



Towards Autism Subtype Detection Through Identification of Discriminatory Factors Using Machine Learning

Tania Akter^{1,2}, Mohammad Hanif Ali¹, Md. Shahriare Satu³,
Md. Imran Khan², and Mufti Mahmud⁴

¹ Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh

² Department of Computer Science and Engineering, Gono Bishwabidyalay, Savar, Dhaka 1344, Bangladesh

³ Department of Management Information Systems, Noakhali Science and Technology University, Sonapur, Noakhali 3814, Bangladesh

⁴ Department of Computing and Technology, Nottingham Trent University, Clifton Campus, Clifton, Nottingham NG11 8NS, UK

mufti.mahmud@ntu.ac.uk

Abstract. Autism spectrum disorder (ASD) is a neuro-developmental disease that has a lifetime impact on a person's ability to interact and communicate with others. Early discovery of autism can assist to prepare a plan for suitable therapy and reduce its impact on patients at an appropriate time. The aim of this work is to propose a machine learning model which generates autism subtypes and identifies discriminatory factors among them. In this work, we use Quantitative Checklist for Autism in Toddlers-10 (Q-CHAT-10) of toddler and Autism Spectrum Quotient-10 (AQ-10) datasets of child, adolescent, and adult screening datasets respectively. Then, only autism records are merged and implemented k-means algorithm to extract various autism subtypes. According to Silhouette score, we select the best autism dataset and balance its subtypes using random oversampling (ROS) and synthetic minority oversampling technique for numeric and categorical values (SMOTENC). Afterwards, various classifiers are employed into both primary dataset and its balanced subtypes. In this work, logistic regression shows the highest result for primary dataset. Also, it achieves the greatest results for ROS and SMOTENC datasets. Hence, shapely adaptive explanation (SHAP) technique is used to rank features and scrutinized discriminatory factors of these autism subtypes.

Keywords: Autism · K-means clustering · Machine learning · SHAP analysis · Discriminatory factors

1 Introduction

ASD is a neurological disorder that impacts on communication, interaction, and learning processes of affecting individuals [12]. It is a lifetime ailment that cannot

be totally healed. But, many significant traits of this disorder are often observed to the children. Thus, instant diagnosis and treatment (i.e. therapy and medicine) is useful to reduce multifarious complexity of this disease more precisely.

Non clinical screening tools such as AQ, Q-CHAT, Social Communication Questionnaire (SCQ), and Modified Checklist for Autism in Toddler (M-CHAT), M-CHAT/Revised with Follow-Up (RF) are more effective to detect autism at early stages [13]. However, these tools are more efficient because autistic patient or their guardians (i.e., parents, teachers, and others) can directly use them without any specific qualification. Therefore, physicians identify autism and provide proper treatment using different red flags. However, machine learning is an useful technique to investigate previous records and detect autism automatically. In many existing works, various machine learning methods were applied into Q-CHAT-10 and AQ-10 datasets to detect autism. Thabtah et al. [16] provided a computational intelligence (CI) technique named Variable Analysis (VA) that lessen several features of ASD screening datasets to predict autism more efficiently. Further, Akter et al. [4] generated several transformed datasets from Q-CHAT-10 and AQ-10 datasets where individual classifiers shows best performance. Then, these datasets were used to identify significant ASD risk factors for toddler, child, adolescent and adult subjects using various feature selection methods respectively. Thabtah and Peebles [17] represented rules based machine learning approach that detects autism traits and extracts rules to understand the reasons of this disorder. Akyol [5] identified many significant attributes for detecting autism using recursive feature elimination and stability selection methods. Wiratsin et al. [18] proposed a feature selection technique and extracted significant attributes of child, adolescent, and adult, respectively. Hossain et al. [10] evaluated these datasets with various state-of-the-art methods to find out the best classifier and feature set for Toddler, Child, Adolescent and Adult subjects respectively. Baadel et al. [7] proposed clustering based autistic trait classification (CATC) for AQ-10 datasets where they optimized input and identified relevant features based on similarity measures. Again, Akter et al. [3] used correlation based analysis to eliminate highly co-linear features of autism and investigated the predictivity of different classifiers using data transformation methods. In the application of Q-CHAT-10 and AQ-10 datasets, researchers were inspected their characteristics for both case and controls, respectively. However, the exploration of different ASD subtypes are required to realize major discriminatory characteristics of autism and ensure proper therapy of it.

In this study, we amalgamated only autism records from Q-CHAT-10 of toddler and AQ-10 datasets of child, adolescent, and adult, respectively. Then, several state-of-the-art methods were used to identify autism subtypes and investigate discriminative factors of them. This effort is beneficiary for numerous scientists and ASD welfare organizations. The most important contributions of this work are:

- Propose a machine learning model that extracts possible autism subtypes and characterizes this disorder more precisely.

- Explore these subtypes using various widely used machine learning methods (i.e., clustering, data transformation, classification, and feature selection) to evaluate this work.
- Use explainable machine learning methods to determine which features are important for getting the best result.
- Identify the key characteristics of autism subtypes by investigating discriminative factors.

2 Materials and Methods

2.1 Dataset Description

Baron-Cohen et al. [8] formulated Autistic-Spectrum (AQ) screening method to detect autism. Then, Allison et al. [6] proposed a shorten version of AQ (i.e. Q-CHAT 10, AQ-10 Child, AQ-10 Adolescent, and AQ-10 Adult) for identifying autism of different types of people respectively. Further, Thabtah et al. [14] developed a smart phone based app named ASDTests based on AQ-10 method for detecting ASD at early stage. Then, there were gathered individual's instances between 18–36 months, 4–11 years old, 12–16 years old and above 16 years respectively. However, two versions of AQ-10 dataset versions were gathered where version-1 contains 20 attributes and version-2 has 23 attributes (i.e., except Q-CHAT-10 Toddler dataset which hold 18 attributes). But, version-1 dataset is unavailable for Toddler. Besides, several questionnaires (A1 to A10) remain same in both versions where version-2 has more records than version-1 [10]. Moreover, child and adolescent datasets have similar queries while toddler and adult consume some particular questions. In this study, we work with version-2 based toddler (N = 1054), child (N = 509), adolescent (N = 248) and adults (N = 1118) screening datasets [7]. For every dataset, the class value “No” indicates this record have no ASD (i.e., control). Additionally, the instance having ASD denotes as class value “Yes”. These values are assigned based on the responses of AQ-10 questions. When the final score of them is less than or equal to 7, then the class value is declared as “No”, otherwise “Yes”.

2.2 Proposed Autism Subtype Detection Model

In this Section, proposed autism subtypes detection model is described step by step briefly as follows (see details in Fig. 1).

- **Data Preprocessing:** There are merged only autism instances of toddler, child, adolescent, and adult from Q-CHAT-10 (i.e. for toddler) and AQ-10 datasets respectively. This work is completely used the attribute notation of following works [7, 10, 17]. Then, several missing values of features are replaced with mean values. According to the Hossain et al. [10], we remove some irrelevant attributes [10] such as “Case”, “Used App Before”, “User (who completed the screening)”, “Language”, “Why taken the screening”, “Age Description”, “Screening Type”, and “Score”. In this work, some features

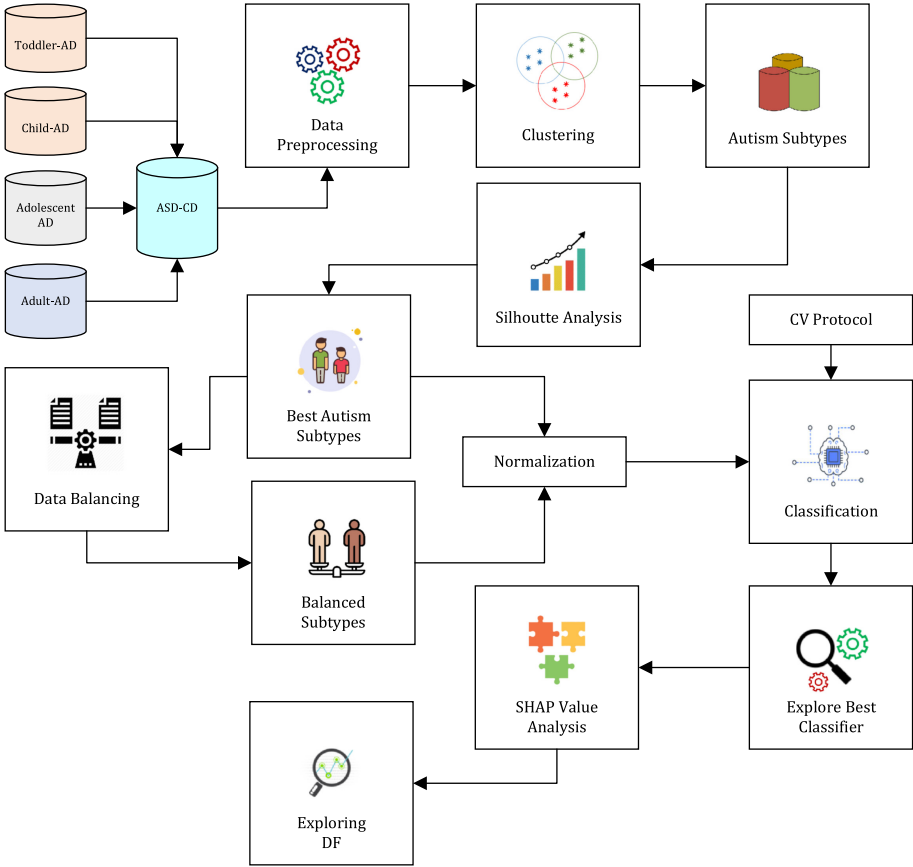


Fig. 1. Proposed autism subtype model

such as Sex, Jaundice, “Family_ASD” are encoded into numeric format where toddler age group has been converted from month to years.

- **Applying Clustering Technique:** Cluster analysis is used to generate similar groups from individual instances according to the data characteristics [1]. Consequently, k-means algorithm is created various clusters by calculating the distances among different records. However, we apply this method into combined autism dataset and generate various autism sub-types by changing the values of k from 2 to 10 in each iteration. Later, these subtypes are considered as individual “class labels” to reveal the predictability of proposed model.
- **Silhouette Analysis:** The estimation of the detachment gap and data consistency are measured in the individual clusters using silhouette analysis [1]. If this value is high, the object is well matched to its own cluster. In this work, this method is implemented on different autism subtypes in each iteration (i.e., depending on the value of k) to identify the best group.

- **Data Balancing:** In the best autism subtypes, the instances of majority and minority clusters are not equal and need to balance these subtypes for further investigation. In this work, we employ Random OverSampling (ROS) and Synthetic Minority Oversampling Technique for Nominal and Categorical Features (SMOTE-NC) and generate two balanced datasets. ROS randomly identifies various examples from minority class with replacement and appends them to the training dataset. On the other hand, SMOTENC is used to estimate class distribution by casually increasing minority instances for both continuous and categorical attributes.
- **Normalization:** Data transformation method is converted into another value to integrate and manage relevant features for machine learning analysis. In this work, we normalize (i.e. applying Z-Score method) baseline (i.e., autism subtypes) and its balanced datasets into suitable structures.
- **Classification Approaches:** Then, we apply several widely used classifiers [12, 15] such as Decision Tree (DT), Naïve Bayes (NB), K Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB) and Gradient Boost (GB) on normalized baseline, ROS and SMOTENC dataset. These classifiers which are implemented in this investigation have been widely used in earlier works relating to autism [11]. Then, the best classifier have been determined for each dataset where all classifier’s results are justified using some evaluation metrics like accuracy, f-measure and AUC (see details in Sect. 2.3).
- **Exploring Significant Features:** In a nutshell, SHapley Additive exPlanations (SHAP) is a game theoretic approach where shapley values evaluate the degree of contribution of each feature for the comprehensive machine learning model. In this work, we implement SHAP method to manipulate the priority of individual features of each autism dataset employing the best classifier. Afterwards, the discriminatory factors of autism subtypes are determined by counting frequency of individual items for both primary and balanced dataset, respectively.

2.3 Evaluation Metrics

In this work, several evaluation metrics such as accuracy, area under the curve (AUC), f-measure are implemented to evaluate the performance of various classifiers. These metrics are enumerated by true positive (TP), true negative (TN), false positive (FP) and false negative (FN), which are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$\text{F-Measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

$$\text{AUC} = \frac{\text{TP rate} + \text{TN rate}}{2} \quad (3)$$

3 Result and Discussion

In this study, we have implemented k-means and its silhouette analysis using Orange data mining toolkit version 3.29. Then, data balancing, and classification process (i.e., apply DT, NB, KNN, SVM, LR, RF, XGB and GB using 10 fold cross validation) have been employed using scikit-learn library [2] and feature interpretation has been done using SHAP library. Without clustering, all experiments are employed at Google Colaboratory in python [9].

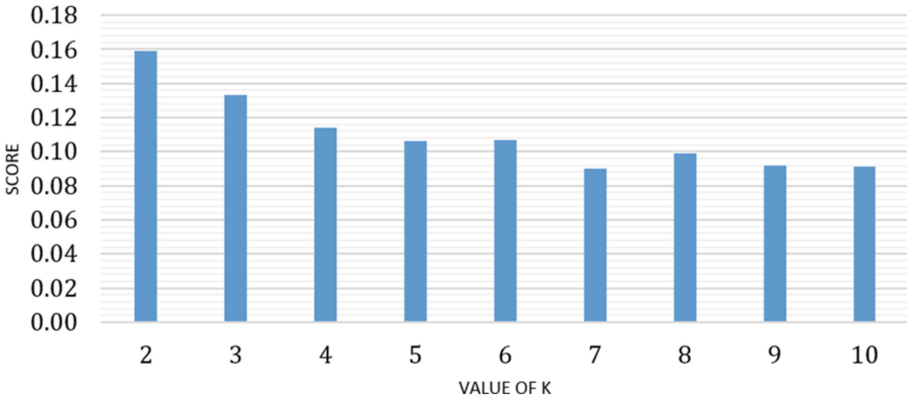


Fig. 2. Silhouette score for individual clusters depending on the values of k (for k = 2 to 10)

In this model, different autism subtypes are formulated depending on k values of k-means algorithm. In Fig. 2, we observe the silhouette values of generated clusters for k = 2 to 10, respectively. Then, the highest score 0.159 is obtained for k = 2 autism subtypes. Besides, other scores are gradually reduced for different k based autism subtypes. So, we take k = 2 subtypes as the best group for further machine learning analysis. These subtypes are denoted as subtype-1 and subtype-2, respectively.

3.1 Comparison of Performance of Individual Classifiers

Primary k = 2 autism subtypes (baseline) are balanced with ROS and SMOTENC methods and formulated two datasets. Then, various classifiers which is mentioned in Sect. 2.2 are used to investigate these datasets. Several evaluation metrics like accuracy, f-measure and AUC are used to justify the results (see Table 1). In this experiment, all classifiers produce high result (i.e. above 90%) where LR shows the highest 99.25% accuracy, 99.25% F-Measure, and 99.16% AUC, for baseline. On the other hand, these classifiers also provide good results (i.e. above 90%) for ROS and SMOTENC dataset respectively. Again, LR outperforms other classifiers where it shows 98.80% accuracy, 98.80% F-Measure

Table 1. Experimental results of individual datasets

| | Baseline | | | ROS Dataset | | | SMOTE Dataset | | |
|-----|--------------|--------------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|
| | Accuracy | F-Measure | AUC | Accuracy | F-Measure | AUC | Accuracy | F-Measure | AUC |
| DT | 97.35 | 97.37 | 97.37 | 98.61 | 98.61 | 98.61 | 98.06 | 98.06 | 98.06 |
| NB | 94.08 | 94.23 | 95.40 | 93.02 | 92.99 | 93.02 | 88.72 | 88.64 | 88.72 |
| KNN | 92.52 | 92.44 | 89.38 | 91.27 | 91.25 | 91.27 | 93.62 | 93.62 | 93.62 |
| SVM | 97.76 | 97.77 | 97.57 | 98.11 | 98.10 | 98.11 | 98.01 | 98.01 | 98.01 |
| LR | 99.25 | 99.25 | 99.16 | 98.80 | 98.80 | 98.80 | 99.03 | 99.03 | 99.03 |
| RF | 98.03 | 98.03 | 97.83 | 98.61 | 98.61 | 98.61 | 98.57 | 98.57 | 98.57 |
| XGB | 97.76 | 97.76 | 97.48 | 98.48 | 98.47 | 98.48 | 98.75 | 98.75 | 98.75 |
| GB | 97.62 | 97.62 | 96.89 | 98.11 | 98.10 | 98.11 | 98.29 | 98.29 | 98.29 |

and 98.80% AUC for ROS dataset. Also, this classifier presents 99.03% accuracy, 99.03% F-Measure and 99.03% AUC for SMOTENC dataset. Besides, the performance of almost all classifiers for balanced datasets are comparatively well than baseline.

3.2 Interpretation of Features for Discriminatory Factors

In primary autism subtypes, the ranks of shapely values of existing features (i.e. implying LR [15]) has been shown at Fig. 3(a). According to this illustration and statistical analysis, the identification of age group is the most significant discriminatory factor to distinguish its subtypes. Hence, the ages of subtype-1 are less than 12 years old and subtype-2 represent its age group greater than 11 years old. Then, Sex is considered as the second discriminatory factors where the number of female is greater than male in subtype-2. Therefore, shapely values of another features are prioritized from high to low values and explained how LR achieves the best performance in this work (see Fig. 3(a)). Besides, both subtypes contain more autistic traits than controls. However, subtype-1 shows more instances (i.e., yes/no responses) than subtype-2 (i.e., not found more DFs between two subtypes). In this experiment, “Pretending Capability (A8)”, “Sound Sense (A1)”, “Making Friends (A10)”, and “Interpreting Conversation (A5)” can be taken as next prior features for both subtypes (see Details in Fig. 3(a)). However, jaundice and “family_ASD” are not so such useful to distinguish autism subtypes.

Then, we determine shapely values to extract significant factors applying best performing LR for ROS and SMOTENC dataset which are shown at Fig. 3(b) and (c). Almost all autistic features (i.e., except A6 and A7) including no jaundice and “family_ASD” are oversampled to the minor subtype-2 for both data balancing techniques. In this case, subtype-1 contains more records than subtype-2. Like primary autism subtypes, age group is found as the most discriminatory factor for both balanced datasets. Therefore, the age group of subtype-1 is less than 22 years old and subtype-2 is more than 11 years old. Then, sex is the second best discriminatory factor where the female samples are also increased into subtype-2. Therefore, it preserve almost similar ranking of autistic features (i.e. “Pretending Capability (A8)”, “Character’s Intention (A7)”, “Making Friends (A10)”, “Sound Sense (A1)”, and “Interpreting Conversation (A5)”)

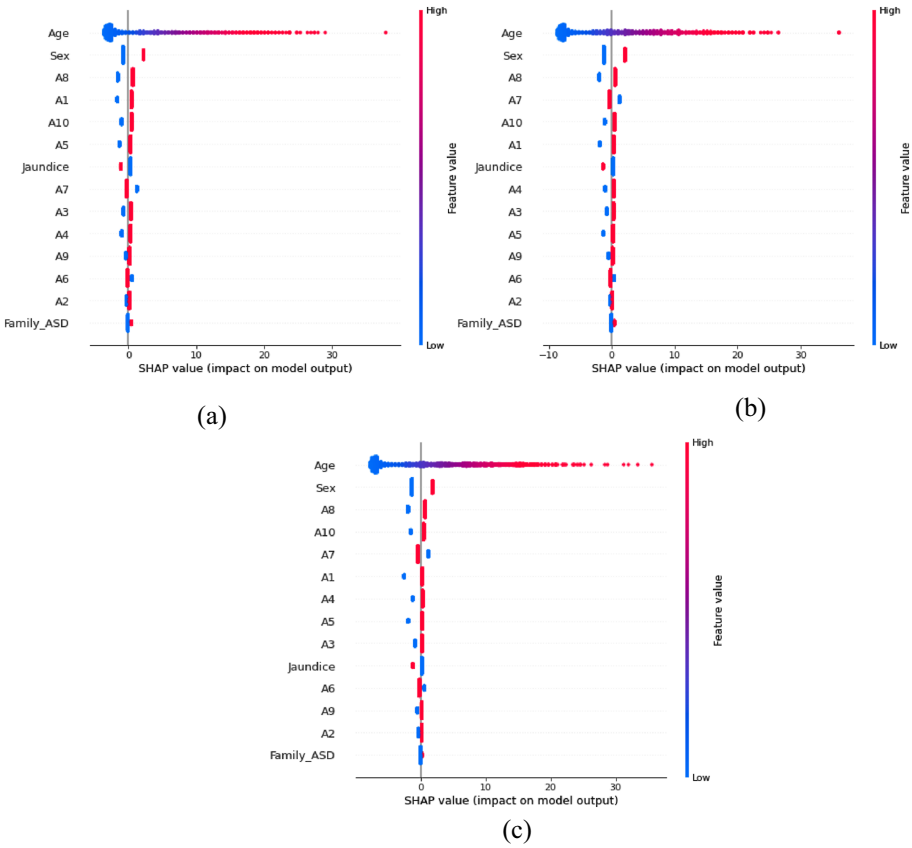


Fig. 3. SHAP values analysis using best performing LR for (a) primary, (b) ROS, and (c) SMOTENC generated autism subtypes

like primary subtypes (see details Fig. 3(b) and (c)). Then, “track conversation (A3)”, “back to the activities (A4)” are found as the medium level discriminative factors. Afterwards, “Social Chit-Chat (A6)”, “eye contact (A2)”, and “Family_ASD” do not more significant discriminatory factor in this analysis.

3.3 Comparative Studies and Implication

Many researchers investigated Q-CHAT-10 and AQ-10 datasets to explore significant outputs for autistic cases and some of them are briefly described at Sect. 1. In that works, they investigated common features between autistic and normal cases. However, the variation of autism can be determined by exploring autism subtypes. Nevertheless, previous models didn’t concern about these issue. In this work, we have separated only autism data from controls and extracted subtypes from Q-CHAT-10 and AQ-10 based datasets. Then, the best autism subtypes

are formulated and different significant features are prioritized to explore discriminative factors between two subtypes. This key characteristics are useful to understand the level of autism among various types of patients that ensure proper treatments of them. It enhances technical skills and managerial strategy of clinicians to detect real features. In addition, the economical cost of patients are reduced because of detecting discriminatory factors more quickly.

4 Conclusion and Future Work

We integrate and preprocess only autistic cases to categorize them into different groups and explore the best autism subtypes. Then, different classifiers have been applied into the best autism subtypes where LR demonstrates the best results (99.25% accuracy, 99.25% f-measure and 99.16% AUC) for baseline, (98.80% accuracy, 98.80% f-measure, and 98.80% AUC) for ROS, and (99.03% accuracy, 99.03%f-measure and 99.03% AUC) for SMOTENC datasets. The rapid identification of autism subtypes is useful for clinicians and psychiatrists to realize distinguished factors and specify proper therapy for them. Further, it helps to produce relevant drugs and cure various abnormal behaviors of ASD patients. Besides, Q-CHAT-10 and AQ-10 datasets which are used in this study are not huge. In future, we will collect more records about autism and use more advanced machine learning techniques (i.e., deep/reinforcement learning approaches) to increase performance of detecting autism subtypes.

References

1. Akter, T., et al.: Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. *Brain Sci.* **11**(6), 734 (2021)
2. Akter, T., Ali, M.H., Khan, M.I., Satu, M.S., Moni, M.A.: Machine learning model to predict autism investigating eye-tracking dataset. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 383–387. IEEE (2021)
3. Akter, T., Khan, M.I., Ali, M.H., Satu, M.S., Uddin, M.J., Moni, M.A.: Improved machine learning based classification model for early autism detection. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 742–747. IEEE (2021)
4. Akter, T., et al.: Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access* **7**, 166509–166527 (2019)
5. Akyol, K.: Assessing the importance of autistic attributes for autism screening. *Exp. Syst.* **37**(5), e12562 (2020)
6. Allison, C., Auyeung, B., Baron-Cohen, S.: Toward brief “red flags” for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *J. Am. Acad. Child Adolesc. Psychiatr.* **51**(2), 202–212 (2012)
7. Baadel, S., Thabtah, F., Lu, J.: A clustering approach for autistic trait classification. *Inf. Health Soc. Care* **45**(3), 309–326 (2020)

8. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E.: The autism-spectrum quotient (AQ): evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J. Autism Dev. Disord.* **31**(1), 5–17 (2001)
9. Bisong, E.: Building machine learning and deep learning models on Google cloud platform. Springer, Berkeley (2019). <https://doi.org/10.1007/978-1-4842-4470-8>
10. Hossain, M.D., Kabir, M.A., Anwar, A., Islam, M.Z.: Detecting autism spectrum disorder using machine learning techniques. *Health Inf. Sci. Syst.* **9**(1), 1–13 (2021). <https://doi.org/10.1007/s13755-021-00145-9>
11. Hyde, K.K., et al.: Applications of supervised machine learning in autism spectrum disorder research: a review. *Rev. J. Autism Dev. Disord.* **6**(2), 128–146 (2019)
12. Omar, K.S., Mondal, P., Khan, N.S., Rizvi, M.R.K., Islam, M.N.: A machine learning approach to predict autism spectrum disorder. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–6. IEEE (2019)
13. Tartarisco, G., et al.: Use of machine learning to investigate the quantitative checklist for autism in toddlers (q-chat) towards early autism screening. *Diagnostics* **11**(3), 574 (2021)
14. Thabtah, F.: An accessible and efficient autism screening method for behavioural data and predictive analyses. *Health Inform. J.* **25**(4), 1739–1755 (2019)
15. Thabtah, F., Abdelhamid, N., Peebles, D.: A machine learning autism classification based on logistic regression analysis. *Health Inf. Sci. Syst.* **7**(1), 1–11 (2019). <https://doi.org/10.1007/s13755-019-0073-5>
16. 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)
17. Thabtah, F., Peebles, D.: A new machine learning model based on induction of rules for autism detection. *Health Inf. J.* **26**(1), 264–286 (2020)
18. Wiratsin, I.O., Narupiyakul, L.: Feature selection technique for autism spectrum disorder. In: Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence, pp. 53–56 (2021)