scientific reports

OPEN

Feature replacement methods enable reliable home video analysis for machine learning detection of autism

Emilie Leblanc[1], Peter Washington[2], Maya Varma[3], Kaitlyn Dunlap[1,4], Yordan Penev[1,4], Aaron Kline[1,4] & Dennis P. Wall[1,4,5]✉

Autism Spectrum Disorder is a neuropsychiatric condition affecting 53 million children worldwide and for which early diagnosis is critical to the outcome of behavior therapies. Machine learning applied to features manually extracted from readily accessible videos (e.g., from smartphones) has the potential to scale this diagnostic process. However, nearly unavoidable variability in video quality can lead to missing features that degrade algorithm performance. To manage this uncertainty, we evaluated the impact of missing values and feature imputation methods on two previously published autism detection classifiers, trained on standard-of-care instrument scoresheets and tested on ratings of 140 children videos from YouTube. We compare the baseline method of listwise deletion to classic univariate and multivariate techniques. We also introduce a feature replacement method that, based on a score, selects a feature from an expanded dataset to fill-in the missing value. The replacement feature selected can be identical for all records (general) or automatically adjusted to the record considered (dynamic). Our results show that general and dynamic feature replacement methods achieve a higher performance than classic univariate and multivariate methods, supporting the hypothesis that algorithmic management can maintain the fidelity of video-based diagnostics in the face of missing values and variable video quality.

Autism Spectrum Disorder (ASD) is a complex neuropsychiatric condition affecting an estimated 53 million children under 5 years old worldwide and one million children in the US ten years of age or younger[1–4]. Autism's prevalence in the U.S. rose from 1 in 125 to 1 in 40 children within the last 5 years alone[5,6]. Although early ASD diagnosis is critical to the outcome of behavior therapies before sensitive periods of development have passed, families, particularly in rural and underserved communities, still face waiting times exceeding 12 months[59,60], repeated visits to professionals and out of state trips to establish a final diagnosis[61].

Trained on electronic medical records archived from standard-of-care diagnostic assessments, artificial intelligence algorithms have shown the ability to classify ASD and the potential to reduce waiting times for diagnosis, optimize caregivers' work load, and reach previously underserved populations through digital health tools. Using the Social Responsiveness Scale (SRS)[7], Duda et al. applied feature selection[8,9] to develop an ASD vs Attention Deficit Hyperactivity Disorder classification algorithm, achieving high performance (Area Under the Receiver Operating Characteristic curve (AUC-ROC) = 96.5%). Washington et al. built upon this work to develop an ASD vs neurotypical (NT) neural network classifier, achieving an AUC-ROC of over 92% for SRS-derived diagnoses[10]. Leveraging Autism Diagnostic Observation Schedule (ADOS)[11] scoresheet records, Küpper et al. developed ASD vs NT algorithms, yielding an AUC-ROC of 90% on patients under 21 years old[12]. Levy et al.[13] similarly found reduced sets of features highly predictive for ASD in ADOS Module 2 and Module 3 datasets. Autism Diagnostic Interview-Revised (ADI-R)[14] scoresheets have also been used for developing ASD classifiers. Duda et al.[15] used ADOS scoresheets to create an algorithm exhibiting >97% sensitivity and >77% specificity in comparison to ADOS-based diagnosis. Wall et al.[16] applied feature selection and found that 7 of the 93 items contained in the ADI-R were sufficient to classify autism, achieving 99.9% statistical accuracy with a 7-question

[1]Department of Pediatrics, Stanford University, Palo Alto, CA 94305, USA. [2]Department of Bioengineering, Stanford University, Palo Alto, CA 94305, USA. [3]Department of Computer Science, Stanford University, Palo Alto, CA 94305, USA. [4]Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305, USA. [5]Department of Psychiatry and Behavioral Sciences (by courtesy), Stanford University, Palo Alto, CA 94305, USA. ✉email: dpwall@stanford.edu

classifier using the ADI-R score sheets of 966 individuals. This feature selection work suggests that the number of features can be reduced from 30 questions per module for the standard instrument ADOS (4 modules, 116 questions total) and 93 questions for the standard instrument ADI-R, to 9 or 7 features depending on the model used. This potentially translates to decreased time to diagnosis as well as mutually beneficial opportunities to use alternative modes of feature capture, even by non-experts evaluating video from mobile devices. Finally, the use of these features in models opens opportunities to move the diagnosis from a binary outcome to a more continuous, quantitative measure that can be tracked over time.

Subsequent experiments focused on independent validation using novel independent datasets confirm that the models retain high performance, supporting their generalizability and limiting the potential of overfitting issues. Wall et al.'s 7-feature algorithm achieved 84.8% Unweighted Average Recall (UAR; the mean of the sensitivity and specificity) when tested on a prospective sample of 222 children[17]. This same model was also validated by Bone et al.[18] on a dataset with reasonable balance of autism and non-autism developmental delay cases (BID) of 680 child participants, producing a UAR of 80% in comparison to the clinical standard outcome. Bone et al. also tested an 8-feature alternating decision tree model from Wall et al.[31] in a different slice of their BID containing 1,033 children and found 94% UAR against the standard clinical diagnosis. Tariq et al.[19] validated several logistic regression models on an independent dataset of 1,089 individuals with autism and 66 individuals with no autism diagnosis, achieving 94.1% UAR with 9 features. These experiments support the hypothesis that small numbers of features can be used by machine learning models to produce an accurate classification of autism.

However, moving models to practice for more efficient and accessible autism diagnoses requires methods for rapid and reliable feature measurements. Guided by frameworks such as the one proposed by Stark et al.[20], such algorithms are now being integrated as quantitative classification tools into real world settings and embedded into mobile and telemedicine solutions. Leveraging YouTube videos, Fusaro et al. showed that features used in today's gold standard assesssment tools, such as the ADOS, can be measured with high accuracy from viewing short videos of children[21]. Abbas et al. trained two algorithms to identify autism from mobile inputs: one based on structured parent-reported questionnaires and the other on tagging key behaviors from short, semi-structured home videos of children[22,23]. Their results outperformed baseline clinician screeners[24]. Tariq et al.[19] evaluated the performance of 8 ADOS- and ADI-R-trained machine learning algorithms—including the osmicki et al.[25] and Wall et al.[16] algorithms that we leverage in this paper—on ratings by blinded non-experts of 5-min home videos of children with and without ASD, all achieving over 94.5% sensitivity. To render these "human-in-the-loop" algorithms scalable, Washington et al. leveraged crowdworkers[58] to rapidly obtain the necessary video ratings for minimal feature sets to run models with highest autism classification accuracy[26] and integrated privacy-protection measures to the process[27]. This work has shown the high potential for using mobile video and machine learning models for more scalable, affordable, and repeatable diagnosis; yet, important questions remain on how variability in video content impacts feature measurement.

The algorithms mentioned above rely on a fixed set of features, such as the ability to maintain eye contact or the presence of repetitive speech, and their prediction performance decreases when one or more of these features is not measurable [28]. During an observational assessment of a child at risk for autism, features can be missing for numerous reasons: children may be unable to express certain behaviors (e.g., because of their age), videos used to capture the child's interactions may only display a subset of the features needed (e.g., video length too short or quality too low) and some raters may not understand or may feel uncertain when answering specific questions. To translate machine learning algorithms into everyday healthcare, researchers must develop a robust missing feature strategy. Much like a clinician, an algorithm should adapt to the child's capacities and cannot experience a significant drop in performance if certain features are not available. For example, if "Repetitive Speech" is a feature of the algorithm and it cannot be evaluated on non-verbal children, alternative features with the same predictive power should be used instead and safeguards should be implemented to avoid misclassifications. Feature imputation and NULL-value treatment methods have been analyzed for healthcare- and psychiatry-related classifiers. Abidin et al. compared the performance of three machine learning classifiers (k-nearest neighbors, decision tree, and Bayesian networks) for data imputation accuracy[29]. Aisha et al. analyzed the effect of 9 different missing value treatments on the accuracy of four Bayesian network classifiers used to predict death in acute chronic Hepatitis patients[30]. However, most ASD-related papers are centered around overall algorithm performance, only briefly mentioning the type of feature imputation technique (for instance, Küpper et al. mention leveraging 5 nearest neighbors[12]).

In this paper, we use and evaluate two previously published ASD classification models: logistic regression with 9 features (LR9)[25] and alternating decision tree with 7 features (ADTree7)[16]. We evaluate various methods of treating missing values on the performance of these algorithms, training them on standard-of-care instrument scoresheets (ADOS Module 2 scoresheets for LR9 and ADI-R version 2003 scoresheets for ADTree7) and testing using non-expert ratings of 140 children YouTube videos. We compare (1) standard univariate and multivariate techniques for feature imputation with (2) general feature replacement strategies and (3) dynamic feature replacement methods which adapt to each specific record. Our work highlights the potential of imputation techniques for video-based ASD classification algorithms and the broader potential for use of feature replacement strategies in remote, mobile diagnostics.

Materials and methods

All methods described below were carried out in accordance with global, federal, state, and university guidelines and regulations for research and reviewed and approved by the Stanford University Institutional Review Board (IRB) prior to taking place.

Models. This work relies on a set of previous experiments, building towards the creation of a novel and scalable video-based assessment of ASD. The first set of experiments used electronic medical record data from standard-of-care measurements made during the clinical diagnosis of autism (or the absence of such diagnosis)[16,19,31]. This work focused on feature selection and dimensionality reduction to train 8 models and test their accuracy against the clinical decision for the child at risk for a developmental delay, including autism[19]. Kosmicki et al.[25], Levy et al.[13] and Washington et al.[10] illustrated how feature selection methods can reduce the number of standard-of-care instrument questions needed from 30 questions per module for ADOS (4 modules, 116 questions total) and 93 for ADI-R, to 9 or 7 behavioral features (depending on the model used), all while preserving high model performance. These experiments documented that the features needed for autism diagnosis can be significantly fewer than what is used in today's standard-of-care. The next experiment showed that the feature vectors needed for the models can be objectively measured in home videos of the child at risk for autism. To do so, our team has developed a secure video rating portal in which raters can view short home-video clips of children and submit their answers. In total, non-clinician raters score 30 behavioral features per video to ensure coverage of all 8 machine learning model features analyzed by Tariq et al.[19]. Finally, Tariq et al.[19], Duda et al.[9] and Washington et al.[26] have shown that prediction accuracy was preserved with scores collected through this system, when based on home-video clips with variable manifestations of autism and other developmental delays[8]. Tariq et al.[32] have also explored the adaptability of these models to distinct cultures and countries. Together this body of prior work (a.) found the optimal features and models, and (b.) demonstrated the ability to run the models on home videos quickly, suggesting that autism diagnosis may be possible through video scoring. The present takes the critical next step to address the robustness of the video diagnostic process to variability in home video quality and content.

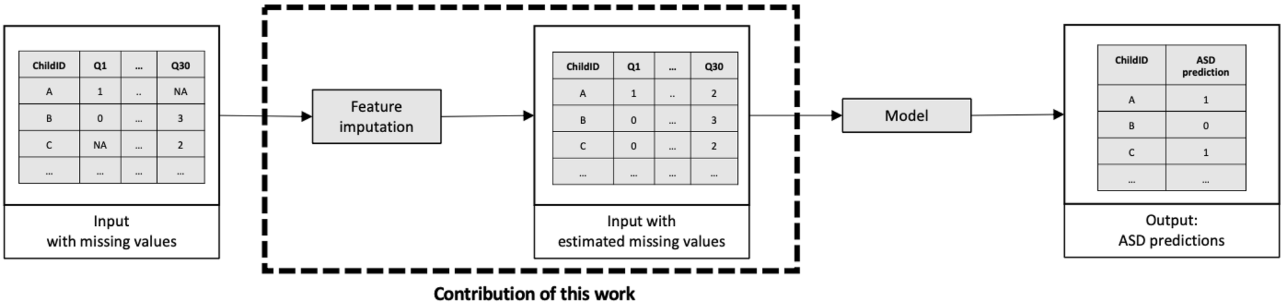
Variability in video length and quality remains a great challenge for the scalability of this ASD screening system, particularly as it increases the risk of missing values—raters being unable to assess some behavioral features. Our goal is to manage and limit missing values in these scores, all while decreasing rating time (i.e. the number of questions) and adapting to the specific content of each video.

This study focuses on and evaluates two published machine learning ASD diagnosis algorithms: logistic regression with 9 features (LR9)[25] and alternating decision tree with 7 features (ADTree7)[16]. LR9 features are "Expressive Language", "Eye Contact", "Joint Attention/Pointing", "Stereotyped Speech", "Spontaneous Gestures", "Indicates Pleasure to Others", "Social Overtures", "Complex Mannerisms" and "Stereotyped Interests/Actions". ADTree7 features are "Expressive Language", "Understands Language", "Eye Contact", "Developmental Delay", "Social Participation", "Pretend Play" and "Indicates Pleasure to Others" (see Supplementary Table S4 for summarizing table). Both models were validated by subsequent independent experiments achieving 98.9% sensitivity and 89.4% specificity for LR9, and at worst 89.9% sensitivity and 79.7% specificity for ADTree7[17,18]. LR9 was tested on independent data from 1,089 individuals with ASD and 66 individuals with no ASD diagnosis and ADTree7 was validated in a clinical trial of 222 participants and in a reasonably balanced independent dataset consisting of 680 individuals (462 with ASD). These models are also structurally quite different. LR9 and ADTree7 rely on two distinct families of machine learning models (Logistic Regression being a linear classifier and Decision Trees a non-linear classifier), they are trained on two different instruments (ADOS Module 2 and ADI-R 2003) and only share 3 common features. Because of these differences, the feature imputation and feature replacement methods' performances are susceptible to vary widely between LR9 and ADTree7. Comparing their performances on these two models thus offers a better assessment of the methods' quality.

Datasets. *Training dataset.* Following approval by the Stanford University IRB, our training dataset[33] was assembled and analyzed. This dataset groups de-identified ADOS and ADI-R electronic health records previously collected by multiple sources: Autism Genetic Resource Exchange[34], Autism Consortium, National Database for Autism Research[35], Simons Simplex Collection[36], Simons Variation in Individuals Project[37]. Under an IRB approved data transfer agreement between Stanford University and Cognoa Inc., we also included a previously collected dataset of de-identified ADI-R responses from Cognoa Inc. As the datasets described above were secondary data sources, informed consent was waived by our IRB.

ADI-R consists of 93 items: 2 free-response items, 14 age of onset items, and 77 ordinal scale items whose responses range from 0 (typical behavior) to 4 (severely atypical). ADOS is administered as four different modules, with each module being appropriate for a different age range and child ability. Responses range from 0 (typical behavior) to 3 (severely atypical). As defined in the initial development of these models[19], we only use ADOS Module 2 for LR9 training and ADI-R 2003 for ADTree7 training. This groups a total of 16,200 instrument ratings, of which 4,343 are ADOS Module 2 score sheets (Supplementary Tables S1a and S2a) and 11,857 ADI-R 2003 score sheets (Supplementary Tables S1b and S2b). The balance of males to females in the dataset matches the increased prevalence of ASD in males compared to females[38].

Testing dataset. With the same methods for video data collection and feature tagging as described in Tariq et al.[19], we collected 140 publicly available YouTube videos of children, 70 ASD and 70 NT balanced for age and gender (Supplementary Tables S3a and S3b). As we collected publicly available data, collection of informed consent was waived by Stanford University IRB. Videos were selected from YouTube using YouTube metatags to confirm the age and diagnosis of the child in the video. If a video did not include a metatag for the age of the child in the video, the age was assigned following full agreement among the estimates made by 3 clinical practitioners in pediatrics. Videos were selected based on whether the video (1) was between 1 and 5 minutes in length, (2) showed the face and hands of the child, (3) showed clear opportunities for or direct social engagement, and (4) involved opportunities for the use of an object such as a utensil, crayon, or toy.



ChildID	Q1	...	Q30
A	1	..	NA
B	0	...	3
C	NA	...	2
...	...	...	...

ChildID	Q1	...	Q30
A	1	..	2
B	0	...	3
C	0	...	2
...	...	...	...

ChildID	ASD prediction
A	1
B	0
C	1
...	...

Figure 1. Pipeline description and contribution.

As in Tariq et al.[19], ratings were performed by either students (high school, undergraduate, or graduate-level) or working professionals with no formal training or certification for detection or diagnosis of autism. The rating questionnaire consisted of 30 behavioral features (e.g., eye contact, social smile), used in previously published machine learning models and shown to be highly predictive of ASD[19] (see Supplementary File Table S4 for the detailed list of features). All raters received minimal training with a trained clinical coordinator prior to performing feature ratings and were blind to the diagnosis of the child in the video. The testing dataset used here included 3 distinct ratings chosen at random from this pool of untrained raters for each of the 140 videos (i.e. 3 distinct 30-feature vectors per video). No optimization was conducted based on the raters' previous performance nor based on rater types. We use this dataset of 3 ratings of the 30 features for each of the 140 5-min YouTube videos as our test set.

Missing values core concepts.

Little and Rubin[39] introduce 3 categories of missing data: (1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) non-ignorable. In MCAR, the probability of missingness is the same for all records. If a variable is MCAR, then ignoring records with missing data should not introduce bias. Ignoring records with missing data (i.e., listwise deletion) is our baseline method for addressing missing values. For MAR cases, the probability of missingness varies for each record but remains predictable from other variables in the database. For example, if we assume males are less likely to fill-in the mental health part of a medical survey, this will induce higher missing values but would be unrelated to the status of their mental health. Therefore, we can fill-in missing values appropriately (i.e. predict mental health responses) by considering the variables affecting the probability of missingness (i.e. gender since it affects the probability of mental health responses' missingness). However, it is difficult to ensure MAR as there may be unobserved variables also influencing the process. We consider our features to be MAR when we predict missing values based on the other features of the algorithm. Finally, in the non-ignorable case, data missingness is non-random, depending on information that has not been recorded in the database. By definition, this makes it extremely difficult to identify and predict. To address this, we expand the dataset we use to predict the missing values to other variables, even if they were not included in the original model's features, through general and dynamic feature replacement techniques. Although this dataset expansion helps reduce risk, it does not entirely rule out the existence of another latent variable. To these three missing data classifications, we can add a fourth: (4) missingness that depends on the missing value itself. For instance, in a survey, the probability of having a missing value in a salary question will most likely depend on the salary amount itself. This introduces a high risk for bias. As our datasets are composed of ratings of videos done by individuals with no particular link to the child in the video, we do not expect them to engage in such self-censoring behaviors.

Methodology.

We describe a pipeline of employing a feature imputation method and feeding the resulting features into a diagnostic classification model (Fig. 1). Our study focuses on finding the best feature imputation method for ASD classification (see Supplementary File "Additional Information—Mathematical Formulation").

Evaluation. We compare LR9 and ADTree7 model performances with each feature imputation method. For every model and every feature imputation method, all items of the pipeline are trained, each time adapting the feature imputation method and tuning model hyperparameters using a 10-fold GridSearch cross validation optimizing for UAR (Fig. 2). The trained pipeline is then tested on the YouTube dataset (Fig. 3), with ratings for each video aggregated with mode (i.e., most frequent value) using the *scikit-learn*[40] library in *Python 3*. This process is repeated 5 times to account for the variability in the generated folds and to be able to report on average and standard deviation performance.

Feature imputation methods compared.

Baseline: listwise deletion. Listwise deletion simply omits records containing missing values. Although it is often the default method in analyses, listwise deletion involves losing information and may introduce a bias if missing values are not missing completely at random (MCAR). It is considered as the baseline algorithm in this paper and we compare the performance of each feature replacement method to this baseline. In our case, a record will be dropped from the training set if at least one of the model's features is NULL and we will not be able to attribute a prediction to the test record if all of the 3 raters have answered NULL in at least one same feature.

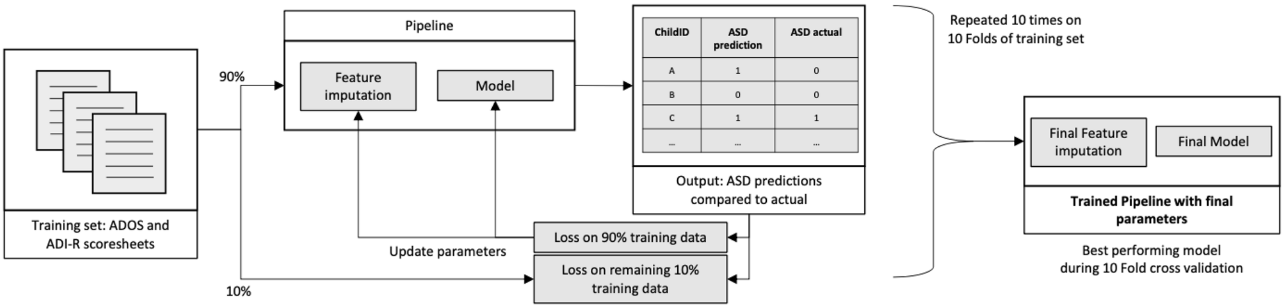

Training set: ADOS and ADI-R scoresheets
90%
10%
Pipeline
Feature imputation
Model
Update parameters
Loss on 90% training data
Loss on remaining 10% training data
ChildID
ASD prediction
ASD actual
A
B
C
Output: ASD predictions compared to actual
Repeated 10 times on 10 Folds of training set
Final Feature imputation
Final Model
Trained Pipeline with final parameters
Best performing model during 10 Fold cross validation

Figure 2. Pipeline training process on instrument scoresheets.

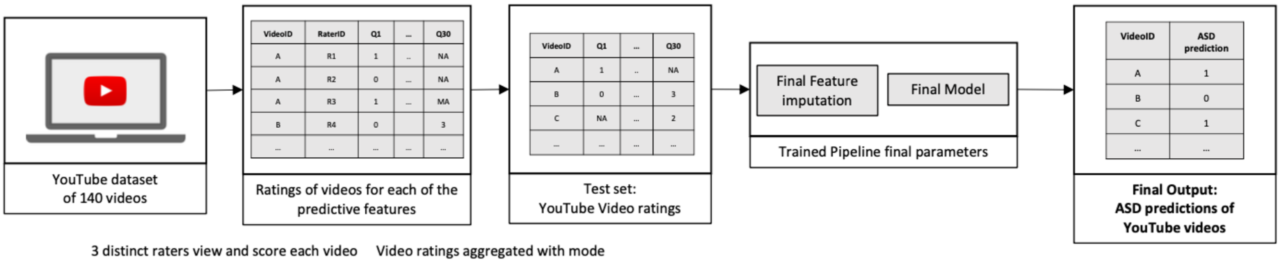

YouTube dataset of 140 videos
VideoID
RaterID
Q1
Q30
R1
R2
R3
R4
NA
MA
Ratings of videos for each of the predictive features
Test set: YouTube Video ratings
Final Feature imputation
Final Model
Trained Pipeline final parameters
ASD prediction
Final Output: ASD predictions of YouTube videos
3 distinct raters view and score each video
Video ratings aggregated with mode

Figure 3. Pipeline testing process on YouTube video ratings.

Classic feature imputation techniques: univariate. Univariate feature imputation methods replace missing values in a feature with a constant obtained by aggregating non-missing values of that same feature. We compare 3 different statistics: **mean**, **median** and **most frequent value** (i.e. mode).

Classic feature imputation techniques: multivariate. Multivariate feature imputation techniques rely on all the model features to fill missing values. In our case, this corresponds to all 9 features for LR9 and all 7 features for ADTree7. In an iterative round-robin fashion, we predict the missing values in feature j^* based on all other j features such that $j \neq j^*$ and $j \in [1, n]$ and n the number of features in the model. We compare 3 commonly used techniques: **Gaussian mixture with expectation-maximization**[41], **ridge regression**[42], and **decision trees**[43] (see Supplementary File "Additional Information—Classic feature imputation techniques: Multivariate").

General feature replacement methods. We expand the feature space by considering all features available in ADOS Module 2 or ADI-R 2003, not limiting ourselves to the 9 or 7 features of the previously published models. Instead of creating a model that would rely on a combination of these features, we simply attempt to replace the missing value with another "close" feature's value as described in Fig. 4. Although this enhances the feature space mathematically, in practice, this method allows us to replace (and not add) a question a rater may be unable to answer with the next best question. This maintains the initial number of questions asked to raters and simultaneously adapts the questionnaire to the video content if a behavioral feature cannot be assessed.

Therefore, we map all n (= 9 or 7) features of the model to their closest feature in the larger set of n^* features available according to a score s (e.g., mutual information (MI)) (see Supplementary File "Additional Information—Mathematical Formulation—General feature replacement methods"). We compare three different scores to select the replacement feature: **correlation**-based, **nearest neighbor** (NN), and **mutual information**-based. Correlation-based selects the feature with the highest Spearman correlation with the feature we wish to fill. Because the input features (i.e. questions answered by raters) are ranked on an ordinal scale, we chose Spearman correlation as it does not assume a linear distribution and measures how well the relationship between two variables can be described as a monotonic function[44]. Similarly, the nearest neighbor method selects the closest nearest neighbor feature as its replacement using Euclidean distance. Finally, MI-based feature replacement selects the feature having the highest MI with the missing feature.

Dynamic feature replacement. In this final method, we realize that the best replacement feature (selected via correlation, NN or MI) may vary depending on the child being rated and the contents of the video. To take this into consideration, for each new test record, we dynamically apply the feature replacement methods described above on a subset of the training set corresponding to records having similar ratings as the test set (described in Fig. 5; see also Supplementary File "Additional Information—Mathematical Formulation—Dynamic feature replacement").

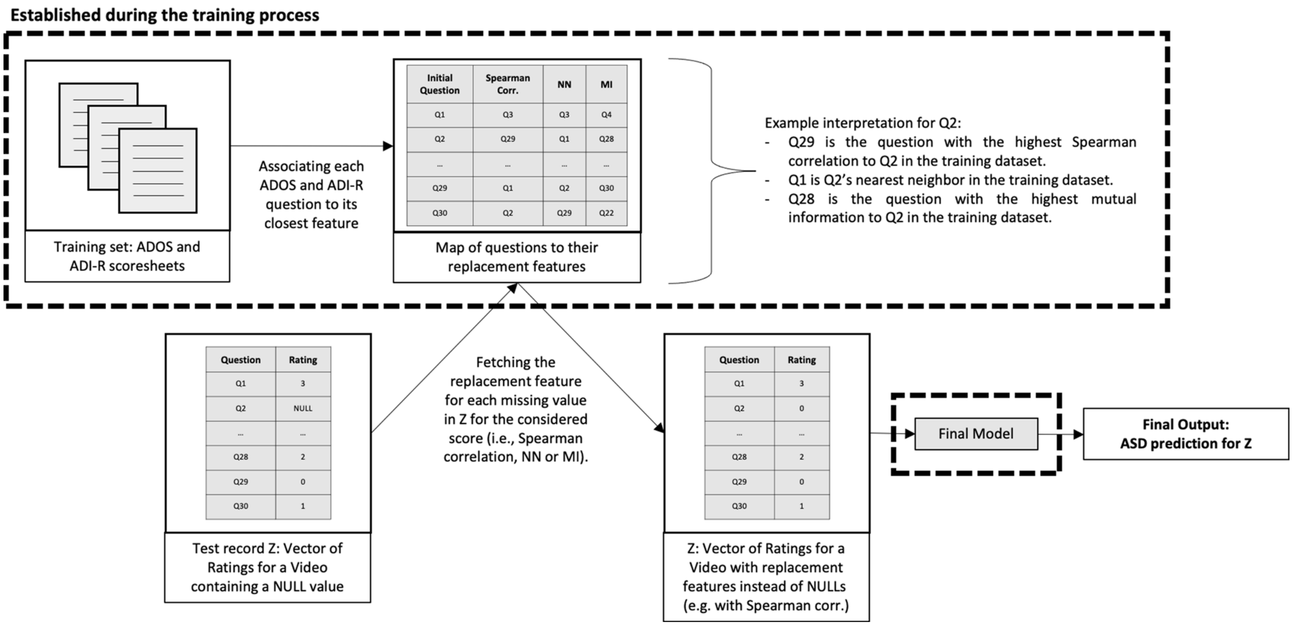

Established during the training process
Training set: ADOS and ADI-R scoresheets
Associating each ADOS and ADI-R question to its closest feature
Initial Question | Spearman Corr. | NN | MI
Q1 | Q3 | Q3 | Q4
Q2 | Q29 | Q1 | Q28
Q29 | Q1 | Q2 | Q30
Q30 | Q2 | Q29 | Q22
Map of questions to their replacement features
Example interpretation for Q2:
- Q29 is the question with the highest Spearman correlation to Q2 in the training dataset.
- Q1 is Q2's nearest neighbor in the training dataset.
- Q28 is the question with the highest mutual information to Q2 in the training dataset.
Question | Rating
Q1 | 3
Q2 | NULL
Q28 | 2
Q29 | 0
Q30 | 1
Test record Z: Vector of Ratings for a Video containing a NULL value
Fetching the replacement feature for each missing value in Z for the considered score (i.e., Spearman correlation, NN or MI).
Q2 | 0
Z: Vector of Ratings for a Video with replacement features instead of NULLs (e.g. with Spearman corr.)
Final Model
Final Output: ASD prediction for Z

Figure 4. General feature replacement process illustration.

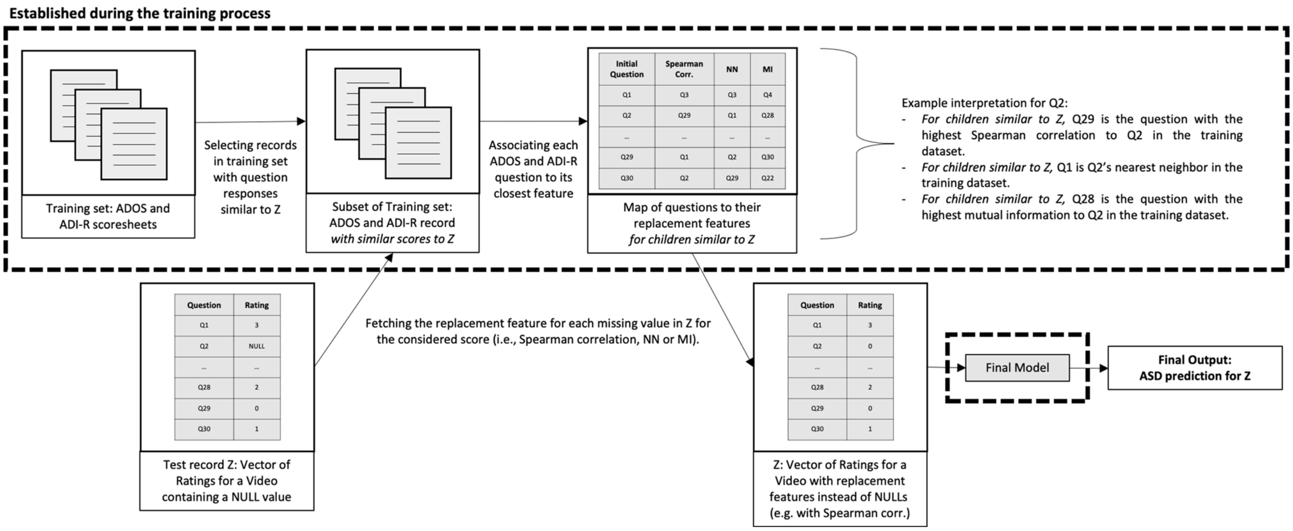

Established during the training process
Training set: ADOS and ADI-R scoresheets
Selecting records in training set with question responses similar to Z
Subset of Training set: ADOS and ADI-R record with similar scores to Z
Associating each ADOS and ADI-R question to its closest feature
Initial Question | Spearman Corr. | NN | MI
Map of questions to their replacement features for children similar to Z
Example interpretation for Q2:
- For children similar to Z, Q29 is the question with the highest Spearman correlation to Q2 in the training dataset.
- For children similar to Z, Q1 is Q2's nearest neighbor in the training dataset.
- For children similar to Z, Q28 is the question with the highest mutual information to Q2 in the training dataset.
Test record Z: Vector of Ratings for a Video containing a NULL value
Fetching the replacement feature for each missing value in Z for the considered score (i.e., Spearman correlation, NN or MI).
Z: Vector of Ratings for a Video with replacement features instead of NULLs (e.g. with Spearman corr.)
Final Model
Final Output: ASD prediction for Z

Figure 5. Dynamic feature replacement process illustration.

Results

Dataset analysis.

We notice a drastic difference in the percentage of missing values in each of the model's training sets: the LR9 model features in the ADOS Module 2 dataset contain on average 2.3% of missing values (Supplementary Figure S1a) while that average increases to 17.8% for ADTree7 (Supplementary Figure S1b). This difference is reduced in testing as the testing dataset is identical for both models. However, the average percentage of missing values per feature for LR9 is still slightly lower at 5.1% (Supplementary Figure S2a) compared to 8.9% for ADTree7 (Supplementary Figure S2b), as the models have different input features. Missing values tend to be concentrated in one or two features: for LR9, "Joint Attention Pointing" and "Spontaneous Gestures" are missing for 13.8% and 14.8% of ratings respectively (Supplementary Figure S2a) and, for ADTree7, "Pretend Play" is missing for over 40.7% of our ratings (Supplementary Figure S2b). We then analyzed the origin of the missing values in our YouTube testing dataset with Welch t-tests and Bonferroni correction. We compared the overall average number of missing values and the number of missing values per feature, between ASD and NT children, between age groups, and between genders. Although no significant difference was noted between the populations in the overall number of missing values, some individual features had an associated p value < 0.05. The following features have significantly more missing values in the ASD population than in the NT population at level of significance $\alpha = 0.05$: "Echolalia" (p value = 0.0235, not significant after Bonferroni correction), "Speech Patterns" (p value = 6.8623e−08), "Communicative Engagement" (p value = 1.4198e−05), "Understands

Model	Sensitivity	Specificity	UAR	AUC-ROC	AUC PR	Ratings dropped
LR9	0.8939 (0.0152)	0.7536 (0.0102)	0.8238 (0.0118)	0.9109 (0.0021)	0.9195 (0.0018)	135 ratings and 5 videos entirely
ADTree7	0.8172 (0.0397)	0.8721 (0.0709)	0.8447 (0.0261)	0.9083 (0.0117)	0.8706 (0.0644)	205 ratings and 21 videos entirely

Table 1. Baseline performance of LR9 and ADTree7 on 420 ratings of 140 YouTube videos (average performance and standard deviation). With listwise deletion, a rating is dropped if it contains at least one NULL value in the model's features (this is the case for 135 ratings for LR9). We are unable to rate a video if at least one model feature is missing in all 3 ratings of this video (this is the case for 5 videos for LR9).

(a) Performance of LR9 (average performance and standard deviation).					
Method	**Sensitivity**	**Specificity**	**UAR**	**AUC-ROC**	**AUC PR**
Mean	0.9029 (0.0064)	0.8571^{*B} (0.0226)	0.8800^{*B} (0.0128)	0.9541^{*B} (0.0012)	0.9658^{*B} (0.0013)
Median	0.9143^{*} (0.0000)	0.8514^{*B} (0.0192)	0.8829^{*B} (0.0096)	0.9577^{*B} (0.0016)	0.9695^{*B} (0.0010)
Mode	0.9171^{*} (0.0064)	0.8400^{*B} (0.0212)	0.8786^{*B} (0.0113)	0.9569^{*B} (0.0006)	0.9671^{*B} (0.0002)
(b) Performance of ADTree7 (average performance and standard deviation).					
Method	**Sensitivity**	**Specificity**	**UAR**	**AUC-ROC**	**AUC PR**
Mean	0.6857 (0.1446)	0.9800^{*} (0.0373)	0.8329 (0.0760)	0.9105 (0.0306)	0.8799 (0.1396)
Median	0.6057 (0.3307)	0.9714^{*} (0.0143)	0.7886 (0.1694)	0.9150 (0.0142)	0.7810 (0.1247)
Mode	0.7000 (0.1325)	0.9457 (0.0445)	0.8229 (0.0859)	0.8823 (0.0591)	0.7646 (0.1548)

Table 2. Univariate feature imputation methods—performance of LR9 and ADTree7 on 420 ratings of 140 YouTube videos.

Language" (p value = 2.3683e−05) and "Stereotyped Speech" (p value = 0.0008). No feature was significant at α = 0.05 when comparing female and male participants and, when comparing age groups, only "Imitates Actions" had significantly more missing values in the 1 to 3 year old age group (p value = 0.0455), although this does not stand after Bonferroni multiple hypothesis correction. This analysis shows that missing values tend to occur more often for ASD children than NT children in specific features ("Speech Patterns", "Communicative Engagement", "Understands Language", "Calls Attention to Objects", "Stereotyped Speech", "Stereotyped Interests and Actions") and that the diagnosis of the child is an unobservable latent variable influencing their missingness. Throughout this paper and as a proxy, we impute missing values with the help of features that are themselves predictors of ASD vs NT diagnosis.

Baseline listwise deletion. In the YouTube dataset, 135 ratings have at least one missing value in the LR9 features and 205 ratings have at least one missing value in the ADTree7 features. As we apply listwise deletion as a baseline, these ratings were dropped, thus reducing the number of ratings per video to 1 or 2 (since we started with 3 ratings per video). 5 videos for LR9 and 21 videos for ADTree7 could not be rated at all as at least one feature was missing in all 3 of the ratings of these videos. The model achieves an average UAR of 82.38% (0.0117 standard deviation) for LR9 and 84.47% (0.0265 standard deviation) as seen in Table 1. However, ADTree7 has a high variability in results, generating higher standard deviations, particularly increasing the Area Under the Precision-Recall curve (AUC PR). We use a Welch t-test for all results to evaluate the significance of the performance increase or decrease compared to baseline. A * symbol indicates significance at level $\alpha = 0.05$ and a ^{*B} symbol indicates the significance at level $\alpha = 0.05$ stands after Bonferroni correction.

Classic feature imputation techniques. *Univariate.* All univariate feature imputation methods yield significant improvements for LR9 for all metrics except sensitivity (Table 2a). Sensitivity only significantly improves for median and mode methods, although this does not stand after Bonferroni correction. Median achieves the best UAR of the univariate feature imputation methods with 88.29% (0.0096). However, univariate feature imputation methods do not perform as well for ADTree7, which contains more NULL values than LR9. The only significant improvements are achieved by mean and median in specificity and do not stand after Bonferroni correction (Table 2b). The top UAR of 83.29% (0.0760) is achieved with mean feature imputation but does not match the baseline model's performance.

Multivariate. Multivariate imputation methods tend to perform as well as univariate imputation methods for LR9, except for feature imputation with Gaussian mixtures. Although ridge regression and decision trees both achieve more than 88.7% UAR, neither generate a jump in sensitivity significant enough for the Welch test (Table 3a). For ADTree7, decision trees seem to perform better than other multivariate methods, but it does not pass the significance test (Table 3b).

(a) Performance of LR9 (average performance and standard deviation).					
Method	**Sensitivity**	**Specificity**	**UAR**	**AUC-ROC**	**AUC PR**
Gaussian mixture	0.9029 (0.0186)	0.8457 (0.1085)	0.8743 (0.0467)	0.9477^{*B} (0.0074)	0.9604^{*B} (0.0072)
Ridge regression	0.9114 (0.0064)	0.8657^{*B} (0.0128)	0.8886^{*B} (0.0064)	0.9549^{*B} (0.0012)	0.9660^{*B} (0.0012)
Decision trees	0.9057 (0.0128)	0.8686^{*B} (0.0445)	0.8871^{*B} (0.0185)	0.9576^{*B} (0.0023)	0.9680^{*B} (0.0015)
(b) Performance of ADTree7 (average performance and standard deviation).					
Method	**Sensitivity**	**Specificity**	**UAR**	**AUC-ROC**	**AUC PR**
Gaussian mixture	0.7429 (0.1313)	0.8514 (0.128)	0.7971 (0.0597)	0.8067 (0.0869)	0.7659 (0.1483)
Ridge regression	0.8229 (0.0594)	0.8886 (0.0356)	0.8557 (0.0321)	0.8937 (0.0508)	0.8724 (0.0782)
Decision trees	0.7571 (0.0562)	0.9829* (0.0235)	0.8700 (0.0222)	0.9289 (0.0353)	0.9214 (0.0806)

Table 3. Multivariate feature imputation methods—performance of LR9 and ADTree7 on 420 ratings of 140 YouTube videos.

(a) Performance of LR9 (average performance and standard deviation).					
Method	**Sensitivity**	**Specificity**	**UAR**	**AUC-ROC**	**AUC PR**
Most correlated feature	0.9114 (0.0064)	0.8629^{*B} (0.0239)	0.8871^{*B} (0.0117)	0.9597^{*B} (0.0012)	0.9708^{*B} (0.0008)
Nearest-neighbor feature	0.9171* (0.0120)	0.8286^{*B} (0.0286)	0.8729^{*B} (0.0185)	0.9582^{*B} (0.0028)	0.9696^{*B} (0.0024)
Highest mutual information feature	0.9086 (0.0078)	0.8657^{*B} (0.0192)	0.8871^{*B} (0.0128)	0.9599^{*B} (0.0010)	0.9706^{*B} (0.0008)
(b) Performance of ADTree7 (average performance and standard deviation).					
Method	**Sensitivity**	**Specificity**	**UAR**	**AUC-ROC**	**AUC PR**
Most correlated feature	0.7771 (0.0217)	0.9886* (0.012)	0.8829* (0.0081)	0.9445* (0.0252)	0.9406 (0.0484)
Nearest-neighbor feature	0.8257 (0.031)	0.9743* (0.0293)	0.9000^{*B} (0.0231)	0.9295 (0.0323)	0.9073 (0.0813)
Highest mutual information feature	0.7343* (0.0603)	0.9629* (0.0217)	0.8486 (0.0234)	0.9316^{*B} (0.0054)	0.8150 (0.1454)

Table 4. General feature replacement methods—performance of LR9 and ADTree7 on 420 ratings of 140 YouTube videos.

General feature replacement techniques.

General feature replacement methods use features that were rated but not included in the initial models as substitutes to model features if marked NULL. This methodology unlocks significant improvements that were not achieved with the classic feature imputation methods. For LR9, when using the nearest neighbor feature, we achieve a significant improvement for all of the metrics considered and only sensitivity does not pass the Bonferroni correction (Table 4a). For ADTree7, general feature replacement methods are the first to achieve significant performance with Bonferroni correction: the highest mutual information method yields 93.16% AUC-ROC and the nearest neighbor method reaches the 90.00% balanced accuracy mark (Table 4b). These methods also help analyze which questions are easier to rate. As seen in Supplementary Tables S5 and S9, "Quality of Social Overtures" is often replaced with "Amount of social overtures / maintenance of attention", hinting that rating quantity may be easier than quality. We also note the presence of simpler concepts such as "Pointing" and "Showing", which are preferred to concepts like "Complex mannerisms", "Joint attention" and "Descriptive gestures", which could help make the questions more accessible to non-expert raters.

Dynamic feature replacements.

When applying our new dynamic feature replacement scheme on the YouTube dataset, we notice a significant improvement for all metrics but sensitivity for LR9 when compared to the base model, achieving a maximum of 89.57% UAR with dynamic mutual information (Table 5a). For ADTree7, we notice a significant increase of UAR compared to baseline, although passing the Bonferroni correction only with mutual information (Table 5b). When comparing general and dynamic feature replacement methods, we notice equivalent performances for LR9, overall equivalent UAR for ADTree7 and an increase in sensitivity for ADTree7. This new method therefore appears to maintain the high performances of the general feature replacement method on top of allowing for an automatic selection of the replacement feature. The increase in ADTree7 sensitivity may also indicate that including information on the individual record when choosing the best feature replacement reduces the imbalance in false positive vs false negatives.

Discussion

In anticipation of the widespread use of machine learning classifiers as detection tools for autism[45], here we studied the impact of missing values on the performance of two previously published ASD classifiers, a logistic regression using 9 features (LR9) and an alternating decision tree model using 7 features (ADTree7), using a dataset of non-expert ratings of 140 YouTube child videos. We compared common univariate and multivariate feature imputation methods to general and dynamic feature replacement techniques. For LR9, general feature

(a) Performance of LR9 (average performance and standard deviation).

Method	Sensitivity	Specificity	UAR	AUC-ROC	AUC PR
Dynamic—most correlated feature	0.9171* (0.0839)	0.8286* (0.0156)	0.8729^{*B} (0.0070)	0.9585^{*B} (0.0018)	0.9690^{*B} (0.0015)
Dynamic—nearest-neighbor feature	0.9171* (0.0057)	0.8543^{*B} (0.0305)	0.8857^{*B} (0.0143)	0.9596^{*B} (0.0020)	0.9697^{*B} (0.0016)
Dynamic—highest mutual information feature	0.9114 (0.0057)	0.8800^{*B} (0.0280)	0.8957^{*B} (0.0132)	0.9613^{*B} (0.0033)	0.9711^{*B} (0.0023)

(b) Performance of ADTree7 (average performance and standard deviation).

Method	Sensitivity	Specificity	UAR	AUC-ROC	AUC PR
Dynamic—most correlated feature	0.8343 (0.0333)	0.9371 (0.0400)	0.8857* (0.0175)	0.9200 (0.0168)	0.9168 (0.0448)
Dynamic—nearest-neighbor feature	0.8429 (0.0239)	0.9343 (0.0400)	0.8886* (0.0154)	0.9164 (0.0117)	0.9325 (0.0079)
Dynamic—highest mutual information feature	0.8714* (0.0286)	0.9171 (0.0262)	0.8943^{*B} (0.0177)	0.9120 (0.0356)	0.8871 (0.0658)

Table 5. Dynamic feature replacement methods—performance of LR9 and ADTree7 on 420 ratings of 140 YouTube videos.

replacement methods achieve a similar performance as classic univariate and multivariate methods (general feature replacement methods achieve at best 88.71% UAR vs 88.86% for multivariate methods). However, when confronted with even more missing values, as is the case with ADTree7, general feature replacement methods achieve a higher average UAR than classic and multivariate approaches (general feature replacement methods achieve 90.00% UAR vs 87.00% for multivariate feature imputation). General feature replacement methods also help elucidate which questions are easier to rate and may point to ways to improve their formulation. Dynamic feature replacement methods allow a jump in average UAR for LR9 (achieving 89.57% when the replacement feature is dynamically selected via mutual information) and an improvement in sensitivity for ADTree7 (87.14% with replacement feature dynamically selected via mutual information). Overall, we see that using algorithmic-driven replacement questions in place of missing values and dynamically personalizing feature imputation methods to the YouTube video considered allows for an increase in UAR for both LR9 and ADTree7.

One main concern is the risk of overfitting in the general and dynamic feature replacement methods we introduced. Dynamic feature replacement in particular relies on a large training set containing enough diversity in children profiles to provide an accurate replacement feature to each test record considered. There is a risk for a drop in performance for rare rating vectors and for types of ratings significantly different from our training set (for instance if we consider raters or children from outside of the United States).

To limit the risk of overfitting, we used machine learning models constrained to have few features and assigned a lower bound to the regularization parameters during hyperparameter tuning while minimizing the model error. We also used 10-fold cross-validation, which enables sequential evaluation of the UAR with different partitions of the dataset. In addition, we were careful to use a gender and age balanced dataset (41.4% females, 58.6% males; 53.6% 1–3 y.o., 42.9% 4–6 y.o. children; Supplementary File Table 3) to avoid overfitting to one type of demographic. An important additional point is that prior experiments have shown that the LR9 and ADTree7 classifiers generalize to large independent datasets, achieving a sensitivity of 98.9% and 89.9% resp. and a specificity of 89.4% and 79.7% resp.[17–19], therefore further minimizing their potential of overfitting.

Despite these safeguards, the overfitting risk introduced by the feature replacement methods remains. Methods such as shuffling labels, adding noise or randomly removing part of the training data can be helpful in measuring the impact of overfit. Other methods, such as bagging, help limit its impact. In particular, we welcome future work that tests these methods on independent datasets, as it is the best way to measure their tendency to overfit.

Future work is also needed to validate these findings using actual at-home videos instead of YouTube videos. Digital mobile[46–50] and wearable[51–57] therapeutics are increasingly collecting highly structured videos of children with ASD and matched neurotypical controls, such as unaffected siblings. This process is building the necessary video database for the validation of our methods. More work is also needed to validate our findings when using crowdsourced ratings. Indeed, realistic and representative videos of children in naturalistic settings can be fed into a crowdsourced pipeline leveraging the methods discussed in the present study to extract diagnostic features of children with autism, enabling for the remote, scalable, and speedy screening of ASD. Another important next step will be to compare the performance of feature replacement methods (both general and dynamic) to the rater's intuition and "best guess" by prohibiting NULL values in ratings.

Received: 18 August 2020; Accepted: 2 November 2020
Published online: 04 December 2020

References

1. Baio, J. *et al.* Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveill. Summ.* **67**, 1 (2018).
2. Baxter, A. J. *et al.* The epidemiology and global burden of autism spectrum disorders. *Psychol. Med.* **45**, 601 (2015).
3. Olusanya, B. O. *et al.* Developmental disabilities among children younger than 5 years in 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet Glob Health* **6**, e1100–e1121 (2018).

4. Boyle, C. A. *et al.* Trends in the prevalence of developmental disabilities in us children, 1997–2008. *Pediatrics* **127**, 1034–1042 (2011).
5. Kogan, M. D. *et al.* The prevalence of parent-reported autism spectrum disorder among us children. *Pediatrics* **142**, e20174161 (2018).
6. Xu, G. *et al.* Prevalence and treatment patterns of autism spectrum disorder in the united states, 2016. *JAMA Pediatr.* **173**, 153–159 (2019).
7. Constantino, J. N. Social responsiveness scale. In *Encyclopedia of Autism Spectrum Disorders* 2919–2929 (2013). https://doi.org/10.1007/978-1-4419-1698-3_296.
8. Duda, M., Ma, R., Haber, N. & Wall, D. Use of machine learning for behavioral distinction of autism and ADHD. *Transl. Psychiatry* **6**, e732–e732 (2016).
9. Duda, M., Haber, N., Daniels, J. & Wall, D. Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Transl. Psychiatry* **7**, e1133–e1133 (2017).
10. Washington, P. *et al.* Feature selection and dimension reduction of social autism data. In; *Pac Symp Biocomput*, vol. 25, 707–718 (World Scientific, 2020).
11. Lord, C. *et al.* Austism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J. Autism Dev. Disord.* **19**, 185–212 (1989).
12. Küpper, C. *et al.* identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. *Sci. Rep.* **10**, 1–11 (2020).
13. Levy, S., Duda, M., Haber, N. & Wall, D. P. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. *Mol. Autism* **8**, 65 (2017).
14. Lord, C., Rutter, M. & Le Couteur, A. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **24**, 659–685 (1994).
15. Duda, M., Kosmicki, J. & Wall, D. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Transl. Psychiatry* **4**, e424–e424 (2014).
16. Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y. & DeLuca, T. F. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE* **7**, e43855 (2012).
17. Duda, M., Daniels, J. & Wall, D. P. Clinical evaluation of a novel and mobile autism risk assessment. *J. Autism Dev. Disord.* **46**, 1953–1961 (2016).
18. Bone, D. *et al.* Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *J. Autism Dev. Disord.* **45**, 1121–1136 (2015).
19. Tariq, Q. *et al.* Mobile detection of autism through machine learning on home video: a development and prospective validation study. *PLoS Med.* **15**, e1002705 (2018).
20. Stark, D. E., Kumar, R. B., Longhurst, C. A. & Wall, D. P. The quantified brain: a framework for mobile device-based assessment of behavior and neurological function. *Appl. Clin. Inform.* **7**, 290 (2016).
21. Fusaro, V. A. *et al.* The potential of accelerating early detection of autism through content analysis of YouTube videos. *PLoS ONE* **9**, e93533 (2014).
22. Abbas, H., Garberson, F., Glover, E. & Wall, D. P. Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening. In: *2017 IEEE International Conference on Big Data (Big Data)* 3558–3561 (IEEE, 2017).
23. Abbas, H., Garberson, F., Glover, E. & Wall, D. P. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* **25**, 1000–1007 (2018).
24. Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E. & Wall, D. P. Multi-modular AI approach to streamline autism diagnosis in young children. *Sci. Rep.* **10**, 1–8 (2020).
25. Kosmicki, J., Sochat, V., Duda, M. & Wall, D. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Transl. Psychiatry* **5**, e514–e514 (2015).
26. Washington, P. *et al.* Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. *J. Med. Internet Res.* **21**, e13668 (2019).
27. Washington, P. *et al.* Crowdsourced feature tagging for scalable autism diagnoses. *Under Review* (2021).
28. Acuna, E. & Rodriguez, C. The treatment of missing values and its effect on classifier accuracy. In: *Classification, Clustering, and Data Mining Applications* 639–647 (2004).
29. Abidin, N. Z., Ismail, A. R. & Emran, N. A. Performance analysis of machine learning algorithms for missing value imputation. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **9**, 442–447 (2018).
30. Aisha, N., Adam, M. B. & Shohaimi, S. Effect of missing value methods on bayesian network classification of hepatitis data. *Int. J. Comput. Sci. Telecommun.* **4**, 8–12 (2013).
31. Wall, D. P., Kosmicki, J., Deluca, T., Harstad, E. & Fusaro, V. A. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl. Psychiatry* **2**, e100–e100 (2012).
32. Tariq, Q. *et al.* Detecting developmental delay and autism through machine learning models using home videos of bangladeshi children: Development and validation study. *J. Med. Internet Res.* **21**, e13822 (2019).
33. Paskov, K. M. & Wall, D. P. A low rank model for phenotype imputation in autism spectrum disorder. *AMIA Summ. Transl. Sci. Proc.* **2018**, 178 (2018).
34. Geschwind, D. H. *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.* **69**, 463–466 (2001).
35. Hall, D., Huerta, M. F., McAuliffe, M. J. & Farber, G. K. Sharing heterogeneous data: the national database for autism research. *Neuroinformatics* **10**, 331–339 (2012).
36. Fischbach, G. D. & Lord, C. The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
37. Individuals Project (Simons VIP Consortium), S. V. A genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. *Neuron* **73**, 1063–1067 (2012).
38. Loomes, R., Hull, L. & Mandy, W. P. L. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *J. Am. Acad. Child Adolesc. Psychiatry* **56**, 466–474 (2017).
39. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data* Vol. 793 (Wiley, Hoboken, 2019).
40. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**, 1–22 (1977).
42. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
43. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
44. Myers, J. L., Well, A. & Lorch, R. F. *Research Design and Statistical Analysis* (Routledge, Abingdon, 2010).
45. Washington, P. *et al.* Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 759–769 (2019).
46. Kalantarian, H. *et al.* Labeling images with facial emotion and the potential for pediatric healthcare. *Artif. Intell. Med.* **98**, 77–86 (2019).

47. Kalantarian, H. *et al.* The performance of emotion classifiers for children with parent-reported autism: quantitative feasibility study. *JMIR Mental Health* **7**, e13174 (2020).
48. Kalantarian, H. *et al.* Guess what?. *J. Healthc. Inform. Res.* **3**, 43–66 (2019).
49. Kalantarian, H., Jedoui, K., Washington, P. & Wall, D. P. A mobile game for automatic emotion-labeling of images. *IEEE Trans. Games* **12**, 213–218 (2018).
50. Kalantarian, H. *et al.* A gamified mobile system for crowdsourcing video for autism research. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)* 350–352 (2018).
51. Washington, P. *et al.* A wearable social interaction aid for children with autism. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* 2348–2354 (2016).
52. Voss, C. *et al.* Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatrics* **173**, 446–454 (2019).
53. Washington, P. *et al.* Superpowerglass: a wearable aid for the at-home therapy of children with autism. *Proc. ACM Interact. Mob. Wear. Ubiquitous Technol.* **1**, 1–22 (2017).
54. Daniels, J. *et al.* Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. *NPJ Dig. Med.* **1**, 1–10 (2018).
55. Voss, C., Haber, N. & Wall, D. P. The potential for machine learning-based wearables to improve socialization in teenagers and adults with autism spectrum disorder-reply. *JAMA Pediatrics* **173**, 1106–1106 (2019).
56. Daniels, J. *et al.* Feasibility testing of a wearable behavioral aid for social learning in children with autism. *Appl. Clin. Inform.* **9**, 129 (2018).
57. Nag, A. *et al.* Toward continuous social phenotyping: analyzing gaze patterns in an emotion recognition task for children with autism through wearable smart glasses. *J. Med. Internet Res.* **22**, e13810 (2020).
58. Washington, P. et al. Precision Telemedicine through Crowdsourced Machine Learning: Testing Variability of Crowd Workers for Video-Based Autism Feature Recognition. *Journal of Personalized Medicine* **10**, 86 (2020).
59. Gordon-Lipkin, E., Foster, J., & Peacock, G. Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. *Pediatr Clin North Am.* **63**, 851–859. (2016). https://doi.org/10.1016/j.pcl.2016.06.007
60. Wiggins, LD., Baio, J., & Rice, C. Examination of the time between first evaluation and first autism spectrum diagnosis in a population-based sample. *J Dev Behav Pediatr.* **27**, S79–S87. (2006)
61. Chiri, G. & Warfield, M. E. Unmet Need and Problems Accessing Core Health Care Services for Children with Autism Spectrum Disorder. *Maternal and Child Health Journal* **16**, 1081–1091 (2012)

Acknowledgements

We thank all raters that viewed and scored the 140 YouTube videos and enabled this study. The work was supported in part by funds to DPW from the National Institutes of Health (1R01EB025025-01, 1R21HD091500-01, 1R01LM013083), the National Science Foundation (Award 2014232), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, Lucile Packard Foundation, the Weston Havens Foundation, and program grants from Stanford's Human Centered Artificial Intelligence Program, Precision Health and Integrated Diagnostics Center (PHIND), Beckman Center, Bio-X Center, Predictives and Diagnostics Accelerator (SPADA) Spectrum, Spark Program in Translational Research, and from the Wu Tsai Neurosciences Institute's Neuroscience:Translate Program. We also acknowledge generous support from David Orr, Imma Calvo, Bobby Dekesyer and Peter Sullivan. P.W. would like to acknowledge support from the Stanford Interdisciplinary Graduate Fellowship (SIGF).

Author contributions

Conceptualization, E.L., P.W. and D.P.W.; data curation, E.L., P.W., K.D., Y.P, A.K. and D.P.W.; formal analysis, E.L., P.W. and D.P.W.; funding acquisition, P.W., K.D., Y.P., A.K. and D.P.W.; investigation, D.P.W.; methodology, E.L., P.W., M.V. and D.P.W.; resources, K.D., Y.P., and D.P.W.; software, E.L., P.W. and D.P.W.; supervision, D.P.W.; validation, P.W., M.V., K.D., Y.P., A.K. and D.P.W.; visualization, E.L.; writing-original draft, E.L. and D.P.W.; writing-review and editing, P.W., M.V., K.D., Y.P., A.K. and D.P.W.. All authors have read and agree to the published version of the manuscript.

Competing interests

D.P.W. is the founder of Cognoa.com. This company is developing digital health solutions for pediatric care. A.K. works as part-time consultant to Cognoa.com. All other authors declare no conflict of interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-76874-w.

Correspondence and requests for materials should be addressed to D.P.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.