Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Intelligence-Based Medicine 6 (2022) 100057

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Machine learning models using mobile game play accurately classify children with autism

Nicholas Deveau [a,*], Peter Washington [b], Emilie Leblanc [c], Arman Husic [c], Kaitlyn Dunlap [c], Yordan Penev [c], Aaron Kline [c], Onur Cezmi Mutlu [d], Dennis P. Wall [a,c]

[a] Biomedical Data Science, Stanford University, Stanford, 94305, California, United States
[b] Bioengineering, Stanford University, Stanford, 94305, California, United States
[c] Pediatrics, Stanford University, Stanford, 94305, California, United States
[d] Electrical Engineering, Stanford University, Stanford, 94305, California, United States

A B S T R A C T

Digitally-delivered healthcare is well suited to address current inequities in the delivery of care due to barriers of access to healthcare facilities. As the COVID-19 pandemic phases out, we have a unique opportunity to capitalize on the current familiarity with telemedicine approaches and continue to advocate for mainstream adoption of remote care delivery. In this paper, we specifically focus on the ability of GuessWhat? a smartphone-based charades-style gamified therapeutic intervention for autism spectrum disorder (ASD) to generate a signal that distinguishes children with ASD from neurotypical (NT) children. We demonstrate the feasibility of using "in-the-wild", naturalistic gameplay data to distinguish between ASD and NT by children by training a random forest classifier to discern the two classes (AU-ROC = 0.745, recall = 0.769). This performance demonstrates the potential for GuessWhat? to facilitate screening for ASD in historically difficult-to-reach communities. To further examine this potential, future work should expand the size of the training sample and interrogate differences in predictive ability by demographic.

1. Introduction

Remote treatment and progress tracking has transformed the way in which clinicians deliver care to their patients [1]. This trend was catalyzed by COVID-19, and as the pandemic phases out, remote health is positioned to remain a primary component of many forms of care [2,3]. One of the best established forms of remote treatment, telemedicine, has been shown to increase access to care across geographic regions [4]. Telemedicine's success during the COVID-19 pandemic provided a glimpse of a future in which access to care is not determined by one's ability to physically visit a care provider. This specific moment in history presents the biomedical community with an opportunity to drastically improve access to care for individuals who have historically been underserved by the medical system. If we are to realize a future of more equitable healthcare delivery, it is critical that we focus on developing new forms of remote care at a moment in time when both patients and clinicians are familiar with remote care workflows.

Autism Spectrum Disorder (ASD) serves as a clear example of a condition that is an ideal substrate for remote care. Although the prevalence of ASD is similar in both rural (0.9 pct.) and urban (1.0 pct.) areas, individuals in rural communities face limited access to identification and intervention services [5]. Data mining studies have suggested that up to 80 pct. of counties in the U.S. lack sufficient diagnostic resources [6]. Moreover, early diagnosis and intervention of ASD can significantly improve the quality of life for individuals with ASD and their families [7]. As such, the invention of novel technologies that allow for remote screening for ASD could address the disparity in early diagnosis of the disorder.

The development of technology for remote care also allows us to experiment with novel ways of conceptualizing the patient-care interaction. It is not a stretch to say that simply providing the current physician interaction through a smartphone may not be the best way of providing remote care. In fact, delays in the adoption of telehealth have been attributed to "unengaged" and "resisting" users [8].

Moreover, current methods for screening for ASD typically include subjective caregiver-report questionnaires. Feature selection on electronic health records have identified salient behavioral features for predicting ASD [9–12], and these features can be reliably acquired through crowdsourcing by non-expert raters [13–17,17]. These non-expert feature tags have been used to train machine learning models which can detect ASD with high accuracy, precision, and recall [10,11,18–23]. While digitization of these questionnaires may be one way of

* Corresponding author.
E-mail address: nick.deveau.94@gmail.com (N. Deveau).

https://doi.org/10.1016/j.ibmed.2022.100057
Received 6 October 2021; Received in revised form 10 January 2022; Accepted 29 March 2022
Available online 24 August 2022

addressing gaps in screening for ASD, such questoinaries require literacy and perform worse with non-white and lower education caregivers [24]. Consequently, naively deploying digitized versions of diagnostic questionnaires risks exacerbating current disparities in the early identification of ASD. A clear need exists for objective methods of screening for ASD that perform equally well across demographic groups.

Our lab developed GuessWhat? a mobile charades-style gamified therapeutic intervention that acquires structured video data from children with ASD for use in behavioral diagnostics research [25–29]. We designed the gamified therapeutic to be an engaging and fun way for parents and children to interact while having the option to support behavioral research and remote therapy by sharing objective gameplay and video data. Computer vision classifiers have been developed with the resulting data streams [29,30], and other computer vision efforts for detecting behavioral features related to early time point diagnostics and longitudinal outcome tracking are possible [13,31,32].

In addition to active data collection and monitoring of structured gameplay sessions, passive data collection and device usage measures can potentially be used for diagnostic purposes. Detecting behavioral and mental health conditions through passive device usage has been termed "digital phenotyping" in the literature [29,30,33–35]. Here, we explore the feasibility of using device usage data during gameplay to predict the presence of ASD in a semi-passive manner. Although this falls outside of the traditional definition of "digital phenotyping", we argue that this passive and semi-passive prediction of behavioral health from device usage also falls into the broad category of "digital phenotyping".

The main goal of this study is to identify the ability for GuessWhat? To generate a signal that distinguishes children with ASD from neurotypical (NT) children. Using only objective behavioral data captured by the game, we successfully demonstrated the ability to train a classifier that distinguishes the two groups, a critical step toward formalizing the game as an objective and easily-deployed remote screening tool for ASD.

2. Methods

2.1. Data collection

We collected behavioral data through at-home gameplay of the GuessWhat? game, a game developed to acquire structured video from children with ASD for behavioral disease research [25–29]. During gameplay, a parent shows a child a prompt–an image–and the child is asked to act out the image, much as they would in a game of charades. As the child acts out the prompt, the parent guesses what the image is based off of the child's acting, and if the parent guesses correctly, the parent

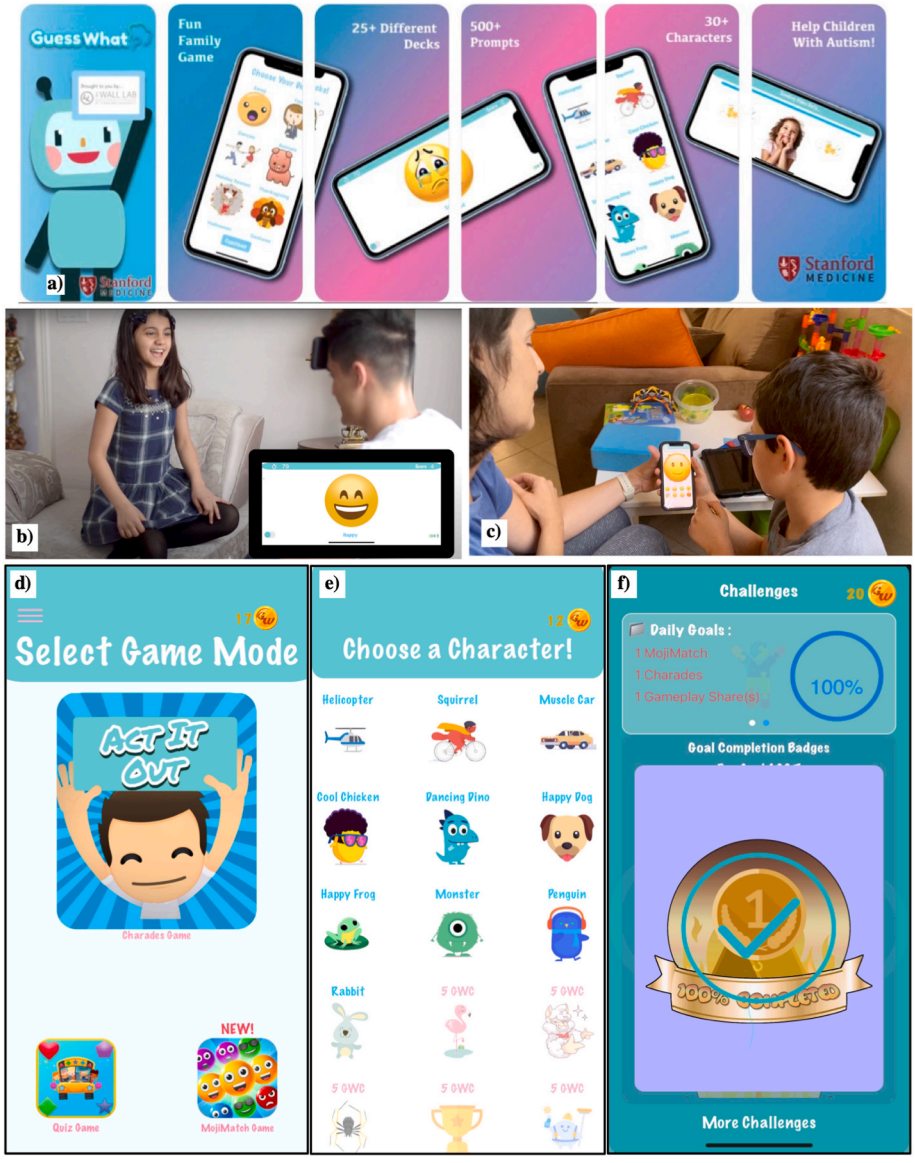

Fig. 1. Mobile Intervention User Experience. a) GuessWhat is a charades-style mobile game available for any a smartphone device. In a typical game session, b) the parent holds the smartphone to their forehead and tries to guess the emotion mimicked by the child in response to the prompt shown on the phone's screen. Upon guessing, the parent tilts the phone to proceed to the next prompt through the end of the 90-second session. c) After each 90s game, parent and child can review together. In-app d) game modes, e) unlocking deck and character choices based on coins earned, and f) activity-based achievement badges reinforce positive progression and ensure optimal child engagement through time.

labels the prompt by tilting the phone forward (top of phone away from forehead) if successful and backward (bottom of phone away from forehead) if unsuccessful. These steps are detailed in Fig. 1.

During gameplay, the app collects metadata regarding whether the parent successfully guessed the prompt acted out by the child. We define a trial to be the delivery of a single prompt to a child during a unique session of gameplay (need figure/graphic showing this).

The unprocessed data stored by the game and used in this study includes event-level data that tracks the following aspects of gameplay:

- Start time of a prompt
- End time of a prompt
- Whether the prompt was correct
- Whether the prompt was skipped

2.2. Participant recruitment

The Stanford Institutional Review Board approved the study prior to any research activities taking place. The recruitment methods for this study are identical to Ref. [36] This study was conducted remotely. Participants were recruited through GuessWhat's existing userbase, The Hartwell Foundation's KidsFirst autism research database, Research Match.org, and Facebook advertisements.

All participating were required to meet the following criteria: 1) they were able to read and speak English, 2) they had a compatible iOS or Android device with internet access, 3) the parent was 18+ years old, 4) the child was between 3 and 12 years of age and diagnosed with ASD.

To safeguard against the potential for self-reporting bias, we required the caregiver to confirm that their child's autism diagnosis came from a formal medical assessment. We asked the caregiver to choose a diagnostic label from a menu of choices including Autism Spectrum Disorder, Autistic Disorder, Pervasive Developmental Disorder-Not Otherwise Specified, Asperger Syndrome, ADHD/ADD, Anxiety, Speech and Language Delay. In addition, we required participants to report on the specific type(s) of therapy being administered to their child. This information requires a specialized understanding of the autism diagnosis and subsequent treatment prescriptions.

2.3. Data sample, feature engineering and preprocessing

We collected gameplay from children with autism spectrum disorder (ASD, n = 28) and neurotypical children (NT, n = 21). 19 of the ASD children were male and 9 were female. 10 of the NT children were male, 7 were female, and 4 did not provide sex information. Data were acquired between April 2017 and February 2021. Children were classified as ASD or non-ASD based on parent-provided information collected when the parent signed up for GuessWhat? Children had a mean age of 7.10 +- 5.82 years.

In order to minimize missing-data imputation, we focused our analysis only on the most commonly presented prompts, which were images from the CAFE dataset, a collection of images of young children displaying angry, fearful, sad, happy, surprised, disgusted and neutral faces [37]. Future work will expand this image dataset to include prompts derived from videos served on the social media platform TikTok.

2.4. Feature engineering

Previous work has shown that differences in emotion recognition tasks can be used to distinguish children with autism from NT children [38,39]. This delta stems from the ability to correctly identify an emotion and the reaction time required to do so (e.g., how long it takes a child to recognize a facial emotion). Consequently, we developed features that allowed us to measure these two constructs.

The first set of features (N = 17, detailed in appendix A1) measured the accuracy with which a child successfully acted out a specific prompt. Each one of these features corresponded to one of 17 faces shown during gameplay. A prompt was considered correct if the parent labeled it as such during gameplay. It should be noted that, although parents received instructions for how to correctly label a prompt, we had no way to confirm that they did in fact label it correctly. Thus, for m trials of each face type, we calculated percent correct p to be:

$$p = 100 * \frac{1}{m} \sum_{i}^{m} \mathbb{1} \quad \text{where } \mathbb{1} = \begin{cases} 1 & \text{if correct} \\ 0 & \text{if incorrect} \end{cases} \tag{1}$$

For a single session of gameplay, we calculated a percent correct feature for each of the prompts shown from the CAFE dataset.

The second set of features (n = 17) measured the amount of time it took for a child to act out the prompt and for the parent to label it. . We call this prompt duration, d, and we calculated it as follows:

$$d = \frac{1}{m} \sum_{i}^{m} e_i - s_i \quad \text{where } \begin{cases} e_i = \text{prompt end time} \\ s_i = \text{prompt start time} \end{cases} \tag{2}$$

where m is the number of times a prompt corresponding to a specific emotion was shown to a child. In other words, it was the average amount of time it took a child to identify a specific emotion. We calculated d for each of the 17 types of faces shown, regardless of whether a child correctly identified the face. Appendix A2 illustrates the input data schema.

2.5. Preprocessing

To avoid incorporating information from the distribution of the training data into the test set (i.e., "data leakage"), we carried out the following preprocessing steps separately for each test-train split of the data. The steps included, in order: outlier removal, imputation, standardization, upsampling. For all features, we considered values greater than 3 standard deviations away from the mean value to be an outlier. We removed these values and then used k-nearest-neighbor imputation (k = 3) to fill missing observations. We then standardized our data by subtracting feature-wise means from each observation and dividing by feature-wise standard deviation. Finally, due to compounding effects of moderately imbalanced classes within a small size dataset, we used SMOTE [40] to upsample the minority class and ensure balanced classes (equal ASD and NT) in each test-train split.

2.6. Modeling

We tested the performance of 4 classifiers on our set of 34 features. We trained and tested our models in Python and using the packages scikit-learn [41]. We chose to test models from 3 families of classifiers: linear models, support vector machines (SVM) and tree-based methods. Three main criteria drove the choice of these families of models. First, we had no *a priori* belief about the linearity (or lack thereof) of the relationships between our features, so it was important to model our data using a set of methods that would perform well under various conditions of linearity. Second, our sample size was not particularly large, so it was important to test model types that offered considerable flexibility to prevent overfitting through regularization. Third, we wanted to choose an interpretable model to gain insight into the specific aspects of gameplay that predict ASD. Table 1 describes the types of models we used in our analysis and their relevant attributes.

We used a repeated, nested grid search to simultaneously identify the best performing set of hyperparameters for our models as well as to understand the statistical accuracy of the performance metrics we obtained. The first iteration of the outer loop of our cross-validation procedure randomly splits the dataset into 4 equal-sized partitions. Then, using only 3 of these 4 partitions, the inner loop tunes the model hyperparameters using a grid search and 4-fold cross validation in order to select the best performing model with respect to AU-ROC. It should be noted that, because of a low sample size, we limited our search space to

Table 1
Summary of tested classifiers. Hypermarameter names correspond to those used by scikit-learn v. 1.0.1

	Hyperparameter	Values Tested
XGBoost	learning_rate	0.05, 0.10, 0.15, 0.20, 0.25, 0.30
	max_depth	1, 2, 3
	min_child_weight	1, 3, 5, 7
	gamma	0.1, 0.2, 0.3, 0.4
	colsample_bytree	0.3, 0.4, 0.5, 0.7
Random Forest	max_depth	1, 2, 3
	min_samples_leaf	2, 3, 4, 5
Logistic Regeession	penalty	L1, L2
	C	0.1 to 10, 20 values log-spaced
Linear SVM	C	−7 to 4, 50 values log-spaced

include hyperparameters that would lead to more regularization and less overfitting. After the best performing model was selected by the inner loop, its performance (measured by AU-ROC and recall) was tested on the unseen data included in the 4th partition from the outer loop. After four iterations of the outer loop, we obtained 4 classifiers, each with a corresponding set of performance metrics. Finally, we repeated the outer loop 7 times to obtain a total of 28 sets of performance metrics from which we bootstrapped distributions of each metric of interest. Fig. 6 provides a visualization of the repeated, nested cross validation procedure.

2.7. Feature selection

A simpler model (e.g. a model with fewer features) is often favorable as it is easier to interpret and often can improve performance by eliminating noisy features. We produced heatmaps in order to inspect the importance of each feature across the 28 iterations of cross validation used to train and evaluate each model (e.g., Fig. 2). Finally, we retained the models using only duration-based features, which consistently displayed higher feature importance in the random forest classifier. We noted a clear separation between duration and accuracy-based features: features using the time taken by the child to act out the prompt were significantly more important on average (Fig. 4; $t = 5.15$, $p = 1e\text{-}5$) than features relying on the parent's ability to correctly guess the prompt the child is acting out. This gap in feature importance based on the type of feature (duration or accuracy-based) could be due to the variability in parents' adherence to the GuessWhat? Instructions.

2.8. Final training

After manual feature selection, we followed the same repeated, nested cross-validation procedure as before and re-trained the classifiers using the entire training set on the reduced feature space.

2.9. Model evaluation

We evaluated models by comparing the mean values of AU-ROC and

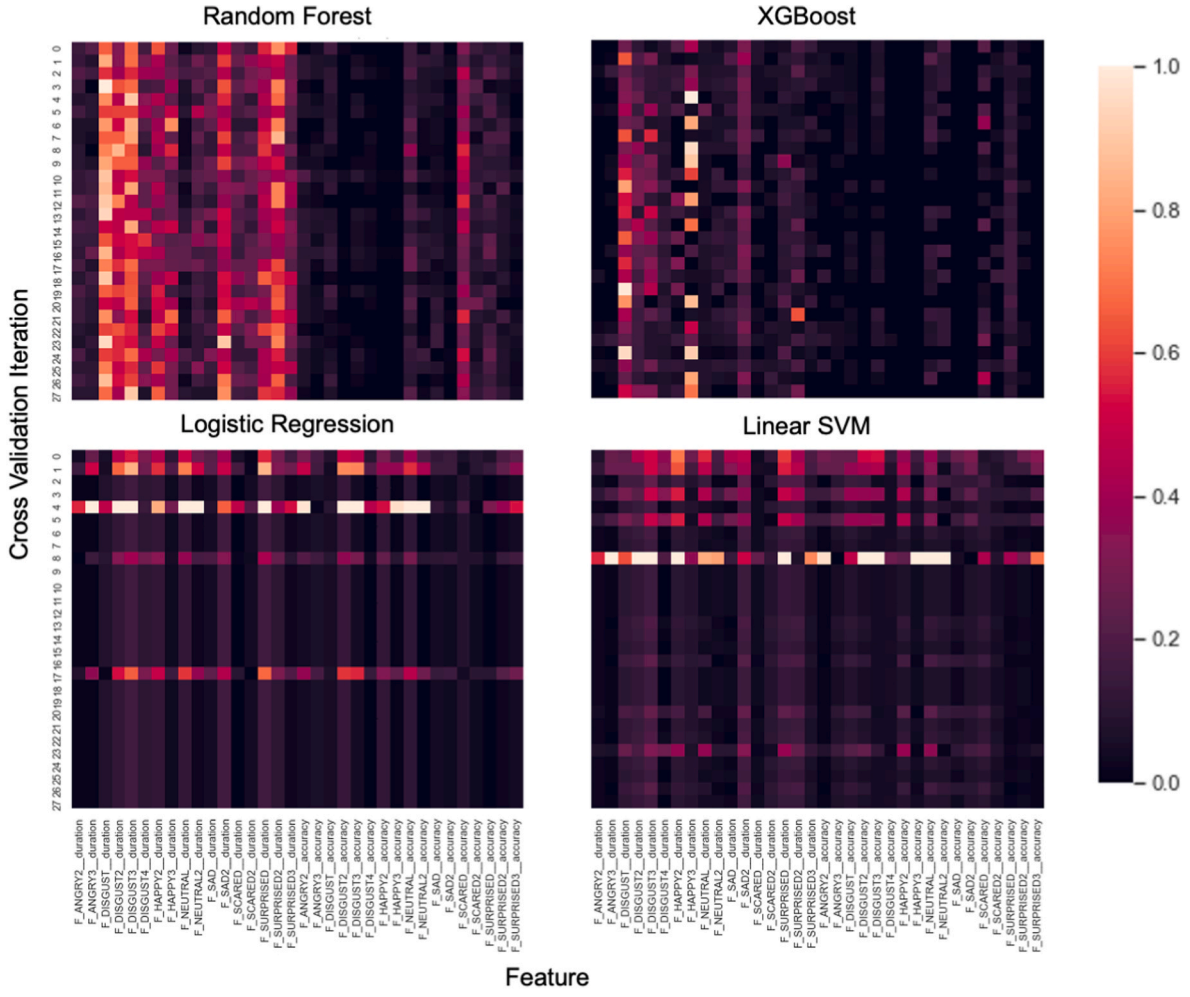

Fig. 2. Heatmaps of relative feature importance for each classier type (all features). The y axis corresponds to the 28 iterations of cross-validation. Plots are presented in decreasing order of ROC-AUC performance (as read left to right). Duration-based features tended to be most important in models that produced a higher AU-ROC.

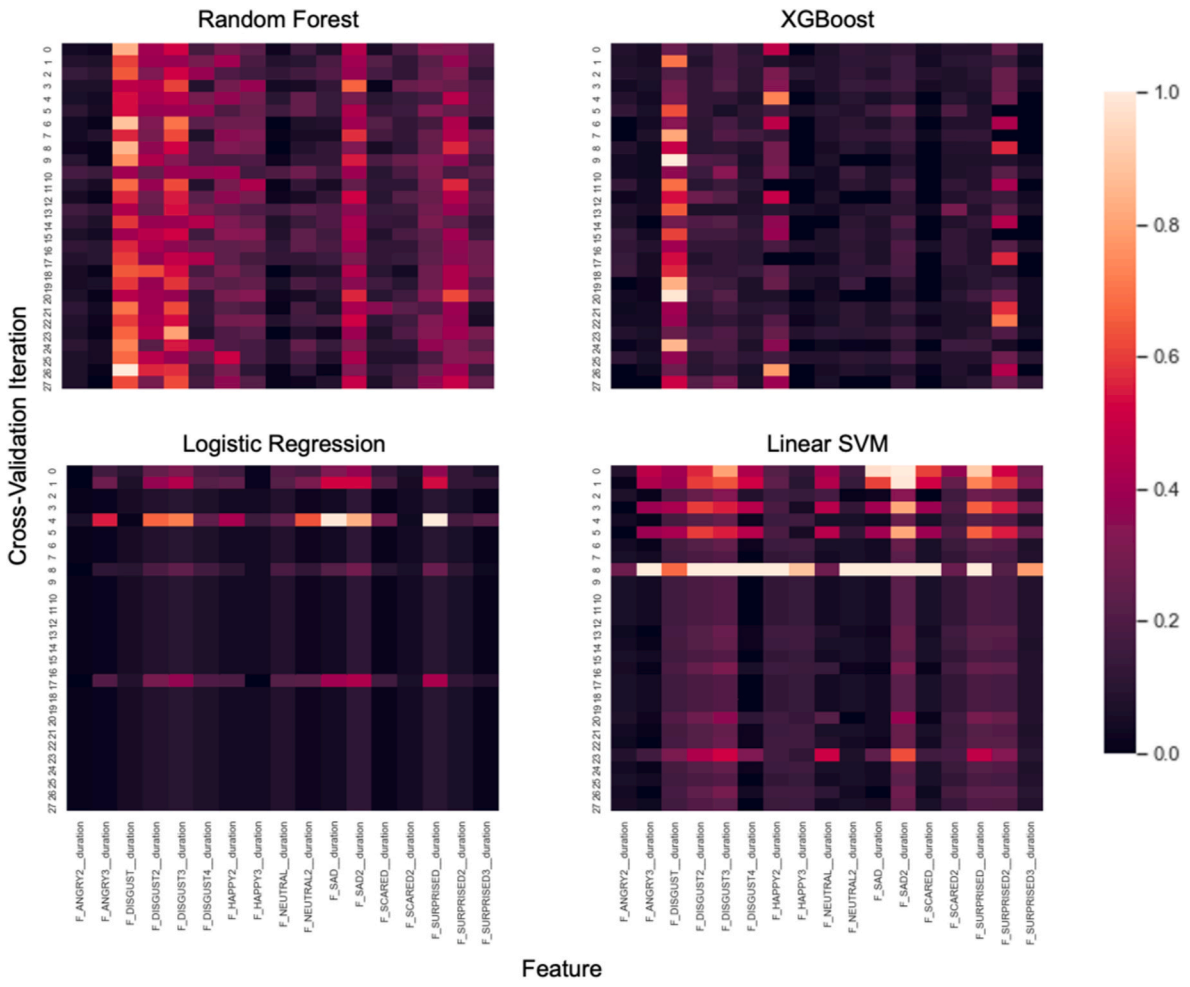

Fig. 3. Heatmaps of relative feature importance for each classier type (duration-only features). The y axis corresponds to the 28 iterations of cross-validation Plots are presented in decreasing order of AU-ROC performance (as read left to right).

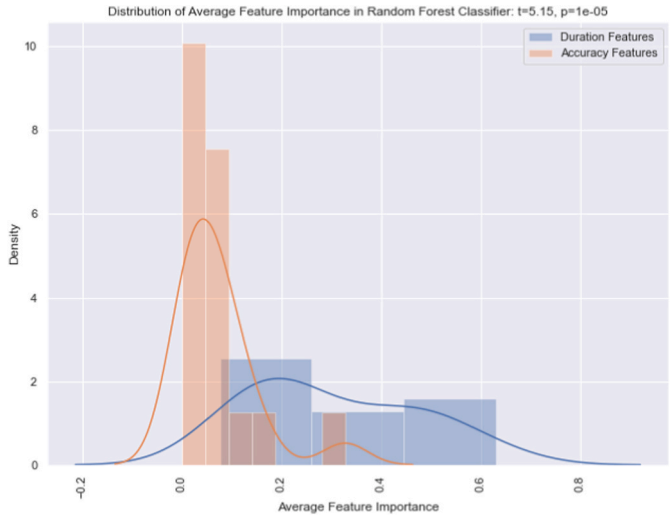

Fig. 4. Difference in average feature importance by feature type.

recall across the 28 iterations of cross validation. Later in this paper we will discuss the practical considerations of each metric and provide an argument for those off of which we should base final model selection.

3. Results

3.1. Model performance using full feature set

Using the full set of 34 features (all accuracy and duration measures), we obtained four heatmaps of relative feature importance across each of the 28 iterations of repeated, nested cross validation (Fig. 2). One heatmap exists for each type of classifier, and each row of the heatmap displays feature importances for the classifier that maximized AU-ROC during the hyperparameter tuning grid search. 28 rows of the heatmap correspond to the 28 models produced through the 28 iterations of repeated, nested cross validation.

When trained on the full set of 34 features, an XGBoost classifier produced the best model with respect to both AU-ROC (AU-ROC = 0.74) and recall (recall = 0.76). Random Forest performed second best with respect to both AU-ROC (AU-ROC = 0.73) and recall (recall = 0.76). Logistic regression and linear SVM performed noticeably worse than tree based methods with respect to recall, but SVM showed better precision than the other methods. In our discussion we will argue for selecting the model that results in the highest average of AU-ROC and Recall. A summary of all performance metrics for all models are found in Table 2.

3.2. Manual inspection of features

Manual inspection of feature importances revealed that duration-based features were, in general, most important in the two best performing tree-based classifiers. This pattern was most pronounced in the

Table 2
Cross-validated performance metrics for each classifier type obtained through hyperparameter grid search. Minority class (NT) was sampled using SMOTE to obtain equal class size as majority class (ASD).

Features	Mean AU-ROC		Mean Recall		Mean Accuracy		Mean Precision	
	All	Duration	All	Duration	All	Duration	All	Duration
Model								
XGBoost	**0.74***	0.70	**0.76***	0.72	0.71	0.67	0.64	0.61
Random Forest	0.73	**0.75***	0.76	**0.77***	**0.72***	**0.72***	0.65	**0.67***
Logistic Regression	0.67	0.69	0.67	0.69	0.65	0.68	0.60	0.67
Linear SVM	0.70	0.71	0.70	0.74	0.69	0.69	**0.70***	0.65

random forest (Fig. 2). Lower performing linear methods tended to spread feature importance out among both duration-based and accuracy-based features (Fig. 2).

More specifically, duration-based features corresponding to faces expressing disgust (in the tree-based methods) most consistently displayed high relative importance across the folds of cross-validation. This is most clearly seen in Fig. 2 Fig. 3, where duration based features are seen on the left side of the figures.

3.3. Model performance using reduced feature set

Three of the four classifiers saw comparable or improved performance with respect to AU-ROC and recall when trained only on the reduced subset of 17 duration-based features. XGBoost was the only classifier that performed worse when trained only on the reduced feature subset. Performance metrics for each model trained using both the full set of features and the duration-based subset are found in Table 2.

3.4. Feature importances using the reduced feature set

Features corresponding to the emotion "disgust" were consistently most important within the highest performing random forest classifier (random forest). Features corresponding to surprise and sadness were consistently highly important across all classifier types except for XGboost (Fig. 5).

When we aggregated feature importance by emotion (e.g., took the average of all features corresponding to a face showing disgust), the difference in importance between the most important feature, disgust, and all other emotions was most drastic in the random forest classifier (Fig. 5).

3.5. Classifier selection

As mentioned, we selected the simplest model resulting in the highest average of AU-ROC and Recall. According to this metric, the best performing model among both the full and reduced feature set was a random forest classifier. Because of the repeated, nested cross validation, we cannot report a single "best" set of hyperparameters.

4. Discussion

4.1. Novelty of method

To the extent of our knowledge, this was the first study to use naturalistic, smartphone-collected game play data to distinguish ASD from NT children in a non-clinical setting. Moreover, the objective nature of the data adds to a growing body of work demonstrating that digital phenotyping can successfully distinguish ASD from NT children [32,42].

Considerable work has focused on researching and developing objective methods of screening for ASD. Many of these methods rely either on genetic information or image and video data collected for use with computer vision algorithms. In this study, we expand on these

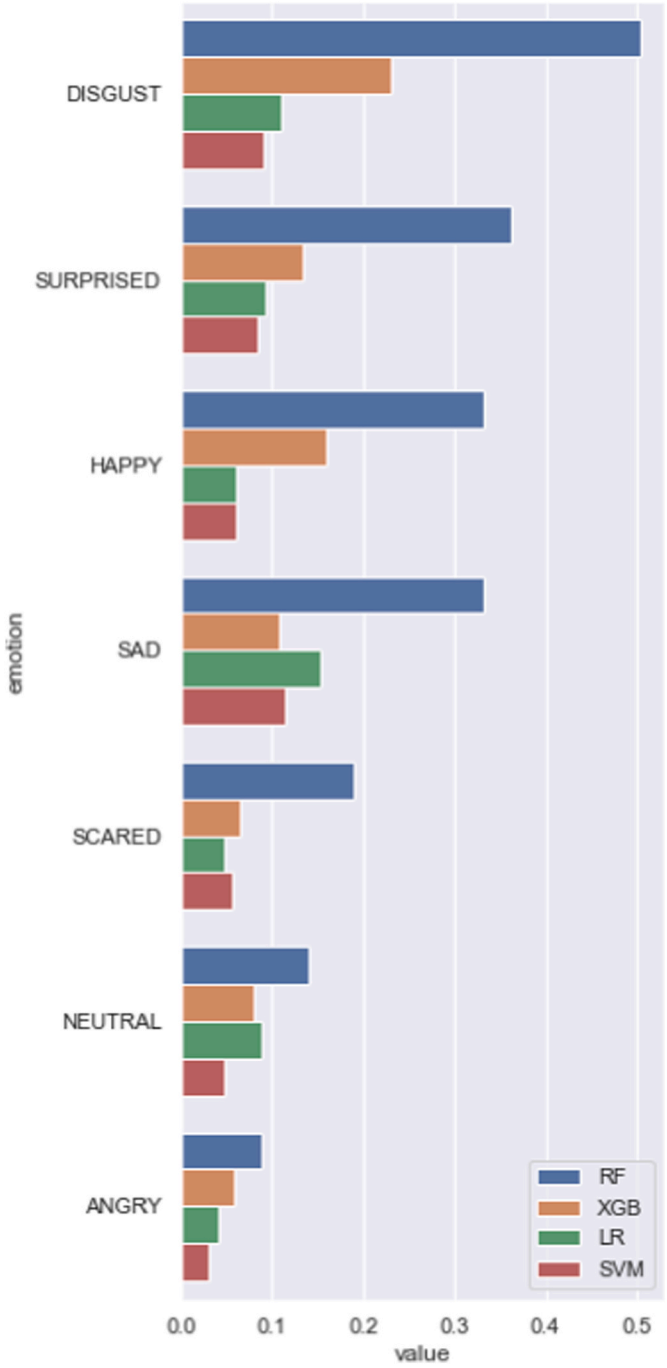

Fig. 5. Feature importances aggregated by emotion across all 4 families of classifiers.

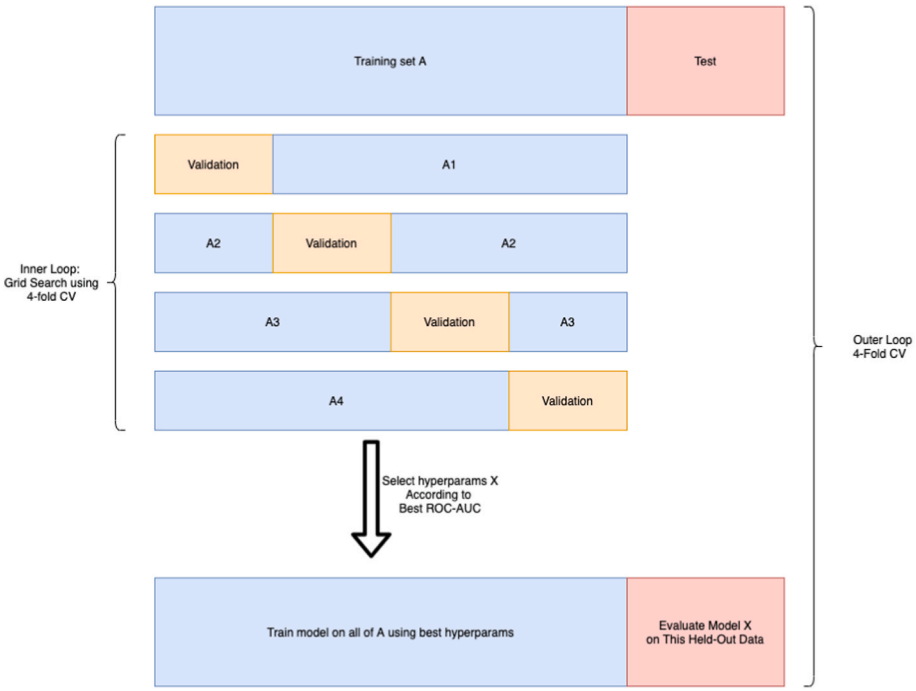

Fig. 6. Repeated Nested Cross Validation Procedure Used to separately tune model hyperparameters and evaluate out-of-sample performance.

modalities, demonstrating that using data collected through the use of a digital therapeutic, we are able to distinguish ASD from NT. These results can potentially generalize to other ubiquitous modalities such as wearable computers, which have proven to be clinically useful for addressing certain symptoms related to autism [43–52].

The significance of this is twofold. First, it is reasonable to expect that supplanting the aforementioned efforts to develop digital screening tools with the modality presented in this paper will produce more expressive models that can be used to screen for ASD in a privacy-preserved manner. Privacy concerns are at the forefront of behavioral phenotyping efforts (Washington et al., 2020). Second, GuessWhat? benefits from capturing information during naturalistic interactions between parents and their children. This has a clear benefit: the parent can intervene if the child begins to lose interest or pay less attention to the game. That said, involving the parent in gameplay introduces potential confounding effects of the parent's method of interacting with the game (e.g., the duration-based measures in this study could capture the speed with which a parent marks a prompt as correct as opposed to the speed with which it takes the child to answer the prompt).

5. Feature discussion

One of the most striking results of the study was the extent to which the highest performing models nearly exclusively found duration-based features important compared to accuracy based features. This may have been due to the low variance of the accuracy features compared to that of the duration based features. For both ASD and NT, accuracy metrics were skewed heavily towards 1.0 (i.e., always correct), suggesting that these features may not discriminate well between the two conditions. This trend could be driven by latent factors such as parents incorrectly labeling or a game design that was too easy for the majority of children regardless of diagnosis.

Positing that the accuracy features may have been introducing noise into our dataset, we opted to train models using only duration-based features, which improved performance in 3 of the 4 types of classifiers. Consequently, our final classifier was a random forest trained on just the duration subset of features (Fig. 3).

When we looked at the mean feature importance of features aggregated by emotion, features corresponding to disgust were the most important in the random forest classifier. It has been shown that younger children are worse at discriminating facial expressions of disgust and surprise when images of these expressions are presented alongside specific other emotions [53], suggesting that cognitive development is required for accurate processing of these emotions. This provides a compelling explanation of why we found features corresponding to disgust and surprise to be best at discriminating between ASD and NT children in our study.

That said, we must take care to not deduce an oversimplified understanding of these features. Specifically, although emotion recognition is necessary for a child to act out a prompt, it is but one component of a complex interaction between child performance and parent interpretation that could drive the signal found in this study. In this paper, we consider this complex didactic process to be a proxy for emotion recognition, but it likely captures components of theory of mind, metacognition and many other phenomena, as well.

6. Strengths and limitations

This study demonstrated that naturalistic gameplay data involving childrens' ability to identify and process facial emotions can be used to distinguish ASD and NT children. Moreover, the features that were most important to distinguishing the children were features corresponding to disgust and surprise, a finding consistent with previous literature [53]. Capturing this objective signal "in the wild" is a promising step forward in successfully developing novel methods of screening for ASD that complement existing instruments, resulting in more accurate and accessible methods of screening for the disorder.

The ability to capture this signal "in-the-wild" without the use of specialized equipment is extremely well situated for translation due to three key factors. First, an emerging model of remote care that emphasizes telehealth visits was catalyzed by COVID-19, and both clinicians and patients became accustomed to receiving care through digital tools. Second, there is increasing awareness of the disparities in diagnosing ASD both in rural areas and among low socioeconomic status groups [5,54]. Requiring only a smartphone with internet access, GuessWhat? could be used as part of a broader strategy of addressing

these disparities in care for ASD among different groups in the United States. Third, while there is some disagreement about the "patient as a consumer" model [55,56], when we narrow the discussion specifically to the ways in which patients interact with digital health tools and products, it would be naive to assume that patients are not sensitive to the experience of using a digital tool, especially when many of the most successful health and wellness apps provide exemplar experiences. As such, the low barrier to use and straightforward experience provided by GuessWhat? positions the platform well for high engagement from the relevant patient populations.

A limitation of this study is that the sample size was too small to evaluate the predictive power of our models across various demographic dimensions, including gender, ethnicity, nationality, age and other diagnosis (e.g. ADHD, dyslexia). As we expand our recruitment efforts, we plan to follow up on this work with models that are validated to assess fairly across demographic subdivisions. Additionally, to mitigate the possible bias introduced by parents being lenient with their own children, we should reproduce this study in a clinical setting in which the person displaying the prompts to the child is neither a parent nor informed of the child's diagnosis.

Finally, future work should attempt to elucidate the impact of the many dimensions of the child-parent dyadic relationship on the signal found in this study. Future work should specifically interrogate 1) the ability for a child to identify a prompt's emotion 2) the child's ability to introspect and express the emotion in a way that the parent would recognize 3) the ability of the parent to identify the relevant emotion and 4) the parent's ability press the button quickly.

Furthermore, the broader autistic phenotype (BAP) is a term that refers to the presence of certain autism-related traits in undiagnosed family members of children with autism [57]. These typically manifest as more mild impairments in social and communication abilities. Parents exhibiting the BAP could potentially drive the results found in this study.

Funding

This work was supported in part by funds to DPW from the National Institutes of Health (1R01EB025025-01, 1R21HD091500-01, 1R01LM013083, 1R01LM013364), 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, Stanford's Precision Health and Integrated Diagnostics Center (PHIND), Stanford's Beckman Center, Stanford's Bio-X Center, Predictives and Diagnostics Accelerator (SPADA) Spectrum, Stanford's Spark Program in Translational Research, Stanford mediaX, and Stanford's 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 Mr. Schroeder and the Stanford Interdisciplinary Graduate Fellowship (SIGF) as the Schroeder Family Goldman Sachs Graduate Fellow.

Declaration of competing interest

D.P.W. is the founder of Cognoa.com. This company is developing digital health solutions for pediatric healthcare. AK works as a part-time consultant to Cognoa.com.

Appendix

A1Prompts shown during gameplay

ANGRY_2	**NEUTRAL_2**
ANGRY_3	**SAD**
DISGUST	**SAD_2**
DISGUST_2	**SCARED**
DISGUST_3	**SCARED_2**
DISGUST_4	**SURPRISED**
HAPPY_2	**SURPRISED_2**
HAPPY_3	**SURPRISED_3**
NEUTRAL	

A2Data Schema

ANGRY_2_duration	ANGRY_2_accuracy	DISGUST_2_duration	DISGUST_2_accuracy	……	SURPRISED_3_duration	SURPRISED_3_accuracy

References

[1] Kadir MA. Role of telemedicine in healthcare during COVID-19 pandemic in developing countries. TMT; Apr. 2020.
[2] Gunasekeran DV, Tham Y-C, Ting DSW, Tan GSW, Wong TY. Digital health during COVID-19: lessons from operationalising new models of care in ophthalmology. Lancet Digit Health 2021;3(2):e124–34.
[3] Inkster B, O'Brien R, Selby E, Joshi S, Subramanian V, Kadaba M, Schroeder K, Godson S, Comley K, Vollmer SJ, Mateen BA. Digital health management during and beyond the COVID-19 pandemic: opportunities, barriers, and recommendations. JMIR Ment Health 2020;7(7):e19246.
[4] Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Publ Health 2021;42:463–81.
[5] Antezana L, Scarpa A, Valdespino A, Albright J, Richey JA. Rural trends in diagnosis and services for autism spectrum disorder. Front Psychol 2017;8:590.
[6] Ning M, Daniels J, Schwartz J, Dunlap K, Washington P, Kalantarian H, Du M, Wall DP. Identification and quantification of gaps in access to autism resources in the United States: an infodemiological study. J Med Internet Res 2019;21(7):e13094.
[7] Elder JH, Kreider CM, Brasher SN, Ansell M. Clinical impact of early diagnosis of autism on the prognosis and parent-child relationships. Psychol Res Behav Manag 2017;10:283–92.
[8] Greenhalgh T, Procter R, Wherton J, Sugarhood P, Shaw S. The organising vision for telehealth and telecare: discourse analysis. BMJ Open Jul. 2012;2(4).
[9] Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry 2015;5:e514.
[10] Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD, Transl. Psychiatry 2016;6:e732.
[11] Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inf Assoc 2018;25(8):1000–7.

[12] Abbas H, Garberson F, Liu-Mayo S, Glover E, Wall DP. Multi-modular AI approach to streamline autism diagnosis in young children. Sci Rep 2020;10(1):5014.
[13] Washington P, Mutlu OC, Leblanc E, Kline A, Hou C, Chrisman B, Stockham N, Paskov K, Voss C, Haber N, Wall D. Using crowdsourcing to train facial emotion machine learning models with ambiguous labels. Jan. 2021. arXiv:2101.03477.
[14] Washington P, Leblanc E, Dunlap K, Penev Y, Varma M, Jung J-Y, Chrisman B, Sun MW, Stockham N, Paskov KM, Kalantarian H, Voss C, Haber N, Wall DP. Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. Pac Symp Biocomput 2021;26:14–25.
[15] Leblanc E, Washington P, Varma M, Dunlap K, Penev Y, Kline A, Wall DP. Feature replacement methods enable reliable home video analysis for machine learning detection of autism. Sci Rep 2020;10(1):21245.
[16] Washington P, Leblanc E, Dunlap K, Penev Y, Kline A, Paskov K, Sun MW, Chrisman B, Stockham N, Varma M, Voss C, Haber N, Wall DP. Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for Video-Based autism feature recognition. J Personalized Med Aug. 2020;10(3).
[17] Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, Kalantarian H, Penev Y, Paskov K, Voss C, Stockham N, Varma M, Husic A, Kent J, Haber N, Winograd T, Wall DP. Crowdsourced feature tagging for scalable and privacy-preserved autism diagnosis. Dec. 2020.
[18] Levy S, Duda M, Haber N, Wall DP. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Mol Autism 2017;8:65.
[19] Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry 2012;2:e100.
[20] Duda M, Kosmicki JA, Wall DP. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Transl Psychiatry 2014;4:e424.
[21] Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, Kalantarian H, Penev Y, Paskov K, Voss C, Stockham N, Varma M, Husic A, Kent J, Haber N, Winograd T, Wall DP. Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Sci Rep 2021;11(1):7620.
[22] Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, Kalantarian H, Khan NZ, Darmstadt GL, Wall DP. Detecting developmental delay and autism through machine learning models using home videos of bangladeshi children: development and validation study. J Med Internet Res 2019;21(4):e13822.
[23] Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: a development and prospective validation study. PLoS Med 2018;15(11):e1002705.
[24] Khowaja MK, Hazzard AP, Robins DL. Sociodemographic barriers to early detection of autism: screening and evaluation using the M-CHAT, M-CHAT-R, and Follow-Up. J Autism Dev Disord 2015;45(6):1797–808.
[25] Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall DP. Guess what?: towards understanding autism from structured video using facial affect. Int J Healthc Inf Syst Inf 2019;3:43–66.
[26] Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall D. A gamified mobile system for crowdsourcing video for autism research. In: 2018 IEEE international conference on healthcare informatics. ICHI; 2018. p. 350–2.
[27] Kalantarian H, Jedoui K, Washington P, Wall DP. A mobile game for automatic Emotion-Labeling of images. IEEE Trans Games 2020;12(2):213–8.
[28] Kalantarian H, Jedoui K, Washington P, Tariq Q, Dunlap K, Schwartz J, Wall DP. Labeling images with facial emotion and the potential for pediatric healthcare. Artif Intell Med 2019;98:77–86.
[29] Kalantarian H, Jedoui K, Dunlap K, Schwartz J, Washington P, Husic A, Tariq Q, Ning M, Kline A, Wall DP. The performance of emotion classifiers for children with Parent-Reported autism: quantitative feasibility study. JMIR Ment Health 2020;7(4):e13174.
[30] Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, Hou C, Mutlu C, Dunlap K, Penev Y, Varma M, Stockham N, Chrisman B, Paskov K, Sun MW, Jung J-Y, Voss C, Haber N, Wall DP. Training an emotion detection classifier using frames from a mobile therapeutic game for children with developmental disorders. Dec. 2020, 08678. arXiv:2012.
[31] Washington P, Kline A, Mutlu OC, Leblanc E, Hou C, Stockham N, Paskov K, Chrisman B, Wall DP. Activity recognition with moving cameras and few training examples: applications for detection of Autism-Related headbanging. Jan. 2021. arXiv:2101.03478.
[32] Washington P, Park N, Srivastava P, Voss C, Kline A, Varma M, Tariq Q, Kalantarian H, Schwartz J, Patnaik R, Chrisman B, Stockham N, Paskov K, Haber N, Wall DP. Data-Driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging 2020;5(8):759–69.
[33] Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318(13):1215–6.
[34] Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatr 2018;17(3):276–7.
[35] Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2:88.
[36] Penev Y, Dunlap K, Husic A, Hou C, Washington P, Leblanc E, Kline A, Kent J, Ng-Thow-Hing A, Liu B, Harjadi C, Tsou M, Desai M, Wall DP. A mobile game platform for improving social communication in children with autism: a feasibility study. Appl Clin Inf 2021;12(5):1030–40.
[37] LoBue V, Thrasher C. The child affective facial expression (CAFE) set: validity and reliability from untrained adults. Front Psychol 2014;5:1532.
[38] Rump KM, Giovannelli JL, Minshew NJ, Strauss MS. The development of emotion recognition in individuals with autism. Child Dev 2009;80(5):1434–47.
[39] Kuusikko S, Haapsamo H, Jansson-Verkasalo E, Hurtig T, Mattila M-L, Ebeling H, Jussila K, Bölte S, Moilanen I. Emotion recognition in children and adolescents with autism spectrum disorders. J Autism Dev Disord 2009;39(6):938–45.
[40] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Jun. 2011. arXiv:1106.1813.
[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in python. Jan. 2012. arXiv:1201.0490.
[42] Nag A, Haber N, Voss C, Tamura S, Daniels J, Ma J, Chiang B, Ramachandran S, Schwartz J, Winograd T, Feinstein C, Wall DP. Toward continuous social phenotyping: analyzing gaze patterns in an emotion recognition task for children with autism through wearable smart glasses. J Med Internet Res 2020;22(4):e13810.
[43] Haber N, Voss C, Wall D. Making emotions transparent: google glass helps autistic kids understand facial expressions through augmented-reaiity therapy. IEEE Spectrum 2020;57(4):46–52.
[44] Kline A, Voss C, Washington P, Haber N, Schwartz H, Tariq Q, Winograd T, Feinstein C, Wall DP. Superpower glass, GetMobile. Mobile Comput Commun 2019;23(2):35–8.
[45] Voss C, Washington P, Haber N, Kline A, Daniels J, Fazel A, De T, McCarthy B, Feinstein C, Winograd T, Wall D. Superpower glass: delivering unobtrusive real-time social cues in wearable systems. Feb. 2020, 06581. arXiv:2002.
[46] Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, Tariq Q, Robinson TN, Desai M, Phillips JM, Feinstein C, Winograd T, Wall DP. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. JAMA Pediatr 2019;173(5):446–54.
[47] Daniels J, Schwartz JN, Voss C, Haber N, Fazel A, Kline A, Washington P, Feinstein C, Winograd T, Wall DP. Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. NPJ Digit Med 2018;1:32.
[48] Washington P, Voss C, Haber N, Tanaka S, Daniels J, Feinstein C, Winograd T, Wall D. A wearable social interaction aid for children with autism. In: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems, CHI EA '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 2348–54.
[49] Washington P, Voss C, Kline A, Haber N, Daniels J, Fazel A, De T, Feinstein C, Winograd T, Wall D, SuperpowerGlass. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):1–22.
[50] Daniels J, Haber N, Voss C, Schwartz J, Tamura S, Fazel A, Kline A, Washington P, Phillips J, Winograd T, Feinstein C, Wall DP. Feasibility testing of a wearable behavioral aid for social learning in children with autism. Appl Clin Inf 2018;9(1):129–40.
[51] Voss C, Haber N, Washington P, Kline A, McCarthy B, Daniels J, Fazel A, De T, Feinstein C, Winograd T, Wall D. Designing a holistic At-Home learning aid for autism. Feb. 2020. arXiv:2002.04263.
[52] Daniels J, Schwartz J, Haber N, Voss C, Kline A, Fazel A, Washington P, De T, Feinstein C, Winograd T, Wall D. 5.13 design and efficacy of a wearable device for social affective learning in children with autism. J Am Acad Child Adolesc Psychiatry 2017;56(10):S257.
[53] Gagnon M, Gosselin P, Hudon-ven der Buhs I, Larocque K, Milliard K. Children's recognition and discrimination of fear and disgust facial expressions. J Nonverbal Behav 2010;34(1):27–42.
[54] Delobel-Ayoub M, Ehlinger V, Klapouszczak D, Maffre T, Raynaud J-P, Delpierre C, Arnaud C. Socioeconomic disparities and prevalence of autism spectrum disorders and intellectual disability. PLoS One 2015;10(11):e0141964.
[55] Goldstein MM, Bowers DG. The patient as consumer: empowerment or commodification? currents in contemporary bioethics. J Law Med Ethics 2015;43(1):162–5.
[56] Gusmano MK, Maschke KJ, Solomon MZ. Patient-Centered care, yes; patients as consumers, no. Health Aff 2019;38(3):368–73.
[57] Gerdts J, Bernier R. The broader autism phenotype and its implications on the etiology and treatment of autism spectrum disorders. Autism Res Treat 2011;2011:545901.