Use of Machine Learning to Investigate The Quantitative Checklist For Autism in Toddlers (Q-CHAT) Towards Early Autism Screening

Gennaro Tartarisco
National Research Council of Italy

Giovanni Cicceri
Universita degli Studi di Messina

Davide Di Pietro
Universita degli Studi di Messina

Stefania Aiello
National Research Council of Italy

Elisa Leonardi
National Research Council of Italy

Flavia Marino
National Research Council of Italy

Flavia Chiarotti
National Institute of Health

Antonella Gagliano
University of Cagliari and "G. Brotzu" Hospital Trust

Giuseppe Maurizio Arduino
ASLCN1, Mondovì, Cuneo

Fabio Apicella
IRCCS Fondazione Stella Maris

Filippo Muratori
IRCCS Fondazione Stella Maris

Toddlers Team
National Research Council of Italy

Dario Bruneo
Universita degli Studi di Messina

Carrie Allison
Autism Research Centre

Simon Baron Cohen
Autism Research Centre

David Vagni

National Research Council of Italy

Giovanni Pioggia

National Research Council of Italy

Liliana Ruta (liliana.ruta@cnr.it)

National Research Council of Italy (CNR) https://orcid.org/0000-0003-4615-7495

Research

Keywords: Q-CHAT, early screening, Machine learning, autism

Posted Date: July 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-40334/v1

Version of Record: A version of this preprint was published at Diagnostics on March 22nd, 2021. See the published version at https://doi.org/10.3390/diagnostics11030574.

Abstract

Background: In the past two decades, several screening instruments have been developed to detect toddlers who may be autistic, both in clinical and unselected samples. Among others, the Quantitative CHecklist for Autism in Toddlers (Q−CHAT) is a quantitative and normally distributed measure of autistic traits which demonstrated good psychometric properties in different settings and cultures. Recently machine learning (ML) has been applied to behavioural science to improve classification performance of autism screening and diagnostic tools, but mainly in children, adolescents and adults.

Methods: In this study, we used machine learning (ML) to investigate the accuracy and reliability of the Q−CHAT in discriminating young autistic children from those without. Three different ML algorithms (Random Forest, Naive Bayes and Support Vector Machine) were applied to investigate the complete set of Q-CHAT items and the best predictive items.

Results: Our results showed that the three selected models outperformed the classical statistical methods of predictive validity and among the three ML classifiers, the Support Vector Machine was the most effective, being able to classify autism with 95% accuracy. Furthermore, using the Support Vector Machine-Recursive Feature Elimination approach we were able to select a subset of 14 items ensuring an accuracy of 93%, while an accuracy of 83% was obtained from the best 3 discriminating items in common to our and the previous reported Q-CHAT-10.

Limitations: Further data collection is needed.

Conclusions: This evidence confirms the high performance and cross-cultural validity of the Q-CHAT and supports the application of ML to create shorter and faster versions of the instrument maintaining high classification accuracy, to be used as a quick, easy and high-performance tool in primary care settings.

1. Background

Autism is a set of neurodevelopmental conditions characterized by impairments in social communication alongside repetitive, restricted interests and behaviours as well as atypical reactivity to sensory stimuli (APA, 2013). Autism is a lifelong condition whose severity and intensity of symptoms are heterogeneous and first signs occur in early childhood with different developmental trajectories [1]. Early screening and developmental surveillance have represented a primary goal in the past two decades and many screening tools have been developed and tested. However, performance, classification accuracy and reliability of those screening instruments vary depending on different settings, samples and screening designs, thus posing critical issues for clinical application [2–6]. Among the most popular and replicated screening tools, the M-CHAT and the subsequent revised M-CHAT/RF [7] have been applied in large mixed samples including both high and low likelihood groups, demonstrating low to moderate accuracy in detecting autism [8]. Other screening tools, such as the Social Communication Questionnaire (SCQ), reported poor balance between sensitivity and specificity in high-likelihood toddlers [9], while measures like the Screening Tool for Autism in Two-years-old (STAT) [10] and the Baby and Infant Screen for Children with aUtIsm Traits (BISCUIT) [11] have been tested only in case-control studies, requiring further prospective population studies. With the shift from a categorical to a dimensional approach to autism diagnosis, a quantitative measure of autistic traits, the Quantitative CHecklist for Autism in Toddlers (Q-CHAT) has been tested in different sample populations and cultures, in both case-control studies and primary care settings, displaying fair to good psychometric properties and predictive validity as well as good cultural stability [12–18]. A short Q-CHAT version, the Q-CHAT-10, including the best 10 predictive items was also developed [19], aiming to create a quick "red flags" tool suitable for the time constraints of pediatric check-ups and to help further reduce the delay for potential referrals.

Most recently, computational intelligence and machine learning (ML) have been applied to behavioural science and provided novel opportunities to improve predictive accuracy and classification reliability in relation to early screening, detection and autism diagnosis. ML algorithms are able to support autism screening and diagnosis by improving sensitivity and specificity of the screening and diagnostic tools and by helping to identify the least number of items maintaining satisfying classification accuracy. Classification accuracy is the number of correct predictions from all predictions made, multiplied by 100 to turn it into a percentage. A classification accuracy of 0.50 indicates random prediction of the independent variable while accuracy > 0.90 indicates excellent predictive validity. One of the first studies related to the use of ML to diagnostic tools was conducted by Wall et al. (2012)[20] and applied ML algorithms to the Autism Diagnostic Observation Schedule (ADOS, Module 1). Eight items were able to classify autism with nearly 100% sensitivity and 94% specificity. However, this study had some conceptual and methodological pitfalls and results were not replicated in another study, underlining the importance of the intersection between computational and behavioural science to ensure the correct application and interpretation of ML approaches [21]. Subsequently, sparsifying ML models were applied to ADOS modules 2 and 3 (for autistic children with verbal communication) finding a classification accuracy of 95% for module 3 and 93% for module 2, selecting the best 10 items [22]. In another study, ML was used to classify autism versus ADHD from the Social Responsiveness Scale items (SRS),[23]. An accuracy of 96.5% from only five items was achieved [24]. Bone and colleagues (2016) [25] applied ML strategies combining codes from both the SRS [23] and the Autism Diagnostic Interview-Revised (ADI-R) [26]. Processing items from multiple instruments, the ML algorithm was able, using only five behavioral codes, to detect autism with a sensitivity of 89.2% and a specificity of 59.0%. Very recently, ML was applied to datasets collected using a mobile application called ASDTests [27]. ASDTests app was developed to screen for autism in children, adolescents and adults using the short forms of the Autism Spectrum Quotient (AQ-10) and the Q-CHAT-10 respectively. In the first study, Thabath and colleagues (2019) [28], analyzed the adult version of the AQ-10 using a new rules-machine learning (RML) and were able to achieve an accuracy of about 90%, sensitivity of 87% and specificity of about 90%, while in the second study, the same authors applied the Naïve Bayes algorithm to the AQ-10 and found a similar accuracy of 92.8%, 91.3% and 95.7% for the child, adolescent and adult versions respectively. To the best of our knowledge, the only study applying ML to toddlers was conducted by Akter and colleagues (2019) [29], who analyzed the Q-CHAT-10 collected using the dataset from the ASDTests app and found that using a range of different classifiers, when optimized, they were able to effectively classify autism with an accuracy of 98%.

Hence, we aim to apply different ML approaches to investigate the accuracy and reliability of the Q−CHAT to classify young children as being autistic or neurotypical. We used three different machine learning (ML) algorithms (Random Forest, Naive Bayes and Support Vector Machine) to analyze the complete set of Q-CHAT items and the best subset of discriminating items in a sample of clinically referred young autistic children, compared to typically developing children. Furthermore, we explored the cross-cultural validity of the results obtained with ML in our Italian sample.

2. Materials And Methods

2.1 Participants

In this study we used a machine learning approach to analyze a previously collected dataset of young autistic and neurotypical children who were administered the Q-CHAT to explore the psychometric characteristics of the instrument in a multicentre study including different Italian regions (Sicily, Tuscany and Piedmont). The study received Ethics approval by the local Committees and all the participants signed a written informed consent form to be enrolled in the study. For all the detailed socio-demographic and clinical characteristics of the sample, refer to Ruta et al. (2019) [13]. A group of n=126 typically developing children (TD) [mean age (SD) = 33.2 (9.3) months] and n=139 autistic children [mean age (SD) = 31.6 (8.0) months] were included in the analysis.

2.2 Machine learning classifier

The classifier was constructed using the italian Q-CHAT data repository of n=265 children with an age range between 22 and 43 months. Any individual with more than 25% of missing answers was excluded from the analysis. N=6 children were excluded for this reason and the final dataset included n=137 subjects with autism and n=122 TD children. The information processed was based on the Q-CHAT questionnaire, consisting of 25 items related to the child's development reflecting autistic traits. Each item (representing a feature for our dataset) is rated on a five-point Likert scale (0–4), with higher ratings indicating more autistic traits and a Q-CHAT total score ranging from 0 to 100. We applied a supervised approach of binary classification, dividing the dataset into two classes according to the diagnostic category (autism vs TD). In this study we tested three different classifiers: Random Forest (RF)[30], Naive Bayes (NB) [31], and Support Vector Machine (SVM) [32] to understand the intrinsic relationship between Q-CHAT items and the diagnostic label. The configuration of each classifier is reported in Table 1. The classification respectively of autism vs TD was carried out performing a 5-fold cross-validation, utilizing 80% (n=207 subjects) for training and the remaining 20% (n=52 subjects) for testing the accuracy of each classifier [33]. We trained the ML algorithms on a laptop equipped with i7-8550U, 8Gb Ram, 256Gb SSD processor, using an Ubuntu 18.04.4 LTS operating system. We used *Pandas* and *Numpy* libraries for data manipulation, and scikit-learn package v.0.22.1 [34] for machine learning in python.

2.3 Feature selection

Once the ML classification with all 25 Q-CHAT items was completed, we tried to select a subset of features in order to identify a faster screening tool, without compromising the screening accuracy and reliability of the Q-CHAT. The Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithm was applied. The RFE is a recursive process that ranks and select features according to some score function with highest score. The SVM with non-linear kernel is run after each RFE iteration to assess all possible subsets of attributes. This process is repeated until the highest classification accuracy is obtained [35].

2.4 Metric for ML performances

The most common metrics for binary classification models are based on standard definitions such as TP and TN which represent respectively the number of tests respectively true positives and true negatives. FP and FN stand for the number of misclassified positive (false positive) and negative (false negative) instances. From these parameters, a number of model performance metrics can be derived. The most common metric is the accuracy, which represent the overall success rate of classifier and is computed as: Accuracy = (TP + TN) / (TP + FP + FN + TN). Other performance metrics include Sensitivity/Recall which is defined as percentage of correctly classified instances and is computed as: Sensitivity/Recall = TP / (TP + FN), and Specificity/PPV defined as percentage of incorrectly classified instances and computed as: Specificity/PPV = TP / (TP + FP). F1- score computed as F1 = (2TP)/(2TP + FP + FN) is the measure of test's accuracy and it's based on the harmonic mean of specificity and sensitivity. It reaches its best value at 1 and worst at 0. For reference and evaluation of classifiers also the mean square error (MSE) is computed.

3. Results

All the three ML algorithms showed satisfactory accuracy, sensitivity and specificity, with the SVM displaying the best discriminant validity. The area under the curve in the ROC analysis, reported in Figure 1, confirmed better performance of the SVM (95%) with respect to the RF (90%) and the NB (89%). When we applied the SVM-RFE algorithm to select the best pool of discriminant items, we found an improvement in accuracy as Q-CHAT items were progressively added, reaching up to 95% of accuracy when all the Q-CHAT items were included (See Figure 2). The SVM-RFE algorithm allowed also to identify the best number of discriminating items, with 14 Q-CHAT items reaching an accuracy of about 93%. The selected 14 items, ordered by accuracy using an integrated rank scoring were: q01, q02, q19, q04, q05, q06, q07, q09, q16, q17, q03, q25, q18, q22. As shown in Table 2, eight items (q01, q02, q19, q05, q06, q09, q17, q25) were in common with those reported by Allison and colleagues in a previous study where the Q-CHAT-10 was composed of the best 10 discriminating items [19]. To explore the replicability of the Q-CHAT-10 results in our sample, we ran the three ML algorithms also on the 10 items selected by Allison and colleagues as well as on the most discriminating 3 items, which were in common between our and Allison's study. Table 3 reports the performance of the three classifiers in relation to the original 25-items

Q-CHAT, the 14-items selected by the SVM-RFE algorithm, the 10 items selected by Allison and colleagues and the 3 most discriminating items in common to the two studies. The SVM algorithm confirmed overall the best accuracy for each Q-CHAT version (25, 14, 10 and 3 items). Finally, in Table 3, we extracted the Positive Predictive Value (PPV), Sensitivity and F1-score for each class (autism vs TD) from the 52 participants of the test-set, to cross check the validity of the trained ML models.

4. Discussion

In this study we applied machine learning and computational intelligence to improve the classification accuracy of the Q-CHAT and to investigate the best sub-set of items able to efficiently discriminate between young autistic and neurotypical children. We tested three different machine learning classifiers (SVM, RF, NB) and we found satisfactory accuracy, sensitivity and specificity for all three algorithms applied. The top-performing SVM model reached an overall accuracy of 95% with a sensitivity and specificity of 90% and 100% respectively, compared to RF (sensitivity=85% and specificity=95%) and NB (sensitivity=82% and specificity=100%). If we compare these results with those obtained applying the standard ROC analysis to the same participant sample [13], we found that ML algorithms were able to improve the classification accuracy of the Q-CHAT. In the previous study [13], the ROC curve showed an accuracy of 89.5% (vs. 95%), sensitivity = 83% (vs. 90%) and specificity = 78% (vs. 100%). Furthermore, by running the SVM-RF algorithm, we were able to select a sub-group of 14 items which maintained a very high accuracy, sensitivity and specificity (93%, 87% and 96% respectively for SVM). In our sample, 8 out of the 14 items (q01, q02, q019, q5, q06, q9, q17, q05) were in common with the Q-CHAT-10 by Allison and colleagues (2012) [19]. To further explore the cross-cultural validity of the instrument we applied the three ML classifiers to the 10 items selected by Allison and colleagues on the Q-CHAT-10 (2012) [19] and we found that, in our sample, SVM algorithm was able to classify autism with 87% accuracy, 65% sensitivity and 86% specificity. These results are in line with those recently reported by Akter et al. (2019) [29], where a dataset of Q-CHAT-10 administered using a mobile application [27] was analyzed using ML and the SVM algorithm was able to classify autism with 98% of accuracy. Together, these findings confirm a satisfactory cross-cultural validity of the Q-CHAT in different samples, countries and languages. Furthermore, when we looked at the specific items in common between the Q-CHAT-10 and our subset of items, interestingly, the three items with the highest ranking in our analysis (q01, q02, q019) were the same as those with the highest PPVs in Allison's study [19]. Taking into account just these 3 items, our SVM algorithm was able to classify autism with an accuracy of 83%, a sensitivity of 78% and a specificity of 93%. These items refer to reduced response to name, eye contact and use of gestures which strongly tap into the core autism symptoms related to social orienting and communication and have been consistently picked up as reliable early "red flags" for autism (see the NICE [36] and the CDC [37] guidelines). Furthermore, "unusual eye contact" was one of the 8 selected ADOS items able to classify autism with nearly 100% sensitivity and 94% specificity [20] and "direct gaze" on the ADI-R was one of the 3 most discriminant items using a novel ML fusion approach in another study by Bone and colleagues (2016) [25].

4.1 Limitations

The study has some limitations. The sample size is relatively small for this kind of computational approach. To face this issue, we used the 5-fold cross validation to increase the estimation of the performance of ML models. Furthermore, we provided a well-balanced dataset between autism and TD children (139 autism vs 126 TD) to train the ML models which helped us to work with the real dataset, avoiding strategies such as repeated random under sampling, that may miss out important samples by chance and create a bias in terms of assessment of metric for ML performance. Nevertheless, a further validation study with a larger cohort is needed and efforts are currently underway, to improve the performance of our models and to pave the way to test other sophisticated ML algorithms which require more data.

5. Conclusions

In this study we investigated the performance of three different classification algorithms such as RF, NB and SVM to correctly detect autism using items from a quantitative screening tool for autism likelihood in toddlers such as the Q-CHAT. Our results show that ML classifiers outperform standard statistical methods of classification in terms of sensitivity and specificity and are able to classify with very high accuracy autism versus TD children with a small subset of Q-CHAT items. In particular we found that ML algorithms were able to correctly detect autism with an accuracy above 90% from a selection of 14 items and above 80% using only 3 items. Furthermore, these 3 items were the best discriminating items already selected in the short form Q-CHAT-10. Taken together, these findings confirm the cross-cultural validity of the Q-CHAT as an early, quantitative screening tool for autism and the potential for the use of ML to improve the efficiency of screening tools, dramatically reducing the number of items. This aspect has important implications for clinical practice in primary care facilitating more effective and quick screening procedures to reach a significantly greater proportion of the population who are likely to be autistic.

Declarations

Ethics approval and consent to participate: The study received Ethics approval (prot.11/2012 and prot.E56/13) by the local committees and all the participants signed a written informed consent form to be enrolled in the study.

Consent for publication: Not applicable.

Availability of data and materials: The informed consent forms signed by the subjects prevent data from being publicly available. Data may be requested via email by researchers, upon reasonable request and verification of all ethical aspects, at: liliana.ruta@cnr.it

Competing interests: All authors declare no potential conflicts of interest, including any financial, personal or other relationships with other people or organizations relevant to the subject of their manuscript.

Funding: The original dataset, analyzed in the present study, was collected in a multicentric study, the "*Toddlers Project*", funded by the Ministry of Health and Tuscany Region (GR-2010-2319668). This research has been partially supported by the grant from the Sicilian Region of Italy (Assessorato Regionale dell'Istruzione e della Formazione Professionale, Avviso 11/2017 Rafforzare l'occupabilità nel sistema R&S e la nascita di spin off di ricerca in Sicilia – P.O. FSE 2014/2020, project no. 2014.IT.05.SFOP.014/3/10.4/9.2.10/0011 - CUP G47B17000100009, entitled SANi) and partially supported by the grant from Azienda Sanitaria Provinciale di Catania (deliberations no. 2005 24/06/2016 and no. 3304 04/10/2017). SBC was funded by the Autism Research Trust, the Wellcome Trust, the Templeton World Charitable Foundation, and the NIHR Biomedical Research Centre in Cambridge, during the period of this work. The Medical Research Council (MRC) funded the Cambridge Autism Research Database (CARD) that made this study possible. SBC also received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 777394. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and AUTISM SPEAKS, Autistica, SFARI. His research was also supported by the National Institute of Health Research (NIHR) Applied Research Collaboration East of England (ARC EoE) programme. The views expressed are those of the authors, and not necessarily those of the NIHR, NHS or Department of Health and Social Care.

Authors' contributions: GT conceptualized the study, worked on statistical data interpretation and drafted manuscript; LR designed and supervised the study and drafted the manuscript; GC and DD contributed with machine learning data analysis; EL and SA participated in the manuscript draft; FC, DV contributed with data interpretation; DB supervised the data analysis; AG, GMA, FA, FM, FM, CA and SBC reviewed the manuscript; *Toddlers Team* enrolled and tested the participants; GP contributed in the coordination of the study. All authors read and approved the final manuscript.

Acknowledgments: we are grateful to the families and children who participated in the study. ***Toddlers Team*** **Author List:** Francesca Isabella Famà, National Research Council of Italy (CNR) - Institute for Biomedical Research and Innovation (IRIB), Messina, Italy; Cristina Carrozza, National Research Council of Italy (CNR) - Institute for Biomedical Research and Innovation (IRIB), Messina, Italy; Natasha Chericoni, IRCCS Stella Maris Foundation, Calambrone, Pisa, Italy; Valeria Costanzo, IRCCS Stella Maris Foundation, Calambrone, Pisa, Italy; Nazarena Turco, Centro Autismo e Sindrome di Asperger ASLCN1, Mondovì (CN), Italy.

References

1. Elsabbagh M, Johnson MH. Getting answers from babies about autism. Trends Cogn Sci. Elsevier; 2010;14:81–7.
2. Thabtah F, Peebles D. Early Autism Screening: A Comprehensive Review. Int J Environ Res Public Health. Multidisciplinary Digital Publishing Institute; 2019;16:3502.
3. McPheeters ML, Weitlauf A, Vehorn A, Taylor C, Sathe NA, Krishnaswami S, et al. Screening for autism spectrum disorder in young children. Agency for Healthcare Research and Quality (US); 2016;

4. Dereu M, Roeyers H, Raymaekers R, Meirsschaut M, Warreyn P. How useful are screening instruments for toddlers to predict outcome at age 4? General development, language skills, and symptom severity in children with a false positive screen for autism spectrum disorder. Eur Child Adolesc Psychiatry. Springer; 2012;21:541–51.
5. Zwaigenbaum L, Penner M. Autism spectrum disorder: advances in diagnosis and evaluation. Bmj. British Medical Journal Publishing Group; 2018;361:k1674.
6. Magán-Maganto M, Jónsdóttir SL, Sánchez-García AB, García-Primo P, Hellendoorn A, Charman T, et al. Building a theoretical framework for autism spectrum disorders screening instruments in Europe. Child Adolesc Ment Health. Wiley Online Library; 2018;23:359–67.
7. Robins DL, Casagrande K, Barton M, Chen C-MA, Dumont-Mathieu T, Fein D. Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). Pediatrics. Am Acad Pediatrics; 2014;133:37–45.
8. Yuen T, Penner M, Carter MT, Szatmari P, Ungar WJ. Assessing the accuracy of the Modified Checklist for Autism in Toddlers: a systematic review and meta-analysis. Dev Med Child Neurol. Wiley Online Library; 2018;60:1093–100.
9. Oosterling I, Rommelse N, De Jonge M, Van Der Gaag RJ, Swinkels S, Roos S, et al. How useful is the Social Communication Questionnaire in toddlers at risk of autism spectrum disorder? J Child Psychol Psychiatry. Wiley Online Library; 2010;51:1260–8.
10. Stone WL, Coonrod EE, Ousley OY. Brief report: screening tool for autism in two-year-olds (STAT): development and preliminary data. J Autism Dev Disord. Springer Science & Business Media; 2000;30:607.
11. Matson JL, Wilkins J, Sharp B, Knight C, Sevin JA, Boisjoli JA. Sensitivity and specificity of the Baby and Infant Screen for Children with aUtIsm Traits (BISCUIT): Validity and cutoff scores for autism and PDD-NOS in toddlers. Res Autism Spectr Disord. 2009;3:924–30.
12. Allison C, Baron-Cohen S, Wheelwright S, Charman T, Richler J, Pasco G, et al. The Q-CHAT (Quantitative CHecklist for Autism in Toddlers): a normally distributed quantitative measure of autistic traits at 18–24 months of age: preliminary report. J Autism Dev Disord. Springer; 2008;38:1414–25.
13. Ruta L, Chiarotti F, Arduino GM, Apicella F, Leonardi E, Maggio R, et al. Validation of the Quantitative CHecklist for Autism in Toddlers (Q-CHAT) in an Italian clinical sample of young children with Autism and Other Developmental Disorders. Front Psychiatry. Frontiers; 2019;10:488.
14. Rutaa L, Arduino GM, Gagliano A, Apicella F, Leonardi E, Famà FI, et al. Psychometric properties, factor structure and cross-cultural validity of the quantitative CHecklist for autism in toddlers (Q-CHAT) in an Italian community setting. Res Autism Spectr Disord. Elsevier; 2019;64:39–48.
15. Devescovi R, Monasta L, Bin M, Bresciani G, Mancini A, Carrozzi M, et al. A Two-Stage Screening Approach with I-TC and Q-CHAT to Identify Toddlers at Risk for Autism Spectrum Disorder within the Italian Public Health System. Brain Sci. Multidisciplinary Digital Publishing Institute; 2020;10:184.

16. Wong HS, Huertas-Ceballos A, Cowan FM, Modi N, Group M for NI. Evaluation of early childhood social-communication difficulties in children born preterm using the Quantitative Checklist for Autism in Toddlers. J Pediatr. Elsevier; 2014;164:26-33. e1.
17. Magiati I, Goh DA, Lim SJ, Gan DZQ, Leong JCL, Allison C, et al. The psychometric properties of the Quantitative-Checklist for Autism in Toddlers (Q-CHAT) as a measure of autistic traits in a community sample of Singaporean infants and toddlers. Mol Autism. Springer; 2015;6:40.
18. Mohammadian M, Zarafshan H, Mohammadi MR, Karimi I. Evaluating reliability and predictive validity of the Persian Translation of Quantitative Checklist for Autism in Toddlers (Q-CHAT). Iran J Psychiatry. Tehran University of Medical Sciences; 2015;10:64.
19. 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 Psychiatry. Elsevier; 2012;51:202-212. e7.
20. Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry. Nature Publishing Group; 2012;2:e100–e100.
21. Bone D, Goodwin MS, Black MP, Lee C-C, Audhkhasi K, Narayanan S. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. J Autism Dev Disord. Springer; 2015;45:1121–36.
22. Levy S, Duda M, Haber N, Wall DP. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Mol Autism. Springer; 2017;8:65.
23. Constantino JN, Gruber CP. Social responsiveness scale: SRS-2. Western Psychological Services Torrance, CA; 2012.
24. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry. Nature Publishing Group; 2016;6:e732–e732.
25. Bone D, Bishop SL, Black MP, Goodwin MS, Lord C, Narayanan SS. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. J Child Psychol Psychiatry. Wiley Online Library; 2016;57:927–37.
26. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop SL. Autism diagnostic observation schedule, (ADOS-2) modules 1-4. Los Angel Calif West Psychol Serv. 2012;
27. Thabtah Fadi. ASDTests: a mobile app for ASD screening [Internet]. 2017. Available from: www.asdtests.com
28. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. Health Informatics J. SAGE Publications Sage UK: London, England; 2019;1460458218824711.
29. Akter T, Satu MS, Khan MI, Ali MH, Uddin S, Lio P, et al. Machine learning-based models for early stage detection of autism spectrum disorders. IEEE Access. IEEE; 2019;7:166509–27.
30. Breiman L. Random forests. Mach Learn. Springer; 2001;45:5–32.

31. Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley & Sons; 2012.
32. Vapnik V, Vapnik V. Statistical learning theory Wiley. N Y. 1998;1:624.
33. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. Montreal, Canada; 1995. p. 1137–45.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. JMLR. org; 2011;12:2825–30.
35. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. Springer; 2002;46:389–422.
36. https://www.nice.org.uk/guidance/cg128/chapter/appendix-signs-and-symptoms-of-possible-autism.
37. https://www.cdc.gov/ncbddd/autism/signs.html.

Tables

Table 1. Hyper-parameters values selected for SVM, RF and NB models

Model	Hyper-parameters	Values selected	Description Parameters
SVM	Kernel	linear	Linear kernel
	C	1	Cost
	γ	scale	Gamma
RF	Ntrees	100	Number of trees
	Criterion	entropy	Functions used to measure the quality of each split
	Max_features	auto	Number of features considered for each split
NB	Var_smoothing	10^{-9}	Smoothing parameter for calculation stability
	Fit_Prior	None	Class prior probabilities. In case of none, a uniform prior was used

Table 2. Comparison between the most discriminating 14 Q-CHAT items (ordered by rank) from the SVM-RFE algorithm and the most predicting 10 items (ordered by PPV) identified by Allison et al. [2012]

Q-CHAT 14 items (ordered by rank SVM-RFE)	Q-CHAT 10 items (Allison et al.) (ordered by PPV)
Does your child look at you when you call his/her name? (1)	Does your child look at you when you call his/her name? (1)
How easy is it for you to get eye contact with your child? (2)	How easy is it for you to get eye contact with your child? (2)
Does your child use simple gestures (e.g. wave goodbye)? (19)	Does your child use simple gestures (eg, wave goodbye)? (19)
Can other people easily understand your child's speech? (4)	Would you describe your child's first words as (typical): (17)
Does your child point to indicate that s/he wants something (eg, a toy that is out of reach) (5)	Does your child point to indicate that s/he wants something (eg, a toy that is out of reach) (5)
Does your child point to share interest with you (eg, pointing at an interesting sight)? (6)	Does your child point to share interest with you (eg, pointing at an interesting sight)? (6)
How long can your child's interest be maintained by a spinning object (e.g. washing machine, electric fan, toy car wheels)? (7)	Does your child follow where you're looking? (10)
Does your child pretend (e.g., care for dolls, talk on a toy phone)? (9)	Does your child pretend (e.g., care for dolls, talk on a toy phone)? (9)
Does your child do the same thing over and over again (e.g. running the tap, turning the light switch on and off, opening and closing doors)? (16)	Does your child stare at nothing with no apparent purpose? (25)
Would you describe your child's first words as (typical): (17)	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (eg, stroking their hair, hugging them)? (15)
When your child is playing alone, does s/he line objects up? (3)	
Does your child stare at nothing with no apparent purpose? (25)	
Does your child echo things s/he hears (e.g. things that you say, lines from songs or movies, sounds)? (18)	
How long can your child's interest be maintained by just one or two objects? (22)	

Table 3. Overall performance of the three selected machine learning classifiers (SVM, RF, NB) with respect to the original 25-items Q-CHAT, the 14-items selected by the SVM-RFE algorithm, the 10 items selected by

Allison and colleagues and the 3 most discriminating items in common to the two studies.

Model	Accuracy	MSE	Sensitivity	Specificity
(nr. selected features)				
SVM (25)	0.95 (+/- 0.02)	0.05	0.90	1.00
SVM (14)	0.93 (+/- 0.03)	0.08	0.87	0.96
SVM (10) [Allison et al.]	0.87 (+/- 0.03)	0.23	0.65	0.86
SVM (3)	0.83 (+/- 0.05)	0.13	0.78	0.93
RF (25)	0.90 (+/- 0.06)	0.09	0.85	0.95
RF (14)	0.88 (+/- 0.04)	0.13	0.82	0.89
RF (10) [Allison et al.]	0.84 (+/- 0.03)	0.17	0.69	0.93
RF (3)	0.83 (+/- 0.05)	0.15	0.83	0.86
NB (25)	0.89 (+/- 0.04)	0.08	0.82	1.0
NB (14)	0.88 (+/- 0.04)	0.11	0.74	1.0
NB (10) [Allison et al.]	0.82 (+/- 0.03)	0.15	0.65	1.0
NB (3)	0.84 (+/- 0.03)	0.17	0.78	0.86

Table 4. Detailed performance metrics of the three selected machine learning classifiers (SVM, RF, NB) in the test-set, with respect to the original 25-items Q-CHAT, the 14-items selected by the SVM-RFE algorithm, the 10 items selected by Allison and colleagues and the 3 most discriminating items in common to the two studies.

Model (nr. of selected features)	Classes	PPV	Sensitivity	F1-score	Nr. of subjects for clinical validation
SVM (25)	Autism	1.00	0.90	0.95	23
	TD	0.91	1.00	0.96	29
SVM (14)	Autism	0.95	0.87	0.91	23
	TD	0.90	0.97	0.93	29
SVM (10) [Allison et al.]	Autism	0.79	0.65	0.71	23
	TD	0.76	0.86	0.81	29
SVM (3)	Autism	0.90	0.78	0.84	23
	TD	0.84	0.93	0.89	29
RF (25)	Autism	0.89	0.74	0.81	23
	TD	0.82	0.93	0.87	29
RF (14)	Autism	0.86	0.83	0.84	23
	TD	0.87	0.90	0.88	29
RF (10) [Allison et al.]	Autism	0.89	0.70	0.78	23
	TD	0.79	0.93	0.86	29
RF (3)	Autism	0.83	0.83	0.83	23
	TD	0.86	0.86	0.86	29
NB (25)	Autism	1.00	0.70	0.82	23
	TD	0.81	1.00	0.89	29
NB (14)	Autism	1.00	0.74	0.85	23
	TD	0.83	1.00	0.91	29
NB (10) [Allison et al.]	Autism	1.00	0.65	0.79	23
	TD	0.78	1.00	0.88	29
NB (3)	Autism	0.82	0.78	0.80	23
	TD	0.83	0.86	0.85	29

Figures

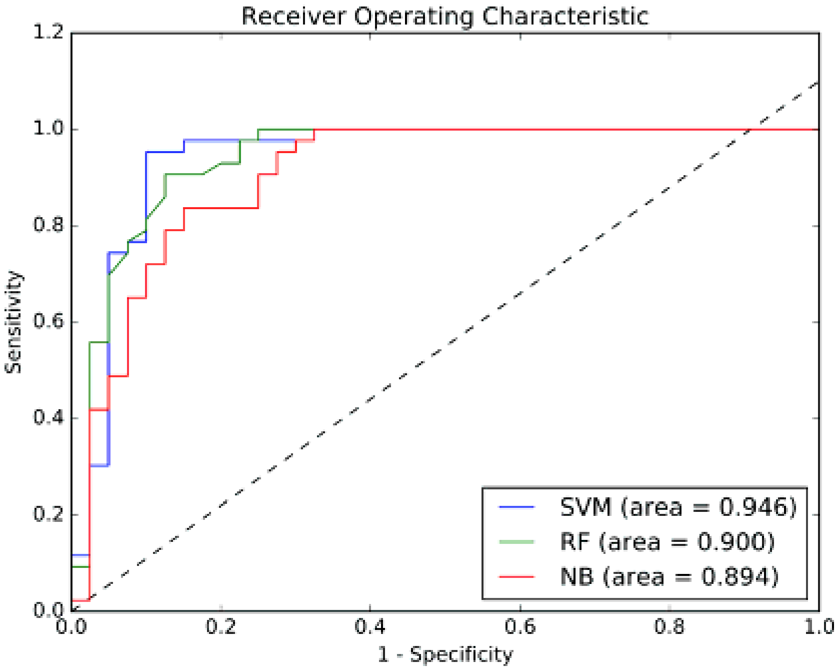

Figure 1

Area under the curve for the Q-CHAT (autism vs TD) comparing all 3 machine learning models

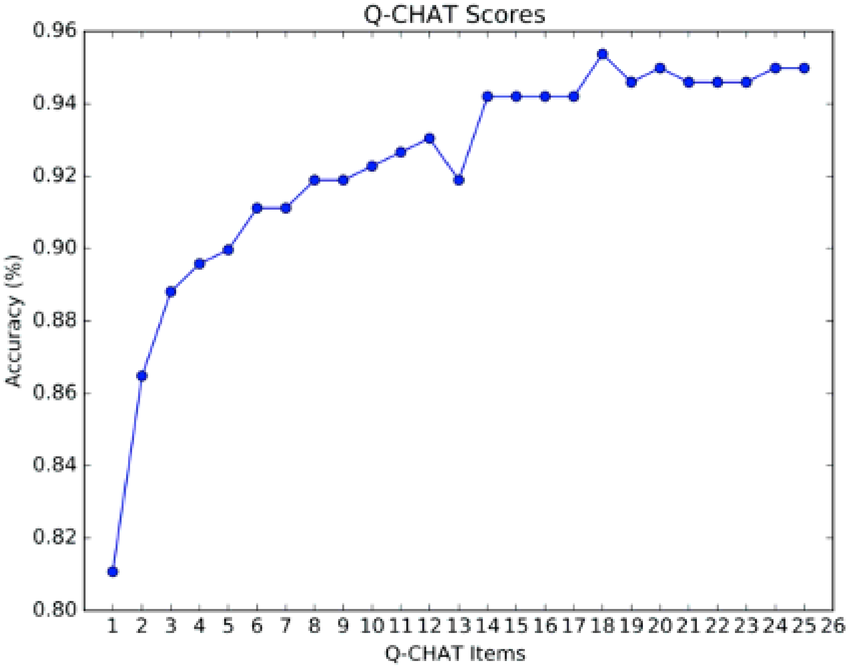

Figure 2

Accuracy selecting an increasing number of Q-CHAT items, using SVM-RFE algorithm