Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Dev Behav Pediatr. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

J Dev Behav Pediatr. 2019 June ; 40(5): 369–376. doi:10.1097/DBP.0000000000000668.

A Machine Learning Strategy for Autism Screening in Toddlers

Luke E. K. Achenie, Ph.D.[1], **Angela Scarpa, Ph.D.**[2,3], **Reina S. Factor, M.S.**[2,3], **Tao Wang, M.S.**[4], **Diana L. Robins, Ph.D.**[5], and **D. Scott McCrickard, Ph.D.**[6]

[1]Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA

[2]Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA

[3]Virginia Tech Center for Autism Research

[4]Department of Economics, University of California, Riverside, CA

[5]A.J. Drexel Autism Institute, Drexel University, Philadelphia, PA

[6]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA

Abstract

Objective: Autism Spectrum Disorder (ASD) screening can improve prognosis via early diagnosis and intervention, but lack of time and training can deter pediatric screening. The Modified Checklist for Autism in Toddlers, Revised (M-CHAT-R) is a widely used screener, but requires follow-up questions and error-prone human scoring and interpretation. We consider an automated machine learning (ML) method for overcoming barriers to ASD screening, specifically employing the feed-forward artificial neural network (fANN).

Method: The fANN technique was applied using archival M-CHAT-R data of 14,995 toddlers (16-30 months, 46.51% male). The 20 M-CHAT-R items were inputs, and ASD diagnosis after follow-up and diagnostic evaluation (i.e., ASD or not ASD) was output. The sample was divided into subgroups by race (i.e., White and Black), sex (i.e., boys and girls), and maternal education (i.e, below and above 15 years of education completed) to examine subgroup differences. Each subgroup was evaluated for best-performing fANN models.

Results: For the total sample, best results yielded 99.72% correct classification using 18 items. Best results yielded 99.92% correct classification using 14 items for White toddlers and 99.79% correct classification using 18 items for Black toddlers. In boys, best results yielded 99.64% correct classification using 18 items, while best results yielded 99.95% correct using 18 items in girls. For the case when maternal education is 15 years or less (i.e., Associate Degree and below), best results were 99.75% correct classification when using 16 items. Results were essentially the

Address correspondence to: Luke E. K. Achenie, PhD, Department of Chemical Engineering, 273 Signature Engineering Building, Room 273, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, [achenie@vt.edu], 540-231-4257.

Financial Disclosure: Diana L. Robins receives royalties for commercial products that incorporate the M-CHAT-R/F. The other authors have no financial disclosure.

Conflict of Interest: Diana L. Robins is co-owner of M-CHAT, LLC, which licenses commercial use of the M-CHAT-R/F. The other authors have no conflicts of interest to disclose.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

same when maternal education was 16 years or more (i.e., above Associate Degree); that is 99.70% correct classification was obtained using 16 items.

Conclusion: The ML method was comparable to the M-CHAT-R with follow-up items in accuracy of ASD diagnosis, while using fewer items. Therefore, ML may be a beneficial tool in implementing automatic, efficient scoring that negates the need for labor-intensive follow-up as well as circumvents human error, providing an advantage over prior screening methods.

Keywords

Autism spectrum disorder; machine learning; artificial neural network; early screening

INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder affecting 1 in 59 children,[1] and is associated with social communication challenges, social interaction difficulties, restricted and repetitive behaviors, and adaptive behavior impairments.[2] In addition to the personal and familial impact, the economic burden of pediatric ASD is substantial due to costs associated with increased use of health services, school supports, ASD-related therapy, family services, and caregiver time. Total societal costs in the United States for children with ASD were estimated at $11.5 billion in 2011.[3]

Early ASD screening and diagnosis leads to early intervention and improved prognosis; however, the average age of diagnosis in the United States is still after 4 years.[4] Additionally, children from rural areas, of a racial/ethnic minority, and of a lower socioeconomic status (SES) often receive a diagnosis later than children from urban areas, of ethnic majorities, and of higher SES.[5–8] As a result, delays in screening and diagnosis lead to missed opportunities for early intervention and improved outcomes, and children of various sociodemographic backgrounds may be at a particular disadvantage. Disparities in ASD screening have ignited efforts to improve access across diverse federal agencies, such as the CDC (Centers for Disease Control and Prevention) and early intervention programs.[9] While there are a number factors that contribute to access to diagnostic resources and interventions, previous literature suggests that further study of the specific variable of race can be informative to ASD diagnostic research.[6–8] Therefore, this manuscript will focus specifically on racial differences in ASD screening.

ASD and M-CHAT-R

The Modified Checklist for Autism in Toddlers (M-CHAT)[10] is a widely-used parent-report ASD screening instrument recommended by the American Academy of Pediatrics for use in primary care at 18 and 24 months.[11, 12] The M-CHAT provides accessible and low-cost screening; however, research indicates it may be less reliable in rural, minority, low socioeconomic samples with low education levels, similar to results found for ASD diagnosis broadly.[13] Additionally, one study found mothers with lower education levels and of racial minority status showed higher initial screen positive rates and were less likely to complete the follow-up interview, in part due to barriers such as phone numbers no longer working. [9]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The original M-CHAT consisted of 23 items and was administered in paper format as part of a 2-stage screener (see Measures section). Recently, the M-CHAT was revised (M-CHAT-R) [10] to reduce false positive responses (see Table 1), and improve clinical utility by identifying varying risk levels[14] to streamline the screening and referral process in busy pediatric settings. Those with low ASD risk are not further evaluated unless there are other reasons for concern, whereas parents of toddlers who score at medium risk are administered the structured Follow-Up (M-CHAT-R/F), consisting of additional interview questions to confirm risk.[11] High-risk toddlers are immediately referred for diagnostic evaluation and early intervention.

The M-CHAT-R/F is psychometrically strong, and most published studies to date continue to use a paper format. Recently, some studies have supported electronic administration of the M-CHAT-R/F.[12] Results suggest that primary care providers can administer the M-CHAT with Follow-Up reliably and efficiently during regular well-child visits using web-based administration,[13] and that administering the M-CHAT on a tablet in a primary care clinic increased acceptability of screening and quality of care.[14]

Despite the ease and widespread availability of the M-CHAT and other tools, ASD screening is still not as common in doctors' offices as one might hope. In one survey, 60% of pediatricians reported using formal ASD screening at 18 months and 50% at 24 months,[15] which is an increase from only 8% using ASD screeners in an earlier study.[16] The main reasons pediatricians reported not screening for ASD included lack of familiarity with tools, referral to a specialist, and lack of time. Pinto-Martin and colleagues[17] similarly identified lack of pediatric provider training and time as barriers to screening. In another study, healthcare providers noted similar barriers to screening (i.e., time, resources, ASD specific training) and also expressed the need for clear signs for ASD symptoms in early child development, screening tools appropriate for sociocultural differences, effective early intervention options, systems to handle potential increases in referrals, and continuing education.[18] In sum, perceived lack of time and training/familiarity with ASD screening tools are two primary barriers to routine ASD screening in pediatric practices.

This study aims to apply machine learning[18] (ML) to M-CHAT-R/F data to examine a potential alternative to assessment barriers assessment, in hopes that this method can improve the precision and application of risk detection in diverse populations in pediatric practices, and ultimately move beyond this setting. To minimize the technical nature of this paper, the bulk of the technical write-up is included in the appendix.

METHODS

Machine Learning

Machine Learning (ML)[1], which shares characteristics with artificial intelligence, is a powerful complementary clinical tool that employs large data sets to systematically learn patterns consistent with ASD traits. Current practices are inefficient in that the paper M-CHAT-R format must be hand-scored, interpreted, and then followed up with structured

[1]Details of the method and discussions are provided in the appendix.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

questions. The ML instrument is automatically scored, makes decisions objectively with minimal or no human bias, and does not require clinical training. Moreover, since ML is data-centric, it is expected to improve scoring as new data become available. A recent study employed an ML technique to classify children with ASD based on upper-limb movement and found ML was able to predict ASD classification with 96.7% accuracy. [18] In this paper, we used the R package (www.r-project.org) to implement the *Feedforward Artificial Neural Network (fANN)*, a ML approach that employs training, cross-validation (CV), and testing.

Participants

Total archival data consisted of 16,168 toddlers (16-30 months, collected during their 18- or 24- month well visit)[14]. Toddlers missing responses to any M-CHAT-R items were excluded; the total sample for the current study included 14,995 toddlers. Participants were 46.51% male, 44.8% female, 50.68% White, 20.30% Black, with 15 years average maternal education (Range = 11 to 20 years) for participants who provided this data (49.94% 11-15 years education, 50.06% 16-20 years education). Data included age at screening, gender, race, maternal education, M-CHAT-R responses, and evaluation outcome if required and obtained. Although data included American Indian, Hispanic, and Bi/Multiracial subjects, there were not enough participants of these races to train the fANN, and so these races/ethnicities have not been considered in the subgroup analyses in the present study.

Measures

Data were obtained from the M-CHAT-R/F validation study [14] in which the M-CHAT-R/F was administered. The M-CHAT-R/F is a 2-stage screener (see www.MCHATscreen.com and Supplemental Appendix), in which parents initially answered 20 yes/no questions. If children screened positive (i.e., 3 or higher), parents were asked structured follow-up questions by research personnel to obtain additional information and examples of at-risk behaviors. If children continued to score at or above the cutoff, they were referred for clinical evaluation (see Robins et al.).[14] Evaluation measures included the Autism Diagnostic Observation Schedule (ADOS), [19] Childhood Autism Rating Scale–2 (CARS-2), [19] Toddler Autism Symptom Interview,[19] Mullen Scales of Early Learning,[19] Vineland Adaptive Behavior Scales–II,[19] Behavioral Assessment System for Children–2, [19] and developmental history.

Procedures

Parents completed informed consent, demographics, and the M-CHAT-R, during their child's 18- or 24-month well-child visit. Pediatricians were asked to indicate ASD concern. Completed M-CHAT-R forms were scored by researchers who contacted parents of screen-positive children to complete follow-up. Children who continued to screen positive on the M-CHAT-R/F or whose physician had concerns were offered a diagnostic evaluation. Final diagnosis integrated all available information and used the psychologist/developmental pediatrician's clinical judgment to assess Diagnostic and Statistical Manual of Mental Disorders, 4th edition, text revision (DSM-IV-TR; APA, 2000) criteria for Autistic Disorder and Pervasive Developmental Disorder, Not Otherwise Specified. Children who did not meet ASD criteria were classified as typically developing or as having other developmental disorders or concerns.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Data Analytic Plan and Feature Selection

The current study used the 20 initial M-CHAT-R items as inputs in the fANN model described below. Therefore, *the follow-up questions were not included as inputs* in the analyses for the present study. By identifying only key features needed for ASD diagnosis, this eliminates redundant features, as we sought to use fewer questions to identify ASD risk-status. The three feature selection algorithms used in this paper were based on the T-test, entropy, or receiver-operating characteristic (ROC).

Model Selection

For every group, we developed 21 different ML models and the top three models, based on specific criteria, are reported in Table 2. Criteria included overall correct and incorrect classification percentage, sensitivity, specificity, PPV (positive predictive value), and NPV (negative predictive value). Based on psychometric criteria, we aimed to have a low overall incorrect classification percentage, with high correct classification and sensitivity. The best of the three models was chosen via CV, as explained in the appendix. [20, 21]

RESULTS

We completed several simulation trials (70% training; 30% testing) with the training data split into k-folds as discussed in the appendix. Best results are reported below and to compare the ML method with M-CHAT-R and M-CHAT-R/F, we also include results from Robins et al. [14] **Total Sample** *(*Tables 1 *and* 2*: Total Sample)*. In the first set of runs, we included all toddlers with complete M-CHAT-R (N = 14,995) data. The test group consisted of 4,498 participants selected by the program to test if the training was successful.

Best results were obtained with an *Entropy* feature selection of 18 features (see Table 2), yielding 99.72% overall correct classification, including 99.27% true negatives (TN), 0.45% true positives (TP), 0.16% false negatives (FN), 0.12% false positives (FP), and 78.90% PPV on test data; details of the calculation of TN, TP, FN, FP, and PPV are shown in Table 2. In several simulations (Table 2: Total Sample), the Entropy method had a slight advantage, yielding the highest overall correct classification with the lowest error rates (i.e., average value of error was 0.0484, see Figure 2a).

White Subgroup (Table 2*: White Subgroup – n with complete data = 8,195; test group subjects n = 2,459)*.

Best results were obtained using a *ROC* feature selection, and we were able to achieve 99.92% overall correct classification (99.59% TN, 0.33% TP, 0.08% FN, 0.00% FP) and 100% PPV using 14 features (Tables 1 and 2). T-test and ROC selection methods out-performed Entropy and therefore, there were two ROC analyses and one t-test conducted (see Table 2: White Subgroup). The average value of error being 0.0075 (Figure 2b).

Black Subgroup *(*Table 2*: Black Subgroup – n with complete data = 3,282, test group subjects n = 985)*.

Best results were obtained using 18 neurons in the input layer selected with a t-test and 15 neurons in the hidden layer (Table 1). With these parameters, we were able to produce

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

99.79% overall correct classification (98.88% TN, 0.91% TP, 0.21% FN, 0.00% FP) and 100% PPV (see Tables 1 and 2: Black Subgroup). After CV, the best model had an average value of error of 0.0505 (Figure 2c).

Male Subgroup (Table 2*: Male Subgroup – n with complete data = 6,966, test group subjects n = 2,089).*

Best results were achieved with 18 M-CHAT-R items in the input layer (Table 1) and 15 neurons in the hidden layer. We were able to produce 99.64% overall correct classification (98.98% TN, 0.66% TP, 0.27% FN, 0.09% FP) and 88.20% PPV (see Tables 1 and 2: Male Subgroup). The best CV model had an average value of error of 0.0130 (Figure 2d).

Female Subgroup (Table 2*: Female Subgroup – n with complete data n= 6,701, test group subjects n = 2,010).*

Best results were achieved with 18 neurons (selected with Entropy-test) in the input layer (Table 1) and 15 neurons in the hidden layer. We were able to produce 99.95% overall correct classification (99.72% TN, 0.23% TP, 0.05% FN, 0.00% FP) and 100% PPV (see Tables 1 and 2: Female Subgroup). The best CV model had an average value of error of 0.0169 (Figure 2e).

Maternal Education, 11 to 15 years of education Subgroup (Table 2: Education 11-15 Subgroup *– n with complete data n= 6,562, test group subjects n = 1,969).*

Maternal education was examined as a proxy for SES. Best results were achieved with 16 neurons (selected with T-test) in the input layer (Table 1) and 6 neurons in the hidden layer. We were able to produce 99.75% overall correct classification (99.04% TN, 0.71% TP, 0.20% FN, 0.05% FP) and 93.3% PPV (see Tables 1 and 2: Education Subgroup). The best CV model had an average value of error of 0.0097 (Figure 2f).

Maternal Education, 16 or more years of education Subgroup (Table 2*: Education 16-20 Subgroup – n with complete data = 6,715, test group subjects n = 2015).*

Best results were obtained using 16 neurons in the input layer (Table 1) selected with a ROC feature selection and 3 and 5 neurons in two hidden layers respectively. With these parameters, we were able to produce 99.70% overall correct classification (99.01% TN, 0.69% TP, 0.30% FN, 0.00% FP) and 100% PPV (see Tables 1 and 2: Education16-20 Subgroup). After CV, the best model had an average value of error of 0.0325 (Figure 2g).

DISCUSSION

Early screening of ASD can improve prognosis via early diagnosis and intervention.[21] We considered the fANN ML method, to improve upon paper hand-scoring of the M-CHAT-R and directly address disparities in ASD screening across diverse populations. [9] Performance was examined for the total sample as well as for subgroups of White, Black, male, female, and lower vs. higher maternal education. Results suggest that fANN can be used as an accurate and potentially improved method of M-CHAT-R analysis.[14] An additional contribution is the ability to tailor questions to diverse subgroups, allowing for future examination of which items are most appropriate to adapt the algorithm to an

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

individual. However, we acknowledge that race is a complex social construct and there are many nuances to be considered when interpreting racial analyses, such as diversity within racial groups. Therefore, future research should attempt to disentangle sociodemographic factors, such as SES or education that may relate to screening outcomes. The current study specifically examines race in the context of an ASD screener to inform future research on appropriate ASD screening for all demographic groups and mitigate racial/demographic disparities in diagnosis.

In our study, fANN scoring compares favorably to the M-CHAT-R/F scoring[14] (Table 3 in Robins et al.).[11] While the same data are used, the projects differ in the metrics used to determine outcome (i.e., ML uses black-box learning, M-CHAT-R uses explicit formulae). The overall correct classification percentage of 99.14% with the M-CHAT-R/F using 20 items in the original study was comparable to our results of 99.72% with 18 items in the total sample. Comparing sensitivity and specificity, the values obtained from the paper version of the M-CHAT-R/F[14] are 0.854 and 0.993, respectively, while values using fANN are 0.738 and 0.999. While the sensitivity is higher for the M-CHAT-R/F, it becomes more comparable in the ML approach when subgroups are analyzed. This finding suggests that with further testing and refinement, the ML approach might improve screening in specific demographic groups relative to the static M-CHAT-R/F scoring and eliminate the need for follow-up questions. The specificity is higher using fANN across all analyses. Comparison of the PPV values for the paper version (0.475) and fANN (0.789) suggests that the PPV is higher for the ML method. One limitation of the current study is the high rate of FN; future studies might prioritize reducing FN to maximize sensitivity. Future studies also should examine ML implementation in pediatric practices. Additionally, while the sample of participants were predominantly from urban and suburban areas, future work could include more participants from rural populations.

In sum, the fANN paradigm appears to be an effective scoring method for the M-CHAT-R. The primary finding provides evidence that ML offers an unbiased and automated way of scoring the M-CHAT-R. Advanced versions of the fANN would allow for refining the fANN structure and therefore, there is potential for determining ASD risk with more ease, accuracy, and specificity for different sociodemographic groups.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Source: This study was funded in part by the Virginia Tech Institute for Society, Culture and Environment (ISCE): Summer Scholars Program and by the Virginia Tech Center for Autism Research. Archival data used in analyses was funded by the Eunice Kennedy Shriver National Institute for Child Health and Human Development, R01 HD 039961.

References

1. CDC. Prevalence of Autism Spectrum Disorder Among Children Aged 8 year- Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2010. (http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6302a1.htm?s_cid=ss6302a1_w), Accessed March 28,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2014., in Morbidity and Mortality Weekly Report (MMWR) 2014, Centers for Disease Control and Prevention.

2. APA. Diagnostic and Statistical Manual (5th edition) DSM-5. 2013, Washington, DC: American Psychiatric Association.
3. Lavelle TA Economic burden of autism spectrum disorders in US. PharmacoEconomics & Outcomes News, 2014 697(1).
4. Baio J Prevalence of Autism Spectrum Disorders: Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. Morbidity and Mortality Weekly Report. Surveillance Summaries. . Centers for Disease Control and Prevention, 2012 61(3).
5. Rhoades RA, Scarpa A, & Salley B The importance of physician knowledge of autism spectrum disorders: Results of a parent survey. BMC Pediatrics, 2007 7.
6. Mandell DS, Listerud J, Levy SE, et al. Pinto-Martin JA, Race differences in the age at diagnosis among Medicaid-eligible children with autism. Journal of the American Academy of Child and Adolescent Psychiatry, 2002 41(12): p. 1447–1453. [PubMed: 12447031]
7. Mandell DS, Novak MM, & Zubritsky CD, Factors associated with age of diagnosis among children with autism spectrum disorders. Pediatrics, 2005 116 (6): p. 1480–1486. [PubMed: 16322174]
8. Morrier MJ, Hess KL, Heflin LJ, & Ethnic Disproportionality in Students with Autism Spectrum Disorders. Multicultural Education, Fall 2008 16(1): p. 31–38.
9. Daniels A, Halladay A, Shih A, et al. Approaches to Enhancing the Early Detection of Autism Spectrum Disorders: A Systematic Review of the Literature. Journal of the American Academy of Child and Adolescent Psychiatry, 2014 53(2): p. 141–151. [PubMed: 24472250]
10. Chlebowski C, Robins DL, Barton ML, et al. Large-scale use of the Modified Checklist for Autism in low-risk toddlers. Pediatrics, 2013 131(4): p. e1121–1127. [PubMed: 23530174]
11. Johnson CP, Myers S, Identification and evaluation of children with autism spectrum disorders. Pediatrics, , 2007 120: p. 1183–1215. [PubMed: 17967920]
12. AMERICAN_ACADEMY_OF_PEDIATRICS, Medical Home Initiatives for Children With Special Needs Project Advisory Committee - The Medical Home. Pediatrics, 2002 110: p. 184–186. [PubMed: 12093969]
13. Scarpa A, Reyes N, Patriquin MA, et al. The Modified Checklist for Autism in Toddlers: Reliability in a diverse rural American sample. Journal of Autism and Developmental Disorders, 2013.
14. Robins DL, Casagrande K, Barton, et al. Validation of the Modified Checklist for Autism in Toddlers, Revised With Follow-up (M-CHAT-R/F). Pediatrics, 2013 133(1): p. 37–45. [PubMed: 24366990]
15. Arunyanart W, Fenick A, Ukritchon S, et al. Developmental and Autism Screening: A Survey across Six States. Infants and Young Children, 2012 25(3): p. 175–187.
16. Dosreis S, Weiner CL, Johnson L, et al. Autism spectrum disorder screening and management practices among general pediatric providers. Journal of Developmental & Behavioral Pediatrics, 2006 27(2): p. S88–S94. [PubMed: 16685190]
17. Pinto-Martin JA, Dunkle M, Earls M, et al. Developmental stages of developmental screening: steps to implementation of a successful program. American Journal of Public Health, 2005 95(11): p. 1928. [PubMed: 16195523]
18. Kotsiantis SB, Supervised Machine Learning: A Review of Classification Techniques. Informatica, 2007 31: p. 249–268.
19. MathWorks_Inc., MATLAB and Statistics Toolbox, Release 2012b. 2012, Natick, MA.
20. Guyon I, Elisseeff A , An Introduction to Variable and Feature Selection. J. Mach. Learn. Res, 2003 3: p. 1157–1182.
21. Orinstein A, Helt M, Troyb E, et al. Intervention History of Children and Adolescents with High-Functioning Autism and Optimal Outcomes. J. Dev. Behav. Pediatr, 2014 35(4): p. 247–256. [PubMed: 24799263]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

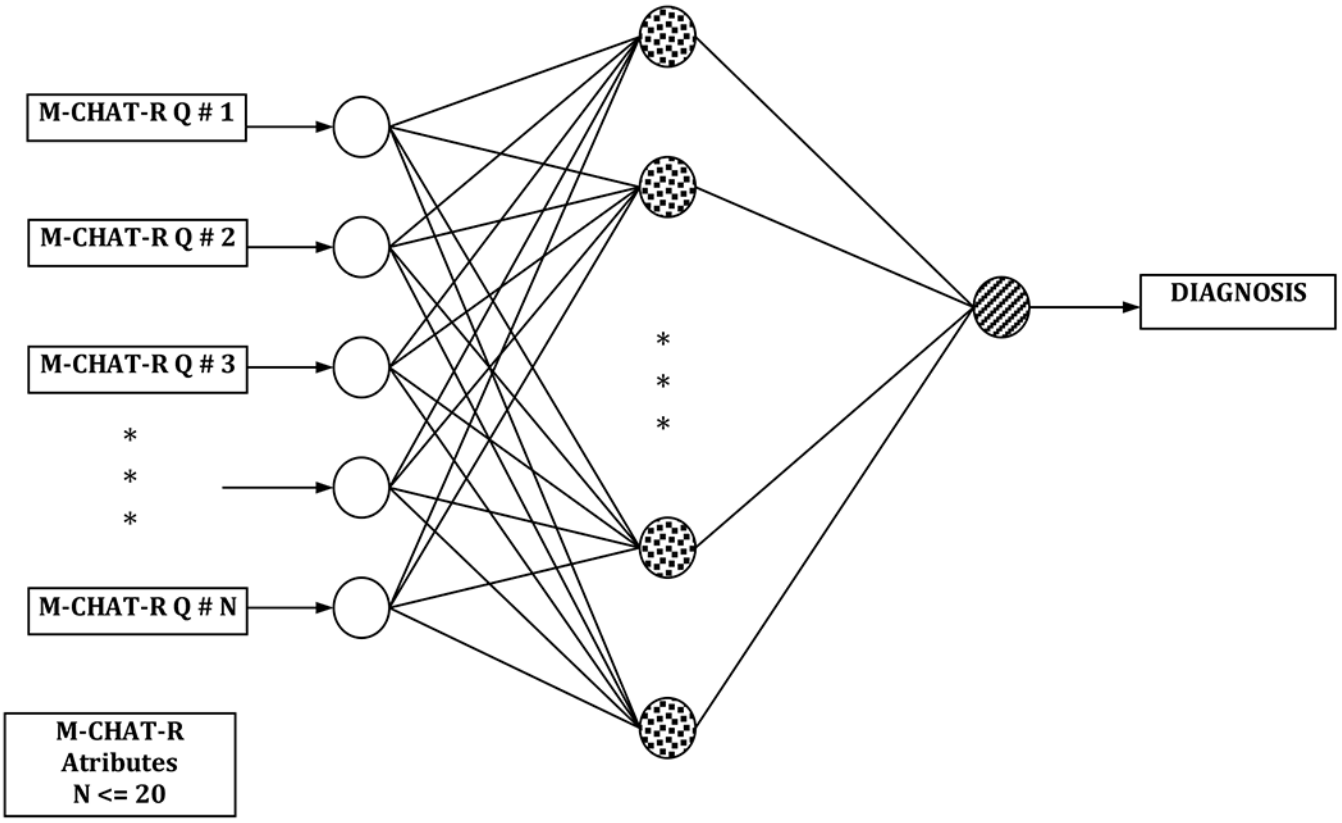

Figure 1. fANN Model:

This is a fANN model with three layers. The input layer (i.e., left most layer) consists of M-CHAT-R items; the hidden layer (i.e., middle layer) has several neurons, and the output layer (i.e., right most layer) has only one neuron associated with ASD diagnosis).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

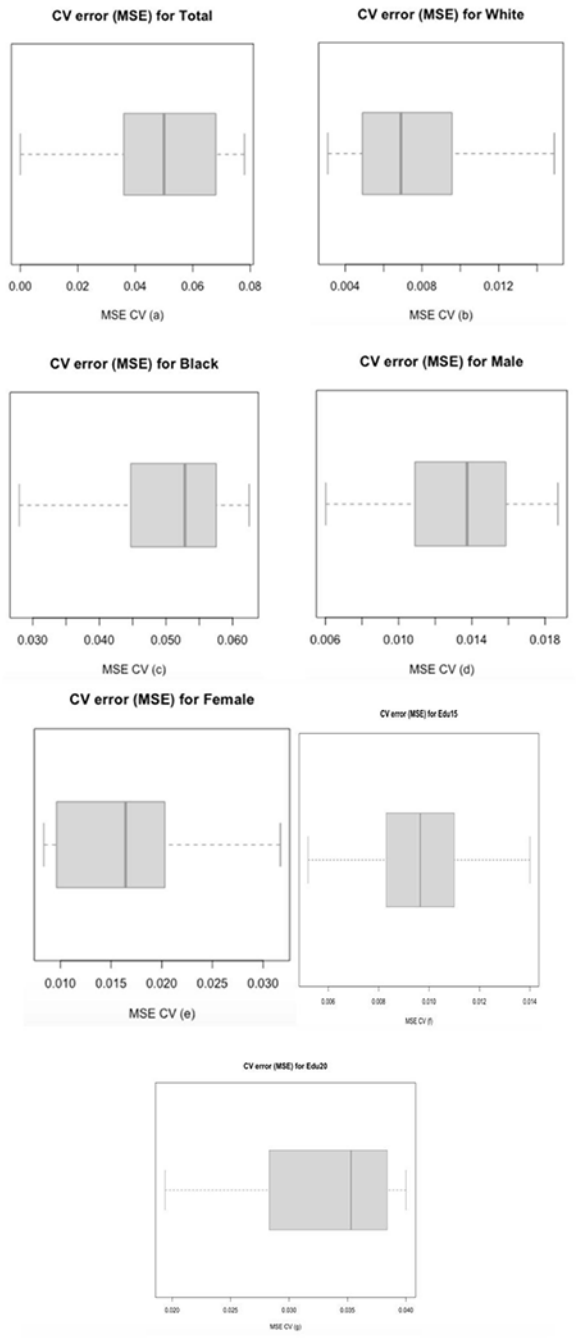

Figure 2. CV Error Box Plot:

a. CV error box plot for ***the Total Sample***. From Table 1, three different ML models were selected; using 10-fold CV for the three models to select the best one. CV process is completed until accuracy is determined for each instance in the dataset, and an overall accuracy estimate is provided, which is the 10 CV errors. The values of 10 CV errors were used to construct the above box plot; the deep blue line (center line) marks the middle value of errors, with the upper and lower limits of the box being the third and first quartile (75th and 25th percentile) respectively, while the ends of the whiskers are the minimum and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

maximum errors. The average value of errors for the 10-fold CV on model 1 was 0.0484, the lowest in the three ML models; the other two ML models (2 and 3) yielded 0.0505 and 0.0625. Therefore, we can conclude model 1 is the best model among these three models.

b. This is the CV error box plot for the ***White sample***. From Table 2, model 1 and 2 perform better than model 3; therefore a 10-fold CV was done for model 1 and 2. The basic function of 10-fold CV is explained in **a**. Based on the 10 CV errors, we drew the box plot, which has the same explanation as in **a**. The average value of errors for the 10-fold CV on model 1 is 0.0075 (the other model (2) is 0.0362). Model 1 is better than model 2.

c. This is the CV error box plot for the ***Black sample***. From Table 2, model 1 and 3 perform better than model 2. Same as explained in **a**, from a 10-fold CV for model 1 and 3, we got 10 CV errors and draw box plot, which has the same meaning as in **a**. The average value of errors of model 1 is 0.0505 (model (3) gave a value of 0.0852). Model 1 does better than model 3.

d. This is the CV error box plot for the ***male sample***. From Table 2, model 1 and 2 perform better than model 3. Using the explanation in **a**, the 10-fold CV on model 1 gave the above box plot and an average value of errors as 0.0130 (the other model (2) is 0.0283), which means model 1 performs better than model 3.

e. This is the CV error box plot for the ***female sample***. From Table 2, three different ML models were obtained. Using the explanation in **a**, the 10-fold CV gave the above box plot and an average value of errors of 0.0169 for model 1, the lowest in the three ML models; the other two ML models (2 and 3) gave 0.0261 and 0.0280. We conclude that model 1 is the best model among these three models.

f. This is the CV error box plot for the ***Education11-15 sample***. From Table 2, model 1 and 2 perform better than model 3; therefore a 10-fold CV was done for model 1 and 2. The basic function of 10-fold CV is explained in **a**. Based on the 10 CV errors, we drew the box plot, which has the same explanation as in **a**. The average value of errors for the 10-fold CV on model 1 is 0.0097 (the other model (2) is 0.0136). Model 1 is better than model 2.

g. This is the CV error box plot for the ***Education16-20 sample***. From Table 2, model 1 and 2 perform better than model 3. Same as explained in **a**, from a 10-fold CV for model 1 and 2, we got 10 CV errors and draw box plot, which has the same meaning as in **a**. The average value of errors of model 1 is 0.0325 (model (2) gave a value of 0.0192). Model 1 does better than model 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1:

M-CHAT-R Items[30]

		Total Sample	White	Black	Males	Females	Education11-15	Education16-20
0	Age (not an M-CHAT-R item but added here for convenience)							
1	Follows point	X	X	X	X	X	X	X
2	Concerns about deafness	X	X	X	X	X	X	X
3	Plays pretend or make-believe	X		X	X	X	X	X
4	Climbs on things	X				X		
5	Unusual finger movements near eyes			X	X			
6	Points with one finger to ask for something or to get help	X	X	X	X	X	X	X
7	Points with one finger to show something interesting	X	X	X	X	X	X	X
8	Interest in other children	X	X	X	X	X	X	X
9	Shows things - not to get help, but just to share	X	X	X	X	X	X	X
10	Responds when you call his or her name	X	X	X	X	X	X	X
11	Reciprocal smile	X	X			X		X
12	Upset by everyday noises			X	X	X	X	
13	Walks	X	X	X	X	X	X	X
14	Eye contact	X	X	X	X	X		X
15	Tries to copy actions	X	X	X	X	X	X	X
16	Follows gaze	X	X	X	X	X	X	X
17	Tries to get to watch him or her	X		X	X	X	X	X
18	Understands when tell him or her to do something	X	X	X	X	X	X	X
19	Social reciprocity	X	X	X	X	X	X	X
20	Likes movement activities	X		X	X		X	

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2:

Comparison of fANN parameters and feature selection methods for the fANN model

Sample	Run (#)	Overall correct classification (%)	Overall incorrect classification (%)	False Positive – Type I Error (%) / True Negatives (%)	True Positives (%) / False Negative – Type II Error (%)	Neurons in input layer (#)	Feature Selection Method	Neurons in hidden layer (#)	Sensitivity	Specificity	PPV	NPV
Total	1	99.72	0.28	0.12 / 99.27	0.45 / 0.16	18	Entropy	[3,5]	0.738	0.999	0.789	0.998
	2	99.76	0.24	0.04 / 99.35	0.41 / 0.20	13	ROC	[4,6]	0.672	0.999	0.911	0.998
	3	99.56	0.44	0.16 / 99.23	0.33 / 0.28	12	T-test	[2,6]	0.541	0.998	0.846	0.997
White subgroup	1	99.92	0.08	0.00 / 99.59	0.33 / 0.08	14	ROC	[3,5]	0.800	1	1	0.999
	2	99.92	0.08	0.00 / 99.59	0.33 / 0.08	16	ROC	6	0.800	1	1	0.999
	3	99.88	0.12	0.00 / 99.59	0.28 / 0.12	16	T-test	6	0.700	1	1	0.999
Black subgroup	1	99.79	0.21	0.00 / 98.88	0.91 / 0.21	18	T-test	15	0.818	1	1	0.998
	2	99.69	0.31	0.00 / 98.88	0.81 / 0.31	16	T-test	6	0.727	1	1	0.997
	3	99.79	0.21	0.00 / 98.88	0.91 / 0.21	18	Entropy	15	0.818	1	1	0.998
Male subgroup	1	99.64	0.36	0.09 / 98.98	0.66 / 0.27	18	T-test	15	0.714	0.999	0.882	0.997
	2	99.64	0.36	0.09 / 98.98	0.66 / 0.27	17	T-test	10	0.714	0.999	0.882	0.997
	3	99.61	0.39	0.04 / 99.02	0.59 / 0.35	17	Entropy	10	0.619	0.999	0.929	0.996
Female subgroup	1	99.95	0.05	0.00 / 99.72	0.23 / 0.05	18	Entropy	15	0.833	1	1	0.999
	2	99.95	0.05	0.00 / 99.72	0.23 / 0.05	17	Entropy	10	0.833	1	1	0.999

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Sample	Run (#)	Overall correct classification (%)	Overall incorrect classification (%)	False Positive – Type I Error (%) / True Negatives (%)	True Positives (%) / False Negative – Type II Error (%)	Neurons in input layer (#)	Feature Selection Method	Neurons in hidden layer (#)	Sensitivity	Specificity	PPV	NPV
	3	99.95	0.05	0.00 99.72	0.23 0.05	17	ROC	10	0.833	1	1	0.999
Education 11-15	1	99.75	0.25	0.05 99.04	0.71 0.20	16	T-test	6	0.778	0.999	0.933	0.998
	2	99.75	0.25	0.00 99.09	0.66 0.25	15	Entropy	[4,6]	0.722	1	1	0.997
	3	99.55	0.45	0.10 98.99	0.55 0.35	17	ROC	[4,5]	0.611	0.999	0.846	0.996
Education 16-20	1	99.70	0.30	0.00 99.01	0.69 0.30	16	ROC	[5,6]	0.700	1	1	0.997
	2	99.65	0.35	0.05 99.01	0.69 0.30	16	T-test	[3,5]	0.700	0.999	0.999	0.997
	3	99.65	0.35	0.05 98.97	0.69 0.30	16	T-test	[4,4]	0.700	0.999	0.999	0.997

Note: 1. For this table, FP (False Positive) indicates toddlers without ASD were incorrectly indicated as being at risk for ASD on the M-CHAT-R/F; TP (True Positives) indicate toddlers with ASD who were correctly diagnosed with ASD on the M-CHAT-R/F; TN (True Negative) indicates toddlers without ASD who were correctly identified as not having ASD risk, and FN (False Negative) indicates toddlers with ASD who were incorrectly identified as not as risk for ASD. Sensitivity = TP/(TP+FN); specificity = TN/(TN+FP); PPV (positive predictive value) = TP/(TP+FP), and NPV (negative predictive value) = TN/(TN+FN). **2.** Usually, PPV values are highly dependent on the population prevalence of the disease. If the test is applied when the proportion of people who truly have the disease is high, then the PPV values improve. Conversely, a very sensitive test (even one which is very specific) will have a large number of false positives if the prevalence of disease is low. It means unlike sensitivity and specificity, predictive value varies with the prevalence of the disease within the population.