




Using machine learning to improve diagnostic assessment of ASD in the light of specific differential and co-occurring diagnoses

Martin Schulte-Rüther,^{1,2}  Tomas Kulvicius,^{1,3} Sanna Stroth,⁴  Nicole Wolff,⁵
Veit Roessner,⁵  Peter B. Marschik,^{1,2,6,7} Inge Kamp-Becker,⁴ and Luise Poustka^{1,2}

¹Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center Göttingen, Göttingen, Germany; ²Leibniz ScienceCampus Primate Cognition, Göttingen, Germany; ³Department for Computational Neuroscience, University of Göttingen, Göttingen, Germany; ⁴Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, University Hospital of Marburg, Philipps-University Marburg, Marburg, Germany; ⁵Department of Child and Adolescent Psychiatry, TU Dresden, Dresden, Germany; ⁶Department of Women's and Children's Health, Center of Neurodevelopmental Disorders (KIND), Centre for Psychiatry Research, Karolinska Institutet, Stockholm, Sweden; ⁷iDN - interdisciplinary Developmental Neuroscience, Division of Phoniatrics, Medical University of Graz, Graz, Austria

Background: Diagnostic assessment of ASD requires substantial clinical experience and is particularly difficult in the context of other disorders with behavioral symptoms in the domain of social interaction and communication. Observation measures such as the Autism Diagnostic Observation Schedule (ADOS) do not take into account such co-occurring disorders. **Method:** We used a well-characterized clinical sample of individuals ($n = 1,251$) that had received detailed outpatient evaluation for the presence of an ASD diagnosis ($n = 481$) and covered a range of additional overlapping diagnoses, including anxiety-related disorders (ANX, $n = 122$), ADHD ($n = 439$), and conduct disorder (CD, $n = 194$). We focused on ADOS module 3, covering the age range with particular high prevalence of such differential diagnoses. We used machine learning (ML) and trained random forest models on ADOS single item scores to predict a clinical best-estimate diagnosis of ASD in the context of these differential diagnoses (ASD vs. ANX, ASD vs. ADHD, ASD vs. CD), in the context of co-occurring ADHD, and an unspecific model using all available data. We employed nested cross-validation for an unbiased estimate of classification performance and made available a Webapp to showcase the results and feasibility for translation into clinical practice. **Results:** We obtained very good overall sensitivity (0.89–0.94) and specificity (0.87–0.89). In particular for individuals with less severe symptoms, our models showed increases of up to 35% in sensitivity or specificity. Furthermore, we analyzed item importance profiles of the ANX, ADHD, and CD models in comparison with the unspecific model revealing distinct patterns of importance for specific ADOS items with respect to differential diagnoses. **Conclusions:** ML-based diagnostic classification may improve clinical decisions by utilizing the full range of information from detailed diagnostic observation instruments such as the ADOS. Importantly, this strategy might be of particular relevance for older children with less severe symptoms for whom the diagnostic decision is often particularly difficult. **Keywords:** Autism spectrum disorders; diagnosis; Machine learning.

Introduction

Autism spectrum disorder (ASD) is an umbrella term for a set of highly heterogeneous neurodevelopmental conditions characterized by impairments in social interaction and communication and restricted, repetitive behaviors (American Psychiatric Association, 2013). The estimated prevalence is ~1% of the population (Baxter et al., 2015) with a considerable amount of overlap with other disorders affecting social interaction (Hossain et al., 2020; Thom, Keary, Kramer, Nowinski, & McDougle, 2020). These characteristics call for efficient yet comprehensive diagnostic procedures.

Diagnosing ASD is a challenging and time-consuming task and requires a high level of clinical expertise and experience. The combination of behavioral observation, anamnestic interviews (e.g.,

Autism Diagnostic Interview-Revised, ADI-R, Lord, Rutter, & Le Couteur, 1994), and additional clinical information (e.g., co-occurring disorders, differential diagnoses, cognitive abilities, and neuropsychological impairment) is considered the diagnostic 'gold standard'. The Autism Diagnostic Observation Schedule (ADOS, Lord et al., 2012; Poustka et al., 2015) is typically used to assess current behavior using age-adapted modules (from toddlers to young adults) and provides a cutoff score for ASD. Such direct observation via a structured social interactive encounter is of utmost importance not only for ASD-specific symptomatology, but also for co-occurring disorders and the evaluation of potential differential diagnoses.

Diagnostic decisions in ASD are particularly challenging because difficulties in social interaction and communication are common also for a range of other conditions and behaviors including affective and anxiety disorders (Tyson & Cruess, 2012; van Steensel, Bögels, & Wood, 2013; Wittkopf et al., 2021),

Conflict of interest statement: See Acknowledgements for full disclosures.

attention deficit hyperactivity disorder (ADHD, (Ros & Graziano, 2018), and conduct disorder (CD, Gilmore, Hill, Place, & Skuse, 2004; Milledge et al., 2019), all potential alternative diagnoses to ASD. On the other hand, the prevalence of such co-occurring disorders, among individuals with ASD, is high (Hossain et al., 2020). Thus, specific behavior of an individual may be due to ASD alone, a differential disorder, or overlap of both. This decision is particularly difficult for children with less severe ASD symptoms (Davidovitch, Levit-Binnun, Golan, & Manning-Courtney, 2015).

Recent work has explored machine learning (ML) methods to support diagnostic procedures (see Hyde et al., 2019 for a review). Most of these studies aimed at streamlining the diagnostic procedures for ASD by identifying most discriminative items from clinician-coded examinations such as ADOS-2 and ADI-R (Bone et al., 2016; Duda, Kosmicki, & Wall, 2014; Duda, Daniels, et al., 2016; Kosmicki, Sochat, Duda, & Wall, 2015; Küpper et al., 2020; Levy, Duda, Haber, & Wall, 2017; Wall et al., 2012; Witkopf et al., 2021, see Bone et al., 2015 for a critical review). ML models built upon such selected items often retain or even exceed the diagnostic accuracy of the original ADOS-2 algorithm. Identifying subsets of items and exploring their stability and generalizability (Levy et al., 2017) is important to develop time-efficient and sensitive screening instruments (Duda, Daniels, et al., 2016; Kamp-Becker et al., 2017). However, it is an open question whether these would provide enough specificity in the light of differential diagnostic decisions (Bone et al., 2015). Recent prospective studies aimed to validate shortened parent interviews are promising (Duda, Daniels, et al., 2016), but the results for shortened observation-based approaches (Abbas, Garberson, Glover, & Wall, 2018; Fusaro et al., 2014; Tariq et al., 2018) are mixed.

We suggest a complementary approach to exploit the information from the entire coding scheme of the ADOS in order to maximize classification accuracy in the light of specific differential diagnosis and co-occurring disorders. For example, atypical eye contact or less initiative during social interaction is typical for obviously anxious individuals, whereas talkative behavior or a lack of empathic responding may be present in individuals with pronounced externalizing behaviors. However, the standard ADOS algorithm does not take such considerations into account. Providing specific diagnostic algorithms for respective contexts of co-occurring disorders and differential diagnosis would be an essential step in clinical practice and likely increase the overall quality of ASD-specific assessment. This is particularly important for ADOS module 3 because the respective age group (~ages 5–16) presents a particular challenge: First, individuals who seek ASD-specific diagnostic service at an older age tend to have less pronounced ASD symptoms than individuals with earlier

diagnostic visits (Davidovitch et al., 2015). Second, the onset of co-occurring disorders such as anxiety-related disorders, ADHD, and CD is most prevalent at this age (Hossain et al., 2020), often resulting in decreased accuracy of ASD-specific diagnostic procedures (Kamp-Becker et al., 2018).

Thus, the present study aimed, for the first time, at using ML methods to train optimized models for ASD diagnosis in the light of specific co-occurring disorders and differential diagnoses (i.e., anxiety, ADHD, and CD). We tested (a) whether such models improve sensitivity and specificity of classification in comparison with the ADOS-2 algorithm, (b) whether this would be particularly pronounced for those individuals with less severe symptoms of ASD. Furthermore, and (c) we tested whether item importance profiles of the specific models would reveal those ADOS items which are particularly relevant for the differential diagnostic decision of ASD versus anxiety, ADHD, or CD.

Materials and methods

Dataset and preprocessing

Data collection. We used item-level data of the ADOS-G/ADOS-2 module 3 (see Table S1 for a list of ADOS items), representing a subsample of data from a German data repository (ASD-Net, Kamp-Becker et al., 2017). All participants visited one of four specialized outpatient clinics for ASD, integrated within a full-care University Hospital in Germany (Marburg, Dresden, Mannheim, Göttingen) and had been referred due to suspected ASD. A clinical best-estimate diagnosis of ASD was either confirmed or excluded following established guidelines of 'gold standard' diagnostic evaluation of ASD (AWMF, 2016; Falkmer, Anderson, Falkmer, & Horlin, 2013). Similarly, co-occurring ICD-F diagnoses were secured. Experienced clinicians with continuous ADOS coding experience performed the coding of all items. Across the sample, we used the ADOS-2 algorithm of module 3. This algorithm is a summed score of a subset of items (cutoff for autism spectrum: 7 or higher and cutoff for autism: 9 or higher) and can be translated to an age-adapted calibrated severity score (corresponding cutoffs: 4[autism spectrum], 6 [autism]).

Ethics considerations. The ethics committee of the medical faculty, Philipps-University Marburg, gave ethics approval for this work. Due to the retrospective nature of data collection and analysis based on anonymized data, the need for informed consent was waived by the ethics committee.

Sample description. The dataset comprised 1,251 participants (mean age: 10.05 + - 2.73SD), of whom 481 had received an ASD diagnosis (i.e., ICD F84.0, F84.1 or F84.5; a few individuals ($n = 11$) with a diagnosis of F84.9 or F84.8 were excluded). Additional ICD10 F diagnoses comprised ADHD (Attention Deficit Hyperactivity Disorder, $n = 439$), CD (Conduct Disorder, $n = 194$), anxiety-related disorders (ANX, $n = 122$), and further ICD-F diagnoses (OTHER, $n = 233$, i.e., neither of the diagnostic labels ADHD, ANX, or CD). These diagnoses were partly overlapping with an ASD diagnosis and with each other. Some individuals did not receive a diagnosis of ASD but had no further assessment to secure a differential ICD-F diagnosis (NONE, $n = 211$). For further details, see Figure 1, Table 1, and Table S2.

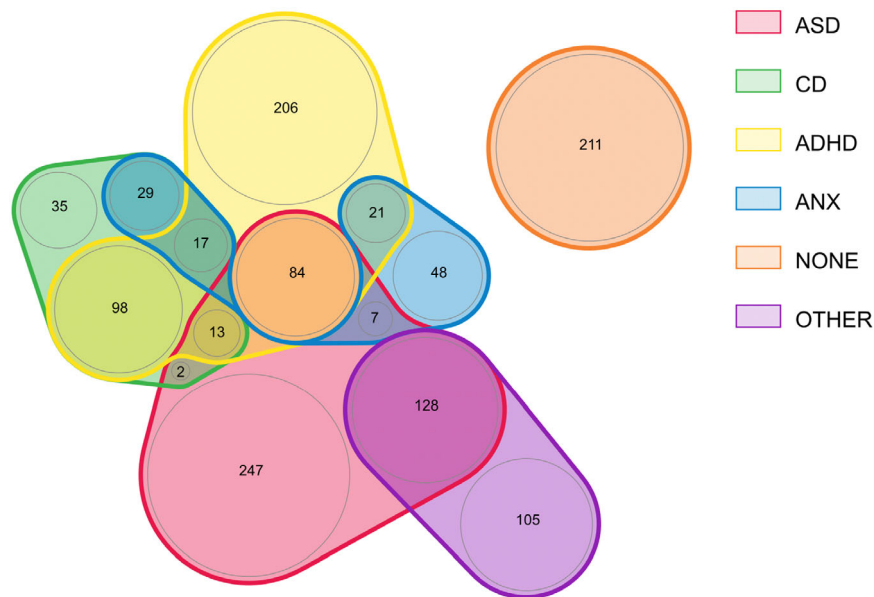


Figure 1 Extent of overlap for diagnostic categories across the sample. For simplicity, the figure only shows overlap across ASD, ADHD, ANX, and CD

Missing data. Individuals with more than 50% missing data within the database ($n = 12$) were discarded. The remaining sample had 1.25% missing data across all cells. 90.41% of individuals had no missing data at all.

Classification models

Item ratings of the ADOS module 3 were used as predictors and the ASD *best-estimate clinical diagnosis* (i.e., not the ADOS implied classification) was the target for classification. Item ratings of 7/8 were recoded to 0, similar to the original ADOS algorithm, whereas item ratings of 1–3 were retained to exploit the full range of coded symptom severity for the classification models (i.e., 3 not recoded to 2, see Figure S1 for the distributions of item codes across groups). We trained random forest models (Breiman, 2001; Wright & Ziegler, 2017), since these provide excellent accuracies for disease prediction from health data (Uddin, Khan, Hossain, & Moni, 2019). We used two different types of models: The first type of models used the full $n = 481$ participants with ASD, but differed with respect to the non-ASD category (ANX models: Non-ASD[ANX] $n = 115$, ADHD models: non-ASD[ADHD] $n = 342$, CD models: non-ASD [CD] $n = 179$, and unspecific models: non-ASD[ANX/ ADHD/ CD/ OTHER/ NONE] $n = 770$), to set up models that may provide insight into the clinical question ‘do the ADOS item ratings suggest a diagnosis of ASD, if there is information available, or a clinical impression the behavior might be explained by a differential diagnosis’. Additionally, we trained models to differentiate between ASD and non-ASD in participants with co-occurring ADHD (ADHD_{co} models: $n_{ASD} = 97$, $n_{Non-ASD} = 342$) to set up a model that may provide insight into the clinical question ‘if a secured diagnosis of ADHD is available, do the item ratings suggest a diagnosis of ASD or not?’. For participants with ANX or CD, this type of model could not be trained because sample sizes were too low ($n_{ANX + ASD} = 7$, $n_{CD + ASD} = 15$). See Table 1 for the sample characteristics of all models.

We employed a strict nested cross-validation approach (optimization of hyperparameters m_try , min_n , and n_trees within inner folds [5x5 repeated cross-validation], model evaluation within outer folds [10 x 5 repeated cross-validation]) to provide an unbiased estimation of model performance and safeguard against overfitting (Vabalas,

Gowen, Poliakoff, & Casson, 2019). Removing of zero-variance predictors and imputation of missing values (bagged-tree imputation) was performed separately for each iteration of parameter optimization and model evaluation to prevent information leakage. Area under the receiver operating characteristic curve (AUC-ROC) was used for optimization and evaluation of models because it is robust against potential distortions during training of small or unbalanced datasets. Evaluation was performed for the full sample and re-calculated for individuals with less severe symptoms (i.e., including only those participants with calibrated severity scores of 2–5, i.e., around the cutoff score 4 for autism spectrum). Evaluation results for the ANX, CD, ADHD, ADHD_{co}, and unspecific models were compared with the standard ADOS score across evaluation folds using Wilcoxon signed-rank tests for paired data (effect sizes were determined using Cliff’s delta). For further details, see Appendix S1, Figures S2, S3, and Table S3. Item importance scores were calculated for each model (Breiman, 2001), normalized, and compared across folds and between model types to test for the specificity of item importance profiles for the specific models (Mann–Whitney tests for each item). For demonstration purposes, we created a web app (<https://msrlab.shinyapps.io/asd-ml-jcpp/>) to showcase the final models. Note, this is not a ready-to-use diagnostic tool but provides a demonstration of feasibility for translation into clinical practice. See Appendix S1 and Figure S4 for more details.

Results

Model evaluation

We observed significantly better performance for the new models in comparison with the ADOS-2 algorithm. For all models, mean AUC was significantly higher for the random forest models than for the ADOS algorithm with large effect sizes (all $p < 1.2 \times 10^{-9}$, cliff’s delta >0.756 , mean increases of 2.7–12.3% for AUC). In particular, the models performed better for those participants with lower ADOS severity levels (see Figure 2, Figure S2 for the

Table 1 Sample characteristics for the ASD and non-ASD groups within respective specific models (ANX, CD, ADHD, and ADHD_{co})

Model	<i>n</i>	Age (mean ± SD, range)	Wilcoxon (<i>p</i>) Cliff's delta	IQ (mean ± SD, range)	Wilcoxon (<i>p</i>) Cliff's delta	No IQ available (<i>n</i>)	Male	Female	χ^2 male / female odds ratio
Unspecific model									
ASD	481	10.3 ± 2.88 5-27	<i>p</i> = .0112 delta = 0.0847 Negligible	99.1 ± 17.0 46-143	<i>p</i> = .426 delta = 0.0299 Negligible	74	438	43	<i>p</i> = .0761 or = 1.43
Non-ASD	770	9.89 ± 2.62 4-18	Negligible	100 ± 17.4 56-148	Negligible	203	675	95	
CD models									
ASD	Diagnoses: ASD and/or F90.1 (hyperkinetic CD, <i>n</i> = 104), F91 (CDs, <i>n</i> = 48), F92 (mixed disorders of conduct and emotions, <i>n</i> = 46) 481 (15 CD)	10.3 ± 2.88 5-27	<i>p</i> = .222 delta = 0.0615 Negligible	99.1 ± 17.0 46-143	<i>p</i> = .559 delta = 0.0337 Negligible	74	438	43	<i>p</i> = .613 or = 1.21
Non-ASD	179 (all CD)	9.98 ± 2.43 5-16	Negligible	100.0 ± 18.2 62-148	Negligible	46	160	19	
ANX models									
ASD	Diagnoses: ASD and/or F40 (Phobic anxiety disorders, <i>n</i> = 7), F41 (Other anxiety disorders, <i>n</i> = 3), F42 (Obsessive-compulsive disorder, <i>n</i> = 14), F92.8 (other mixed disorders of conduct and emotions, <i>n</i> = 34), F93.[0-2,8] (Emotional disorders with onset specific to childhood, <i>n</i> = 65) [Separation anxiety, Social anxiety, Phobic anxiety, generalized anxiety]) 481 (7 ANX)	10.3 ± 2.88 5-27	<i>p</i> = .319 delta = 0.0594 Negligible	99.1 ± 17.0 46-143	<i>p</i> = .490 delta = 0.0465 Negligible	74	438	43	<i>p</i> = .0497 or = 1.89
Non-ASD	115 (all ANX)	10.5 ± 2.27 6-17	Negligible	101 ± 16.8 67-141	Negligible	25	97	18	
ADHD-models									
ASD	Diagnoses: ASD and/or F90.0 (Attention deficit Hyperactivity disorder, <i>n</i> = 335), F90.1 (Hyperkinetic CD, <i>n</i> = 104) 481 (97 ADHD)	10.3 ± 2.88 5-27	<i>p</i> = .00467 delta = 0.115 Negligible	99.1 ± 17.0 46-143	<i>p</i> = 0.732 delta = 0.0157 Negligible	74	438	43	<i>p</i> = .687 or = .873
Non-ASD	342 (all ADHD)	9.75 ± 2.44 5-16	Negligible	99.7 ± 17.9 62-148	Negligible	80	315	27	
ADHD-comorbidity models									
ASD	Diagnoses: ADHD (F90.0 (Attention deficit Hyperactivity disorder, <i>n</i> = 440), F90.1 (Hyperkinetic CD, <i>n</i> = 102) and ASD or Non-ASD) 97 (all ADHD)	10.0 ± 2.37 5-17	<i>p</i> = .236 delta = 0.0783 Negligible	97.9 ± 16.0 58-131	<i>p</i> = .487 delta = 0.0521 Negligible	20	87	10	<i>p</i> = .583 or = .746
Non-ASD	342 (all ADHD)	9.75 ± 2.44 5-16	Negligible	99.7 ± 17.9 62-148	Negligible	80	315	27	

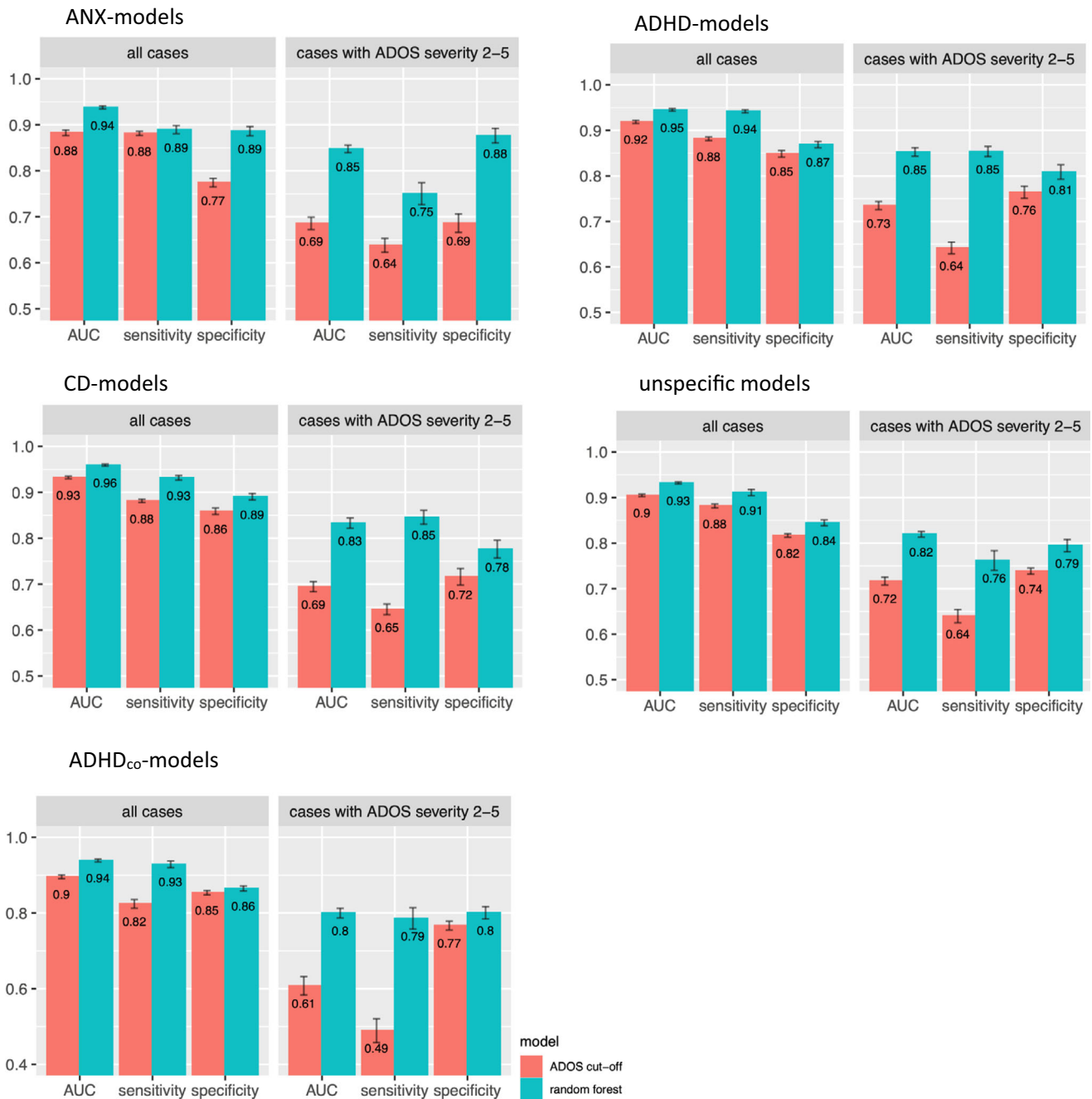


Figure 2 Model evaluation metrics (area under the receiver operating characteristic curve [AUC], sensitivity and specificity) of ANX, CD, ADHD, unspecific, and ADHD_{co} models. Error bars depict the standard error of means across evaluation folds of the cross-validation procedure. Left panels: all available cases, right panels: cases with ADOS severity score 2–5

respective ROC curves, and Table S3 for a complete list of model evaluation scores).

For the full sample, sensitivity was particularly increased for the CD models and ADHD models (CD +4.8%, ADHD +5.9%, ADHD_{co} + 10.4%, large effect sizes, ANX +1.4% small effect size). For those individuals with less severity (i.e., ADOS severity levels 2–5), the increase in specificity was even stronger (CD +19%, ADHD +21.2%, ADHD_{co} + 35.7%, ANX +13.7%, large effect sizes). Increases in specificity were more pronounced for ANX models for both the full sample (ANX_f + 11%, large effect size, CD +3.8% medium effect size, ADHD +2% small effect size,

ADHD_{co} + 1.1%, negligible effect size) and for those individuals with less severity (ANX +8.5%, large effect sizes, CD +2.6%, ADHD +1.7%, ADHD_{co} + 1.0%, small effect sizes).

Similarly, positive likelihood ratios were significantly improved for the random forest models with medium to large effect sizes for the full sample (all $p < .00017$, cliff's delta >0.3392 , mean increases of LR+ for CD +2.91, ADHD +1.83, ADHD_{co} + 1.76, ANX +5.34) and large effect sizes for the sample with less severity (all $p < 1.88 \cdot 10^{-6}$, mean increase in LR+ for CD +6.68, ADHD +6.435, ADHD_{co} + 8.150, ANX +5.66).

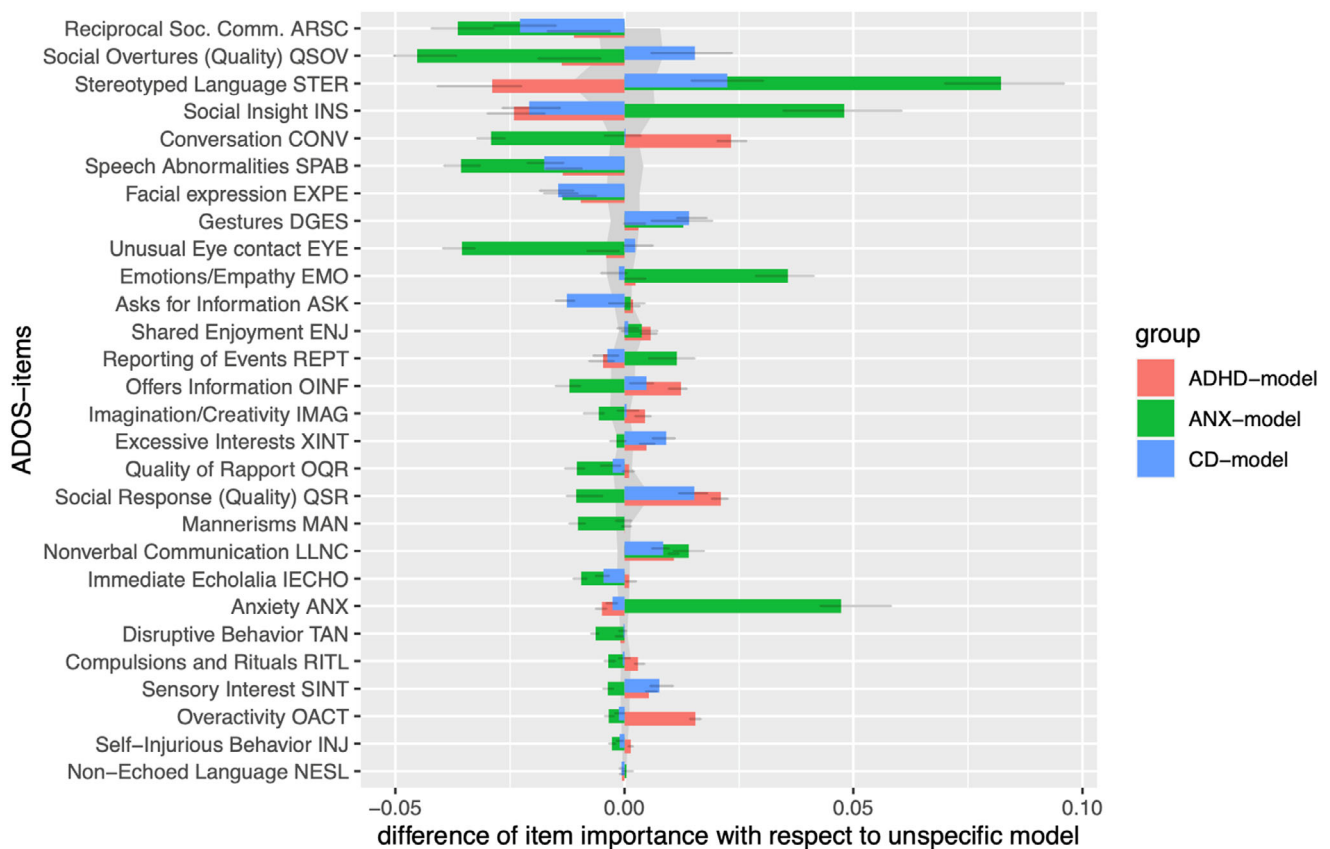


Figure 3 Difference plots of normalized item importance for ANX, CD, and ADHD models, with respect to the unspecific models (zero-line). Error bars indicate a 99% rank-based confidence interval of the median. The gray shading indicates a 99% rank-based confidence interval for the unspecific models. Items are sorted along the y-axis with increasing importance for the unspecific models from bottom to top (see also Figure S5). See Table S1 for the unabbreviated list of ADOS items

Item importance

Comparing ANX, CD, and ADHD models with the unspecific models for all participants, we could observe unique patterns of item importance (see Figure 3 and Figure S5). The item profiles were pretty stable across evaluation model folds, with significant differences for most items (multiple Mann–Whitney tests) in comparison with the profile for the unspecific models ($p < .05$ Bonferroni-corrected for multiple comparisons, see Appendix S2 for details). For a visualization of raw item ratings separately for each diagnostic group, see Figure S1.

Discussion

In this study, we aimed to train optimized ADOS-based models for ASD classification in the light of typical co-occurring disorders and differential diagnoses using ML methods. We provide evidence for improved sensitivity and specificity in comparison with the ADOS algorithm, in particular for individuals with less severe symptoms. Furthermore, we could reveal specific item importance profiles for the classification of ASD against anxiety-related disorders, ADHD, and CD, respectively.

Improved diagnostic accuracy

Model performance. Most previous ML approaches used very small samples of non-ASD individuals (typically less than 10% of analyzed cases) or a clinical best-estimate diagnoses was not available for all ASD individuals (Duda et al., 2014; Kosmicki et al., 2015; Levy et al., 2017; Wall et al., 2012); thus, the possibility of biased or overly optimistic results could not be completely ruled out for some of these findings. (Bone et al., 2015). In addition to these important earlier findings, our results support the idea that ML is a promising tool (Hyde et al., 2019) for *typical populations in clinical practice* and may ultimately increase the diagnostic accuracy of standard ADOS assessments by utilizing the full range of information available (see Bone et al., 2015 for similar findings). Thus, our approach is complementary to previous attempts (focused on the development of new screening instruments by identifying most informative ADOS items, e.g. Duda et al., 2014; Küpper et al., 2020; Levy et al., 2017; Wall et al., 2012). Within our well-characterized clinical sample, our ML models showed excellent performance to classify ASD (0.94–0.96 AUC, 0.90–0.94 sensitivity, 0.87–0.90 specificity, 8.0–10.3 positive likelihood ratio) and performed significantly

better than the ADOS-2 algorithm, which is comparable to other studies with available clinical best-estimate diagnosis (e.g., Duda et al., 2014). Note, previous studies reporting even higher accuracies, sometimes used the ADOS score and not the clinical diagnosis as their classification target, or a mixture of both. (e.g., Kosmicki et al., 2015; Levy et al., 2017), which may inflate performance measures (Bone et al., 2015).

ASD classification in the context of differential diagnoses and co-occurring disorders. Optimized classification algorithms for ASD in the context of co-occurring disorders can be helpful in clinical practice because other available clinical information (such as clinical anxiety symptoms or externalizing symptoms) often directly suggests an alternative diagnostic decisions, or a co-occurring disorder is already secured beforehand. Respective systematic ML approaches are scarce, to date (but see Duda, Ma, et al., 2016, for ASD vs. ADHD, and Wittkopf et al., 2021 for ASD vs. mood and anxiety disorders), likely due to a lack of curated databases with sufficient detail of co-occurring disorders in addition to ASD. The ASD-net (Kamp-Becker et al., 2017) database is unique in this respect. Here, we could demonstrate, for the first time, the feasibility of constructing optimized models of ASD classification for specific differential and co-occurring diagnoses. Importantly, the increase in classification performance was particularly strong for individuals that typically present a difficult decision for the clinician, that is, individuals with lower severity around the ADOS cutoff (up to 35% increase of specificity or sensitivity). The models with the anxiety group showed the largest increase in *specificity* for an ASD diagnosis, whereas the improvement for the models with the CD and ADHD samples was more related to increased *sensitivity*. This finding resonates well with other studies demonstrating lower specificity of the ADOS for mood and anxiety disorders (Wittkopf et al., 2021) and a high amount of symptom overlap (Hartley & Sikora, 2009; Tyson & Cruess, 2012; van Steensel et al., 2013). Thus, several individuals with anxiety-related disorders may be misclassified as having ASD when relying too much on the ADOS cutoff scores of module 3. Furthermore, the sensitivity of the ADOS appears to be particularly low for individuals with less severe ASD symptoms in the context of a potential differential diagnosis of CD or ADHD (Grzadzinski, Dick, Lord, & Bishop, 2016; Hartley & Sikora, 2009) or with ADHD as a co-occurring disorder. Our results demonstrate that optimized classification models can result in substantial improvements of diagnostic accuracy, which may directly translate into improved clinical care. This is particularly relevant for older children and adolescents who often have less severe ASD symptoms (Davidovitch et al., 2015), mixed with or masked by co-occurring disorders

(Frenette et al., 2013) rendering diagnostic decisions particularly challenging.

Specific item importance profiles

Our analyses reveal specific item importance profiles of the ADOS with respect to common differential diagnoses of ASD. In contrast to other approaches comparing single items for their potential to distinguish groups (e.g., (Grzadzinski et al., 2011, 2016) for ADHD), these item importance values reflect the total gain of information, which may help in distinguishing the diagnostic groups, including non-linear *interactions* between item ratings. Thus, for example, when anxiety is high, a diagnosis of ASD might not always be correct despite high scores on other symptom areas (e.g., eye contact). On the contrary, when hyperactivity is high, excessive talking and missing reciprocity during a conversation might be weighed differently than for less active individuals. In clinical practice, this is precisely what experienced clinicians do when evaluating ADOS results against other behavioral symptoms to arrive at a diagnostic decision (often irrespective of the overall ADOS score). Thus, for example, the increase in the importance value of the anxiety item for the ANX model (5th rank in item importance, in comparison with 23rd rank in the unspecific model) might reflect similar tweaks of the decision trees that compose the respective model and ultimately improve the classification.

ANX-model item importance profile. The ANX models had the largest deviations in item importance as compared to the unspecific models. Several items, many of these related to taking the initiative during direct social interaction were less important for the classification of ASD (see Figure 3). Other items, such as *Stereotyped/Ideosyncratic Use of Words or Phrases, Insight into Typical Social Situations and Relationships*, and *Anxiety*, seemed more important for the classification. Interestingly, the item *Stereotyped/Ideosyncratic Use of Words or Phrases* from the RBB domain was the item with the highest importance for the ANX vs. ASD model (see also (Wittkopf et al., 2021)). The RBB domain might thus be particularly important for the distinction between ASD and anxiety disorders.

ADHD model item importance profile. The item importance profile for the ADHD models revealed lower importance for the items *Stereotyped/Idiosyncratic Use of Words or Phrases, Insight into Typical Social Situations and Relationships, Speech Abnormalities, and Quality of Social Overtures*, but higher importance for *Conversation, Offers Information, Quality of Social Response, Overactivity, and Language Production and Linked Nonverbal Communication*. In contrast, Grzadzinski et al. (2016) report five items that discriminate best between ADHD and ASD (according to their definition, that is, that the

item is endorsed in >66% of the ASD group and <33% of the ADHD group): *Quality of Social Overtures, Amount of Reciprocal Social Communication, Unusual Eye Contact, Facial Expressions Directed to Examiner, and Stereotyped/Idiosyncratic Use of Words or Phrases*. Although we partially replicate this result, that is, four of these items were among the top 6 items for the ASD vs. ADHD model, our results suggest that this is not necessarily specific to ADHD. Similar items were among those with the highest importance for the unspecific model. In the specific ADHD model, the importance scores of these items were either in the same range or even lower, whereas, for example, *Conversation* had increased item importance in relation to the unspecific model and was the 3rd most important item for ADHD.

CD-model item importance profile. The profile for the CD models revealed slight decreases in item importance for *Insight into Typical Social Situations and Relationships* and *Conversation*, which is in line with the observation of reduced adherence and understanding of social norms and rules in CD. Furthermore, item importance was lower for *Speech Abnormalities Associated With Autism* (including prosody), *Facial Expression Directed to Examiner, Asks for information, and Amount of Reciprocal Social Communication*. This finding may be associated with observations of decreased emotional empathy and disturbed affective responsiveness in some individuals with CD, in particular those individuals with high callous-unemotional traits. (Klapwijk et al., 2016; von Polier et al., 2020). These characteristics may translate into lower differentiability for respective associated items in the ADOS. Slight increases in item importance could be observed for *Quality of Social Overtures, Stereotyped/ Idiosyncratic Use of Words or Phrases, Descriptive, Conventional, Instrumental, or Informational Gestures, and Quality of Social Response*, potentially suggesting stronger differences in behavioral symptoms related to these items for individuals with CD in comparison with ASD.

Limitations

The approach presented here is limited to the classification of ASD vs. non-ASD, it is not possible to confirm a co-occurring condition such as ADHD, anxiety, or CD using these models. Future studies could extend the approach by incorporating a standardized battery of specific diagnostic instruments for a range of diagnostic categories and test whether differential diagnostic decisions can be enhanced by using multi-label classification approaches. Furthermore, it could be beneficial to include ADI-R items into the models to further enhance diagnostic accuracy. The models currently only allow for the consideration of *one* specific co-occurring disorder (i.e., ANX, ADHD, or CD) to improve the ASD diagnosis. With higher sample sizes, it would also be possible to

include more fine-grained differential diagnostic groups (e.g., multiple co-occurring disorders, further characteristics and potentially confounding factors such as sex and age, and IQ) and reveal respective profiles of symptom clusters.

The samples for the ANX, CD, ADHD models differ with respect to % comorbidity (i.e., number of patients with both ASD and a co-occurring disorder) and only for the ADHD group sample sizes were high enough to train a specific model. Future studies should investigate to what extent this may influence accuracy of the trained models. Lastly, our models cannot predict the actual clinical utility of the trained models, for example, some diagnostic decisions might not actually be difficult for a clinician despite a discrepant ADOS score or an ADOS score around the cutoff. Further studies are needed to validate ML-based algorithms and potential improvements in clinician confidence for ASD diagnosis.

Future directions

To advance clinical care and diagnostic accuracy in the upcoming era of personalized medicine, it is essential not only to create large databases and optimized diagnostic algorithms using ML, but also to make these available for use in clinical practice. Our app (<https://msrlab.shinyapps.io/asd-ml-jcpp/>) demonstrates the potential of ML approaches to advance diagnostic decisions in today's clinical care for ASD. Importantly, it should not be viewed as a ready-to-use diagnostic tool, but as a demonstration of feasibility.

The ADOS represents a condensed clinical evaluation of specific dimensions across a broad range of observable behaviors. Its validity, however, depends on the expertise and experience of the observer (Kamp-Becker et al., 2018). Future developments should incorporate quantifiable indices of behavior, for example, eye gaze (Chong et al., 2019; Hartz, Guth, Jording, Vogeley, & Schulte-Rüther, 2021), facial expression (Drimalla et al., 2020), motion (Budman et al., 2019), or even neural assessment during social interaction (Kruppa et al., 2020). A holistic approach is necessary (Roessner et al., 2021), including intermediate steps: Future studies could aim to set up generative models that describe how quantifiable behavioral indices translate into clinician symptom ratings, and in a second step, relate these to the diagnostic classification. Similar to advances in neuromodelling (Frässle et al., 2018), this approach could use the powerful technique of generative embedding (Shawe-Taylor & Cristianini, 2004) to improve the diagnostic algorithm. Our approach is compatible with such considerations and may encourage further research into this direction.

To conclude, using ML in diagnostic procedures could be an excellent strategy for improving clinical decisions by utilizing the full range of information from comprehensive and detailed diagnostic observation.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Supplementary methods.

Appendix S2. Supplementary results.

Figure S1. Overview of individual item ratings for the different diagnostic groups (ASD: all participants with an ASD diagnosis; ANX, ADHD, CD, OTHER, NONE: participants without an ASD diagnosis, but a specific differential diagnosis (ANX, ADHD, CD), a different ICD-F diagnosis (OTHER), or no ICD-F diagnosis (NONE)).

Figure S2. ROC (receiver operating characteristic) curves for the evaluation of the random forest models.

Figure S3. Specificity of models. Y-axis depicts the mean AUC of the group-specific model evaluation (“specific”) and cross-evaluation across folds, error bars depict standard error of means.

Figure S4. Exemplary output of the web-app. The dashed grey line represents the individual prediction after entering a set of ADOS item scores.

Figure S5. Profile plots for ANX, CD, and ADHD models.

Figure S6. (a) Influence of age on model accuracy (in comparison to ADOS). (b) Influence of sex on model accuracy (in comparison to ADOS). (c) Influence of IQ on model accuracy (in comparison to ADOS).

Table S1. Items and item abbreviations of ADOS.

Table S2. ICD-F diagnoses within the OTHER group, separately for those individuals with and without an ASD diagnosis.

Table S3. (A) Comparison of model evaluation scores and respective scores for the ADOS algorithm for the full data. (B) Comparison of model evaluation scores

and respective scores for the ADOS algorithm for the data around the ADOS cut-off score.

Acknowledgements

The authors would like to thank all families who participated in the study. In addition, they want to thank Gerti Gerber for support in data management as well as Charlotte Küpper for their cooperation in the development of the data base. L.P. has received speaking fees from Takeda, medice and infectopharm and royalties from Hogrefe, Kohlhammer, and Schattauer. V.R. has received payment for consulting and writing activities from Lilly, Novartis, and Shire Pharmaceuticals; lecture honoraria from Lilly, Novartis, Shire Pharmaceuticals, and Medice Pharma; and support for research from Shire Pharmaceuticals and Novartis. He has carried out clinical trials in cooperation with the Novartis, Shire, Servier, and Otsuka companies. The remaining authors have declared that they have no competing or potential conflicts of interest.

Data Availability Statement

The data and the source code for the analyses are available upon reasonable request to the authors.

Correspondence

Martin Schulte-Rüther, Social Interaction and Developmental Neuroscience Lab, Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center Göttingen, von-Siebold-Str. 5, D-37075 Göttingen, Germany; Email: martin.schulte-ruether@med.uni-goettingen.de

Key points

- The diagnosis of autism spectrum disorder in children is particularly challenging in the context of co-occurring disorders with behavioral symptoms affecting social interaction. This is not taken into account by observation instruments such as ADOS-2.
- We trained specific machine learning models based on ADOS-2 items for the classification of ASD in the context of differential and co-occurring diagnoses.
- We found increased diagnostic accuracy, in particular for those patients with less severe ASD symptoms for whom the diagnostic decision is particularly difficult.
- Optimized diagnostic classifier may improve clinical decisions by utilizing the full range of information from comprehensive and detailed diagnostic observation and provide insights into those behavioral symptoms which are particularly relevant for differential diagnostic decisions.

References

- Abbas, H., Garberson, F., Glover, E., & Wall, D.P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, 25, 1000–1007.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®) (5th edn)*. Washington, DC: American Psychiatric Association.
- AWMF. (2016). *Autismus-Spektrum-Störungen im Kindes-, Jugend- und Erwachsenenalter Teil 1: Diagnostik. Interdisziplinäre S3-Leitlinie der DGKJP und der DGPPN sowie der beteiligten Fachgesellschaften, Berufsverbände und Patientenorganisationen*.
- Baxter, A.J., Brugha, T.S., Erskine, H.E., Scheurer, R.W., Vos, T., & Scott, J.G. (2015). The epidemiology and global burden of autism spectrum disorders. *Psychological Medicine*, 45, 601–613.

- Bone, D., Bishop, S.L., Black, M.P., Goodwin, M.S., Lord, C., & Narayanan, S.S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, *57*, 927–937.
- Bone, D., Goodwin, M.S., Black, M.P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2015). Applying Machine Learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders*, *45*, 1121–1136.
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
- Budman, I., Meiri, G., Ilan, M., Faroy, M., Langer, A., Reboh, D., ... & Dinstein, I. (2019). Quantifying the social symptoms of autism using motion capture. *Scientific Reports*, *9*, 7712.
- Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R. M., ... & Reh, J. M. (2019). Detecting gaze towards eyes in natural social interactions and its use in child assessment. *ArXiv:1902.00607 [Cs]*. <http://arxiv.org/abs/1902.00607>
- Davidovitch, M., Levit-Binnun, N., Golan, D., & Manning-Courtney, P. (2015). Late diagnosis of autism spectrum disorder after initial negative assessment by a Multidisciplinary Team. *Journal of Developmental & Behavioral Pediatrics*, *36*, 227–234.
- Drimalla, H., Scheffer, T., Landwehr, N., Baskow, I., Roepke, S., Behnia, B., & Dziobek, I. (2020). Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (SIT). *NPJ Digital Medicine*, *3*, 25.
- Duda, M., Daniels, J., & Wall, D.P. (2016). Clinical evaluation of a novel and mobile autism risk assessment. *Journal of Autism and Developmental Disorders*, *46*, 1953–1961.
- Duda, M., Kosmicki, J.A., & Wall, D.P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational Psychiatry*, *4*, e424.
- Duda, M., Ma, R., Haber, N., & Wall, D.P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Translational Psychiatry*, *6*, e732.
- Falkmer, T., Anderson, K., Falkmer, M., & Horlin, C. (2013). Diagnostic procedures in autism spectrum disorders: A systematic literature review. *European Child & Adolescent Psychiatry*, *22*, 329–340.
- Frässle, S., Yao, Y., Schöbi, D., Aponte, E.A., Heinzle, J., & Stephan, K.E. (2018). Generative models for clinical applications in computational psychiatry. *WIREs Cognitive Science*, *9*, e1460.
- Frenette, P., Dodds, L., MacPherson, K., Flowerdew, G., Hennen, B., & Bryson, S. (2013). Factors affecting the age at diagnosis of autism spectrum disorders in Nova Scotia, Canada. *Autism*, *17*, 184–195.
- Fusaro, V.A., Daniels, J., Duda, M., DeLuca, T.F., D'Angelo, O., Tamburello, J., ... & Wall, D.P. (2014). The potential of accelerating early detection of autism through content analysis of YouTube videos. *PLoS ONE*, *9*, e93533.
- Gilmour, J., Hill, B., Place, M., & Skuse, D.H. (2004). Social communication deficits in conduct disorder: A clinical and community survey. *Journal of Child Psychology and Psychiatry*, *45*, 967–978.
- Grzadzinski, R., Di Martino, A., Brady, E., Mairena, M.A., O'Neale, M., Petkova, E., ... & Castellanos, F.X. (2011). Examining autistic traits in children with ADHD: Does the autism Spectrum extend to ADHD? *Journal of Autism and Developmental Disorders*, *41*, 1178–1191.
- Grzadzinski, R., Dick, C., Lord, C., & Bishop, S. (2016). Parent-reported and clinician-observed autism spectrum disorder (ASD) symptoms in children with attention deficit/hyperactivity disorder (ADHD): Implications for practice under DSM-5. *Molecular Autism*, *7*, 7.
- Hartley, S.L., & Sikora, D.M. (2009). Which DSM-IV-TR criteria best differentiate high-functioning autism spectrum disorder from ADHD and anxiety disorders in older children? *Autism*, *13*, 485–509.
- Hartz, A., Guth, B., Jording, M., Vogeley, K., & Schulte-Rüther, M. (2021). Temporal behavioral parameters of on-going gaze encounters in a virtual environment. *Frontiers in Psychology*, *12*, 982.
- Hossain, M.M., Khan, N., Sultana, A., Ma, P., McKyer, E.L.J., Ahmed, H.U., & Purohit, N. (2020). Prevalence of comorbid psychiatric disorders among people with autism spectrum disorder: An umbrella review of systematic reviews and meta-analyses. *Psychiatry Research*, *287*, 922.
- Hyde, K.K., Novack, M.N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D.R., & Linstead, E. (2019). Applications of supervised Machine Learning in autism Spectrum disorder research: A review. *Review Journal of Autism and Developmental Disorders*, *6*, 128–146.
- Kamp-Becker, I., Albertowski, K., Becker, J., Ghahreman, M., Langmann, A., Mingeback, T., ... & Stroth, S. (2018). Diagnostic accuracy of the ADOS and ADOS-2 in clinical practice. *European Child & Adolescent Psychiatry*, *27*, 1193–1207.
- Kamp-Becker, I., Poustka, L., Bachmann, C., Ehrlich, S., Hoffmann, F., Kanske, P., ... & Wermter, A.-K. (2017). Study protocol of the ASD-Net, the German research consortium for the study of Autism Spectrum Disorder across the lifespan: From a better etiological understanding, through valid diagnosis, to more effective health care. *BMC Psychiatry*, *17*, 206.
- Klapwijk, E.T., Aghajani, M., Colins, O.F., Marijnissen, G.M., Popma, A., Lang, N.D.J., ... & Vermeiren, R.R.J.M. (2016). Different brain responses during empathy in autism spectrum disorders versus conduct disorder and callous-unemotional traits. *Journal of Child Psychology and Psychiatry*, *57*, 737–747.
- Kosmicki, J.A., Sochat, V., Duda, M., & Wall, D.P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry*, *5*, e514.
- Kruppa, J.A., Reindl, V., Gerloff, C., Oberwilling Weiss, E., Prinz, J., Herpertz-Dahlmann, B., ... & Schulte-Rüther, M. (2020). Interpersonal synchrony special issue brain and motor synchrony in children and adolescents with ASD—a fNIRS hyperscanning study. *Social Cognitive and Affective Neuroscience*, *16*, 103–116. <https://doi.org/10.1093/scan/nsaa092>
- Küpper, C., Stroth, S., Wolff, N., Hauck, F., Kliewer, N., Schadhansjosten, T., ... & Roepke, S. (2020). Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. *Scientific Reports*, *10*, 4805.
- Levy, S., Duda, M., Haber, N., & Wall, D.P. (2017). Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. *Molecular Autism*, *8*, 65.
- Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule—2nd edition (ADOS-2)* (p. 284). Los Angeles: Western Psychological Corporation.
- Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*, 659–685.
- Milledge, S.V., Cortese, S., Thompson, M., McEwan, F., Rolt, M., Meyer, B., ... & Eisenbarth, H. (2019). Peer relationships and prosocial behaviour differences across disruptive behaviours. *European Child & Adolescent Psychiatry*, *28*, 781–793.
- Poustka, L., Rühl, D., Feineis-matthews, S., Bölte, S., Poustka, F., & Hartung, M. (2015). *ADOS-2 Diagnostische Beobachtungsskala für Autistische Störungen—2*. Hogrefe AG: Verlag Hans Huber.

- Roessner, V., Rothe, J., Kohls, G., Schomerus, G., Ehrlich, S., & Beste, C. (2021). Taming the chaos?! Using eXplainable artificial intelligence (XAI) to tackle the complexity in mental health research. *European Child & Adolescent Psychiatry*, 30, 1143–1146.
- Ros, R., & Graziano, P.A. (2018). Social functioning in children with or at risk for attention deficit/hyperactivity disorder: A meta-analytic review. *Journal of Clinical Child & Adolescent Psychology*, 47, 213–235.
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge, UK: University Press.
- Tariq, Q., Daniels, J., Schwartz, J.N., Washington, P., Kalantarian, H., & Wall, D.P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLOS Medicine*, 15, e1002705.
- Thom, R.P., Keary, C.J., Kramer, G., Nowinski, L.A., & McDougle, C.J. (2020). Psychiatric assessment of social impairment across the lifespan. *Harvard Review of Psychiatry*, 28, 159–178.
- Tyson, K.E., & Cruess, D.G. (2012). Differentiating high-functioning autism and social phobia. *Journal of Autism and Developmental Disorders*, 42, 1477–1490.
- Uddin, S., Khan, A., Hossain, M.E., & Moni, M.A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19, 281.
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A.J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14, e0224365.
- van Steensel, F.J.A., Bögels, S.M., & Wood, J.J. (2013). Autism spectrum traits in children with anxiety disorders. *Journal of Autism and Developmental Disorders*, 43, 361–370.
- von Polier, G.G., Greimel, E., Konrad, K., Großheinrich, N., Kohls, G., Vloet, T.D., ... & Schulte-Rüther, M. (2020). Neural correlates of empathy in boys with early onset conduct disorder. *Frontiers in Psychiatry*, 11, 178.
- Wall, D.P., Kosmicki, J., Deluca, T.F., Harstad, E., & Fusaro, V.A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, 2, e100.
- Wittkopf, S., Stroth, S., Langmann, A., Wolff, N., Roessner, V., Roepke, S., ... & Kamp-Becker, I. (2021). Differentiation of autism spectrum disorder and mood or anxiety disorder. *Autism*, 136236132110396. <https://doi.org/10.1177/13623613211039673>
- Wright, M.N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77, 1–17.

Accepted for publication: 8 May 2022