Model (Glmboost), Penalized Discriminant Analysis, Mixture Discriminant Analysis (MDA), Linear Discriminant Analysis, Support Vector Machine (SVM) and Classification and Regression Trees). Support Vector Machine this algorithm provided better results for toddlers accurately reached 98.77%, while Glmboost showed better performance for the adolescent accurately reached 93.89%, Adaboost algorithm presented the best performance for the children accurately reached 97.20%, and Adaboost algorithm for the adult accurately reached 98.36%. In this paper, the researcher used more than one age group and these results were improved by building our improved model for two age groups for toddlers and children. The researcher in this proposed system [9], they carried out multiple classifiers to predicate ASD in children. The accuracy of the Naïve Byes – Support Vector Machine (SVM) reached 87.6% and then it was an algorithm Random Forest achieved 87.07%, while the neural network algorithm achieved 86.3%, and the summing neural network algorithm reached 85.08%. The proposed system for this research presented better results compared to previous research.

3. Problem statement
The major problem of this paper is related to problems of healthcare issues which are.
1. Huge, complex, and imbalanced data that are needed to a robust prediction model.
2. Low correlation between attributes and the target of prediction due to the reliability of the measures, the amount of variability in the data and outliers [10].
3. Optimizing the works that need to design a framework that solves the healthcare problems due to diagnostic errors and determination of autism characteristics and development in ASD case classification [11].

To address these problems, we have constructed a robust prediction model that produces a high accuracy in a reasonable time.

4. The Prediction Data Mining Algorithms

4.1. Decision tree (J48)

It is an algorithm which is one family members of the supervised learning algorithms. It is possible to involve to address problems of regression and classification. The aim of employing a Decision Tree is to build a training model that can be used to predict the class or value of the target variable by studying basic decision rules (training data) derived from prior data [10]. One of the advantages of this algorithm requires less effort for data preparation during pre-processing and one of the disadvantages of this algorithm is if a simple change occurs in any part of the data that leads to a major change in the decision-making tree that leads to instability of the algorithm [12].

4.2. Support Vector Machine (SVM)

It is a linear algorithm for dealing the problems of regression and classification. The idea of SVM is simple: The algorithm creates a line or a hyperplane that separates the data into categories [13]. One of the advantages of this algorithm is the efficient handling of memory, but it does not work well with big data [14]. SVM is regarded as the classifier that maximizes the difference between certain important vectors or support vectors. This method is most likely to result in a stronger data separation because the foundation lies in optimizing the distance between the support vectors that provide the ideal hyperplane. The basic Support Vector Machine (SVM) uses a linear decision threshold Data is linearly separated. However, it is attempting to reach a threshold that increases the margin (M) requiring the use of an optimizing procedure, with a restriction that all points be on the correct part of the hyperplane [15].

\[
y = x_iw + b = 0
\]  

(1)

Can extract the margin \( M \) as follows:

\[
M = \frac{(1-b)}{|w|} - \frac{(-1-b)}{|w|} = \frac{2}{|w|}
\]  

(2)

Here

- \( M \): the margin
- \( y \): predicted class label \( \in [-1, 1] \)
- \( x_i \): \( i^{th} \) attribute
- \( w_i \): weights
- \( b \): bias

4.3. Random Forest (RF)

Random forests are an outfit learning strategy by developing a large number of decision trees at preparing time and produce the class. It is ideal for decision trees to overfit their training set [15]. The random forest simulation algorithm applies the common bootstrap strategy of aggregating to tree learners. Instead of a training set \( X = x_1, \ldots, x_n \) with \( Y = y_1, \ldots, y_n \), bagging frequently \( (B \text{ times}) \) picks a random sample to substitute the training set and matches trees for \( b = 1, \ldots, B \) to such samples. Examples, with substitution, \( n \) illustration of \( X, Y \) training; name these \( X_b, Y_b \). Train a tree \( f_b \) on \( X_b, Y_b \) for classification or regression. Predictions for unknown samples \( x' \) can be produced after training.

\[
\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')
\]  

(3)

Where, \( \hat{f} \) is predictions from each tree, \( B \) is bagging repeatedly, \( b \) is sampling for \( (b = 1, \ldots, B) \), \( f_b \) is regression tree and \( x' \) is unseen samples [16].
4.4. Naive Bayes (NB)
This is a Bayes Theorem which based method in mathematical classification. This is one of the easiest algorithms for supervised learning. Classifier Naive Bayes is a quick, accurate, and reliable algorithm. Naive Bayes classifiers on broad datasets provide high accuracy and speed [17]. It is one of the powerful and scalable learning algorithms for machine learning and data mining, with strong predictive efficiency. It can be trained successfully in a supervised environment. This often includes minimal training data, and may also be utilized with various attributes for small and broad training data set. This classifier can accommodate incomplete values. Classifier claims that the attributes relate individually to the likelihood of the model that is described in the formula that follows:

$$P(c|F) = \frac{P(F|c) P(c)}{P(F)}$$ (4)

Where P is a probability, variable c is the dependent class label and F are the several feature variables [18].

5. The Proposed System
The proposed system for this paper to predict ASD in toddlers and children which is shown in Fig. 1. In this paper, the data of autism patients for toddlers and children was collected based on the Q-CHAT data [19] [20]. Dataset was relevant to autism evaluation of children with prominent characteristics to be utilized for further study, in particular in evaluating autistic characteristics and enhancing the diagnosis of ASD cases. Which proved useful in predicting ASD events from behavioural science monitors. The dataset is divided into training data 70% and testing data 30%. Training data is entered into classifiers (Random Forest, Naïve Bayes, Decision tree J48, and Support Vector Machine SVM) to build the model and testing data are recruited to test the model after building the model and testing it we use the measures to evaluate its performance. The datasets include 1565 instances with 17 main features (Case_No, A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, Qchat, Sex, Ethnicity, Jaundice, Family_mem_with_ASD, and Class). The features selection algorithm (correlation) is used to determine the best attributes that used for classification. The features that are selected using correlation algorithm are 12 which are (A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, Qchat, and Sex). These features are merged with other dataset to provide an autism prediction for more than one age group as shown in table (1).

| Name | Type   | Description                                                                 |
|------|--------|------------------------------------------------------------------------------|
| A1   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A2   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A3   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A4   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A5   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A6   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A7   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A8   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
| A9   | Binary (0,1) | The answer to the question, depending on the method used in the diagnosis       |
A10  Binary (0,1)  The answer to the question, depending on the method used in the diagnosis
Qchat  Integer  When patients scores higher than 3 (Q-chat-10-) then there are possible ASD traits otherwise no ASD traits are observed
Sex  String  Male or Female

In this paper we used attributes that are selected by features selection algorithm to estimate of the classifiers in ASD predication. The performance parameters are recruited for binary classification problem to evaluate accuracy and showed as shown in table (2):

| Actual class | Predicted class |
|--------------|-----------------|
| True         | True Positive (TP) True Negative (TN) |
| False        | False Positive (FP) False Negative (FN) |

\[
\text{\textit{RAcc} = Accuracy(\%) = \frac{|TP+TN|}{|TP+TN+FP+FN|} (5)}
\]

Figure (1) and Figure (2) shows the contribution of the preprocessing step for the ASD predation model.

![Figure 1. ASD Preprocessing Steps](image)

Figure 2 shows the splitting data percent into training and testing for the ASD predication model.

![Figure 2: Split data stage into training and testing](image)
The main proposed system and details for ASD prediction model is shown in the figure (3) included all the steps from the dataset to the final results.

![Diagram of the proposed prediction model for ASD data](image)

**Figure 3.** The Proposed prediction model for ASD data

### 6. Results and Discussion

The code used in this paper is on Spyder IDE and Lenovo computers with Core i7, 2.4 GHz and RAM 8 GB. The results of the algorithms (Decision Tree (J48), Naïve Bayes, Random Forest, and Support Vector Machine) showed in Table (3) for both children and toddlers used in this paper to predicate the autism spectrum disorder (ASD).

| Algorithms          | Precision | Recall | Accuracy |
|---------------------|-----------|--------|----------|
| Decision Tree J48   | 1.00      | 1.00   | 1.00     |
| Random forest       | 1.00      | 1.00   | 1.00     |
| Naïve Bayes         | 1.00      | 1.00   | 1.00     |
| Support Vector Machine | 1.00    | 1.00   | 1.00     |

The features selected in this paper by the feature selection correlation algorithm that showed advanced results by building an extremely predictive system for autism spectrum disorder where all the results of all the algorithms used in this paper are 100% and are considered the best results for related works that are searching in this field. It is clearly shown from table (3), the prediction algorithms perform high accuracy due to the robust pre-processing step in figure (1) and superior the related works.
It is clear that the results from Table (3), provided high results due to the classification accuracy classifiers with the important features selected by the feature selection algorithms (Correlation coefficient) improved result compared to the model with related works and converting categorical data to numerical data using label encoding technology.

7. Conclusion
ASD is a neurological disorder and long-term. ASD needs early, rapid, accurate and effective detection to provide appropriate health care. The traditional methods which have been proposed and performed previously in identifying the affected person spend a long time and costly, so data mining techniques were used as an alternative to the traditional methods which are working on a short time and lower cost patient diagnosis. Several studies in this field has been performed successfully to predict ASD, but these studies still did not reach a high performance in less costly time. In this paper, data mining techniques (Random Forest, Naïve Bayes, Decision tree J48, and Support Vector Machine SVM) were applied to perform early predicting ASD among toddlers and children using AQ-10 datasets for toddlers and children. The results show high performance in term of accuracy compared with other related works.

References
[1] H. S. Alarifi and G. S. Young, “Using Multiple Machine Learning Algorithms to Predict Autism in Children,” in Proceedings on the International Conference on Artificial Intelligence (ICAI), 2018, pp. 464–467.
[2] B. I. Seraphim, L. S. Rao, and S. Joshi, “Survey on Early Detection of Autism Using Data Mining Techniques,” Int. J. Eng. Technol., vol. 7, no. 2.24, pp. 79–80, 2018.
[3] K. S. R. Geetha Ramani, “AUTISM SPECTRUM DISORDER IDENTIFICATION USING DATA MINING TECHNIQUES TECHNIQUES.pdf,” Int. J. Pure Appl. Math., vol. 117, no. special, pp. 427–436, 2017.
[4] B. van den Bekerom, “Using machine learning for detection of autism spectrum disorder,” in Proc. 20th Student Conf. IT, 2017, pp. 1–7.
[5] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the ABIDE dataset,” NeuroImage Clin., vol. 17, pp. 16–23, 2018.
[6] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, “A machine learning approach to predict autism spectrum disorder,” in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–6.
[7] S. Selvaraj, P. Palanisamy, and S. Parveen, “Autism Spectrum Disorder Prediction Using Machine Learning Algorithms,” in International Conference On Computational Vision and Bio Inspired Computing, 2019, pp. 496–503.
[8] T. Akter et al., “Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders,” IEEE Access, vol. 7, pp. 166509–166527, 2019.
[9] S. H. Lee, M. J. Maenner, and C. M. Heilig, “A comparison of machine learning algorithms for the surveillance of autism spectrum disorder,” PLoS One, vol. 14, no. 9, 2019.
[10] L. D. Goodwin and N. L. Leech, “Understanding correlation: Factors that affect the size of r,” J. Exp. Educ., vol. 74, no. 3, pp. 249–266, 2006.
[11] A. Aziz and A. U. Rehman, “Detection of Cardiac Disease using Data Mining Classification Techniques,” IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 7, 2017.
[12] D. K, “Top 5 advantages and disadvantages of Decision Tree Algorithm,” medium, 2019. [Online]. Available: https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a. [Accessed: 27-Mar-2020].
[13] R. Pupale, “Support Vector Machines(SVM) — An Overview,” Towards Data Science data science, 2018. [Online]. Available: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989. [Accessed: 08-Mar-2020].
[14] D. K, “top 4 advantages and disadvantages of support vector machine,” medium, 2019. [Online].
Available: https://medium.com/@dhiraj8899/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107. [Accessed: 14-Mar-2020].

[15] W. Jamal, S. Das, I.-A. Oprescu, K. Maharatna, F. Apicella, and F. Sicca, “Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates,” J. Neural Eng., vol. 11, no. 4, p. 46019, 2014.

[16] W. contributors, “Random forest,” Wikipedia, The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=945233128.

[17] A. Navlani, “naive bayes scikit learn,” DataCamp, 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn. [Accessed: 15-Mar-2020].

[18] A. S. Halibas, L. B. Reazol, E. G. T. Delvo, and J. C. Tibudan, “Performance Analysis of Machine Learning Classifiers for ASD Screening,” in 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2018, pp. 1–6.

[19] Fadi, “autism screening for toddlers,” Kaggle, 2018. [Online]. Available: https://www.kaggle.com/fabdelja/autism-screening-for-toddlers. [Accessed: 15-Mar-2020].

[20] F. Thabtah, “autism datasets,” Fadi Thabtah, 2018. [Online]. Available: https://fadifayez.com/autism-datasets/. [Accessed: 15-Mar-2020].