HHS Public Access

Author manuscript

Intell Based Med. Author manuscript; available in PMC 2022 May 27.

Published in final edited form as:

Intell Based Med. 2022 ; 6: . doi:10.1016/j.ibmed.2022.100056.

Crowd annotations can approximate clinical autism impressions from short home videos with privacy protections

Peter Washington[a,*], **Brianna Chrisman**[a], **Emilie Leblanc**[b], **Kaitlyn Dunlap**[b], **Aaron Kline**[b], **Cezmi Mutlu**[c], **Nate Stockham**[d], **Kelley Paskov**[e], **Dennis Paul Wall**[f]

[a]Department of Bioengineering, Stanford University, United States

[b]Department of Pediatrics (Systems Medicine) Stanford University, United States

[c]Department of Electrical Engineering, Stanford University, United States

[d]Department of Neuroscience, Stanford University, United States

[e]Department of Biomedical Data Science, Stanford University, United States

[f]Department of Pediatrics (Systems Medicine) and Biomedical Data Science, Stanford University, United States

Abstract

Artificial Intelligence (A.I.) solutions are increasingly considered for telemedicine. For these methods to serve children and their families in home settings, it is crucial to ensure the privacy of the child and parent or caregiver. To address this challenge, we explore the potential for global image transformations to provide privacy while preserving the quality of behavioral annotations. Crowd workers have previously been shown to reliably annotate behavioral features in unstructured home videos, allowing machine learning classifiers to detect autism using the annotations as input. We evaluate this method with videos altered via pixelation, dense optical flow, and Gaussian blurring. On a balanced test set of 30 videos of children with autism and 30 neurotypical controls, we find that the visual privacy alterations do not drastically alter any individual behavioral annotation at the item level. The AUROC on the evaluation set was 90.0% ±7.5% for unaltered videos, 85.0% ±9.0% for pixelation, 85.0% ±9.0% for optical flow, and 83.3% ±9.3% for blurring, demonstrating that an aggregation of small changes across behavioral questions can collectively result in increased misdiagnosis rates. We also compare crowd answers against clinicians who provided the same annotations for the same videos as crowd workers, and we find that clinicians have higher sensitivity in their recognition of autism-related symptoms. We also find that there is a linear correlation ($r = 0.75$, $p < 0.0001$) between the mean Clinical Global Impression (CGI) score provided by professional clinicians and the corresponding score emitted by a previously validated autism classifier with crowd inputs, indicating that the classifier's output probability is a reliable estimate of the clinical impression of autism. A significant correlation is

*Corresponding author.: peterwashington@stanford.edu (P. Washington).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: DPW is the founder of Cognoa.com. This company is developing digital health solutions for pediatric care. All other authors declare no competing interests.

maintained with privacy alterations, indicating that crowd annotations can approximate clinician-provided autism impression from home videos in a privacy-preserved manner.

1. Introduction

While artificial intelligence (A.I.) approaches are needed for healthcare to achieve scale and consistency, current A.I.-powered solutions to diagnostics in psychiatry and the behavioral sciences are under-performant due to the complexity of the underling behaviors. Until A.I. has advanced to a point where it can seamlessly pass the Turing Test [40] and understand human social behavior, with all its subtleties and nuances, to a degree that surpasses an average human's social acuity (i.e., to the level of a licensed clinician), it is unlikely that A.I. will replace any behavioral healthcare jobs. In the meantime, A.I. solutions which can augment the capabilities of non-expert humans may allow for scalable and accessible remote diagnostics while easing the burden of professional clinicians and the healthcare system to provide initial consultations in person.

A.I. is increasingly being considered to help healthcare solutions scale. For such solutions to be deployed in clinical settings, the system must garner maximal trust from all stakeholders. Privacy rises to the forefront of patient concern when humans are incorporated into the diagnostic pipeline [57]. Many parents are uncomfortable with sharing raw videos of their children with strangers online even if the purpose of sharing is to help the parents receive affordable and accessible diagnostic services [53]. To help ameliorate this uneasiness, privacy concerns must be directly addressed in any scalable solution where the humans in the loop are strangers (i.e., crowdsourced workers).

Here, we address these issues by studying an A.I.-augmented pipeline for detecting Autism Spectrum Disorder, or autism, from unstructured home videos. Autism is a developmental delay which is currently estimated to affect 1 in 40 children [28]. While access to care requires a formal diagnosis, access to diagnostic services is severely limited for many families, thus limiting potential care. Some evidence suggests that as much as 80% of families in the United States lack access to care [35], and underserved populations are disproportionately affected [17]. A.I. powered telemedical solutions therefore have the possibility to help these families, many of whom may otherwise lack access to traditional health services.

Crowd-powered telemedical diagnostic tools can provide parents with a risk score and an associated probability for a diagnosis, and prior works have repeatedly demonstrated the success of A.I. models to successfully detect autism using solely annotations provided by non-clinical human workers [5–8,20,30,38,39,46,48,49,51–54,58]. This family of solutions utilizes a distributed crowdsourced workforce to quickly annotate behavioral features displayed in videos recorded during parent-administered home autism therapy sessions using mobile digital health therapies [21–25] and wearable augmented reality solutions [2–4,13,14,15,27,34,41–44,55,56].

In order for such solutions to truly scale for annotation by a large crowd workforce, the privacy of the patients must be preserved. This is important in general for healthcare

applications, but it is especially critical when the patients in question are young children with a developmental delay and observed by a stranger in their home. Our prior work has shown that applying targeted privacy-preserving modifications to videos, such as pitch shifting and covering the child's face with a virtual box, results in minimal degradation of the quality of crowdsourced annotations used for remote detection of autism-related behaviors [53]. However, such lightweight privacy protections may be insufficient to some patients, such as those who do not want the interior of their home exposed to strangers on the Internet.

Here, we explore the effect of standard visual privacy-preserving mechanisms on the annotation quality of videos of children with autism and matched neurotypical controls. We apply pixelation, dense optical flow, and Gaussian blurring, which are either standard methods for protecting the privacy of human subjects in image and video datasets [36] or for representing visual features for activity recognition algorithms [26]. We study the effect of these video transformations on annotation quality. On a balanced test set of 60 videos of children with autism and matched neurotypical controls, we find that no visual privacy condition deviates from the unaltered condition by more than half of a categorical ordinal severity point out of 4 questions corresponding to behavior severity. We compare crowd responses to professional clinicians and find that the probability score emitted by the classifier is consistent with the clinicians' global impression from watching the same video. We also find that, as expected, clinicians are more adept at identifying autism-related symptoms than crowd workers.

2. Methods

2.1. Balanced video dataset

We leveraged a balanced video dataset of 30 children with autism and 30 controls without autism. Both groups were gender and aged matched: we posted 30 videos of male children (13 with autism and 17 neurotypical) and 30 videos of female children (17 with autism and 13 neurotypical). The mean age in the videos of children with autism was 3.49 years old (SD = 1.58 years old), and the mean age in the videos of children without autism was 3.41 years old (SD = 1.39 years old).

2.1.1. Logistic regression classifier for predicting autism—We utilized a previously validated [38,39,53] logistic regression classifier for predicting autism vs. not autism from the answers to the multiple-choice questions we asked crowd workers. The classifier [29, 31] was derived from the Autism Diagnostic Observation Schedule (ADOS) module 2 [32] and was trained on clinician filled scoresheets from the Boston Autism Consortium (AC), the Simons Simplex Collection v14 (SSC) [10], Autism Genetic Resource Exchange (AGRE) [11], National Database of Autism Research (NDAR) [16] and the Simons Variation in Individuals Project (SVIP) [37].

To generate confidence intervals, we performed 10,000 iterations of a bootstrapping procedure for each video. In each iteration, we sampled with replacement from the 60 videos used for evaluation and computed the accuracy, precision, recall/sensitivity, specificity,

Area Under the Receiver Operating Characteristic (AUROC), and Area Under the Precision-Recall Curve (AUPRC) on the resulting video set.

2.1.2. Privacy conditions—We evaluated three privacy modifications applied to each video: pixelation (Fig. 1), dense optical flow (Fig. 2), and Gaussian blurring (Fig. 3). Pixelation and Gaussian blurring are common methods for protecting the privacy of human subjects in image and video datasets [36]. Dense optical flow is a standard method for representing visual features for activity recognition [26] but also obfuscates much of the visual features of a frame, resulting in privacy preservation. To apply pixelation, we first resized the input frames down to 32 × 32 pixels. We then calculated the final pixelated frame by resizing the smaller frame back to the original frame size using bilinear interpolation. To calculate dense optical flow, we applied Farnebäck's algorithm for two-frame motion estimation based on polynomial expansion [9]. The image was colored through obtaining a 2-channel array with optical flow vectors, where the direction of the vectors corresponds to the hue of the image and the magnitude of the vectors corresponds to the value (lightness) of the image. To apply Gaussian blurring, we used a blurring kernel with 1/4th the width and height of the input image.

We note that pixelation (Fig. 1) makes the background setting and human more discernible (less private) than Gaussian blurring (Fig. 3), and Gaussian blurring is more discernible (less private) dense optical flow (Fig. 2).

2.1.3. Privacy condition analysis—To understand the effects of the strength of the privacy condition on worker rating ability, we explored the effects of altering both blurring and pixelation intensity on crowd ratings. We did not conduct this analysis for optical flow since there is no clear parameterization of intensity level for this method. We used a separate set of 28 videos balanced by diagnosis, child age, and child gender. Crowd workers were randomly assigned groups. As with the main study, no crowd worker was assigned to two separate conditions (alteration, intensity level) of a video.

For blurring, we used a blurring kernel with 1/6th, 1/5th, 1/4th, 1/3rd, and half the width and height of the input video as well as a blurring kernel the size of the entire original video. For pixelation, we set six different intermediate frame sizes when resizing the image using interpolation: 96 × 96, 64 × 64, 48 × 48, 32 × 32, 16 × 16, and 8 × 8 pixels.

2.1.4. Crowd annotation—To recruit crowd workers, we posted a video rating task on the crowdsourcing website Microworkers.com [18]. In this video rating task, crowd workers were asked to answer a series of 13 multiple choice questions, with 4 answer choices each, about the child's behavior exhibited in each of 8 video videos (4 featuring a child with autism and 4 featuring a child without autism). We used the categorical ordinal variables corresponding to each worker's answers as the inputs to a pretrained binary logistic regression classifier for autism (see subsection Logistic regression classifier for predicting autism below for details about classifier training). 1,000 crowd workers completed the task and passed basic quality control checks for answer acceptance. Quality control measures included time spent on the annotation task and deviations in answers between videos [53].

To identify workers who were invited to participate in the final study, we measured the classifier's prediction on all 8 videos for all workers. We then measured the mean probability of the correct class (PCC) for each worker across all 8 videos, where the PCC is the classifier's output probability p when the true class of a video is autism and $1-p$ otherwise. Out of the 1,000 workers evaluated, exactly 40 workers had a mean PCC at or above 80%. We recruited these 40 crowd workers to rate the primary video set of 60 videos used in this study.

Each of the 40 crowd workers were tasked with rating all 60 videos described in the subsection Balanced video dataset. We randomly split the 40 workers into 4 groups, and each group was assigned to one privacy condition per video. Therefore, no worker saw more than one version of each video. All workers saw exactly 15 unaltered videos, 15 videos with pixelation, 15 videos with dense optical flow, and 15 videos with Gaussian blurring. This ensured that no privacy condition was affected by any crowd worker biases.

2.1.5. Clinician annotation—We recruited 19 clinicians to rate the same balanced video dataset and provided the same categorical ordinal video-wide annotations as the crowd workers. All clinicians were licensed professionals who provide diagnoses of autism as part of their job duties. We asked all clinicians an additional question which crowd workers were not asked: "*Do you think the child has autism?*" The answer choices were:

- No, I am confident the child does not have autism (0)
- No, but I am unsure (1)
- Yes, but I am unsure (2)
- Yes, I am confident the child has autism (3)

All 60 videos received at least 1 rating by a clinician. Some videos were rated by more than 1 clinician, in which case we recorded the mean of the clinician answers for that video. For the "*Do you think the child has autism?*" question, we coded the responses from 0 to 3 as shown above.

Clinicians were also asked to provide a Clinical Global Impression (CGI) [12] rating for the children in the videos. The CGI scale measures the "severity of illness" between 1 ("normal, not at all ill") to 7 ("among the most extremely ill patients") and is designed to allow clinicians to provide a global impression without providing a formal diagnosis.

2.1.6. Item-level analysis—For each question that we asked crowd workers, we measured the mean absolute deviation of the mean answer for each privacy condition from the mean answer for the baseline condition. This difference provides a measure of the privacy condition's effect on annotation quality. We hypothesized that some questions would be more susceptible to alteration with certain privacy conditions applied.

3. Results

All procedures performed in studies involving human participants were approved by the Stanford University Institutional Review Board and are in accordance with the 1964 Helsinki declaration and its later amendments.

3.1. Performance of clinicians and crowd workers

Out of the 30 videos of children with autism, clinicians rated 8 videos confidently (with a mean autism rating above 2.5 out of 3.0). By contrast, clinicians rated 18 of the videos of neurotypical children confidently (with a mean autism rating below 0.5 out of 3.0). All 8 confidently rated videos were of children with autism, while 16 of the 18 neurotypical children were correctly identified (only 2 were actually diagnosed with autism). This suggests that clinicians observing remote videos of children are cautious about calling an autism diagnosis, but when they do guess a diagnosis, the child is very likely to actually have autism.

The clinician's classifier correctly identified 28 of the 30 autism cases while only correctly identifying 14 of the 30 neurotypical cases. By contrast, the crowd's classifier correctly identified 25 of the 30 autism cases and 27 of the 30 neurotypical cases. This suggests that clinicians are more sensitive to autism-related symptoms than crowd workers, thus resulting in a higher frequency of autism diagnoses by the binary classifiers.

There is a clear linear correlation ($r = 0.75$, $p < 0.001$) between the mean Clinical Global Impression (CGI) score provided by professional clinicians for each video and the corresponding classifier score emitted by the logistic regression classifier with crowd inputs (Fig. 4). This suggests that crowd responses in conjunction with machine learning algorithms can approximate clinician intuition, and the classifier's output can be interpreted as a reliable approximation of clinical global impressions of autism (see Fig. 5).

Of the 60 unaltered videos, the mean performance metrics were 90.0% ±7.5% accuracy, 92.9% ±8.9% precision, 86.7% ±11.8% recall (sensitivity), and 93.3% ±8.6% specificity. The mean AUROC was 90.0% ±7.5% and the mean AUPRC was 93.1% ±6.2%.

3.2. Effect of privacy conditions

The linear correlation between the mean Clinical Global Impression (CGI) score provided by professional clinicians for each video and the corresponding classifier score emitted by the logistic regression classifier with crowd inputs is maintained with privacy-preserving video modifications. The correlation is weaker for Gaussian blurring ($r = 0.64$, $p = 0.001$ for Gaussian blurring) than for dense optical flow and pixelation ($r = 0.71$, $p = 0.0002$ for pixelation for both).

With pixelation, the mean performance metrics were 85.0% ±9.2% accuracy, 88.9% ±12.1% precision, 80.0% ±14.4% recall (sensitivity), and 90.0% ±10.9% specificity. The mean AUROC was 85.0% ±9.0% and the mean AUPRC was 89.4% ±7.9%.

With dense optical flow, the mean performance metrics were 85.0% ±9.2% accuracy, 81.8% ±13.1% precision, 90.0% ±10.9% recall (sensitivity), and 80.0% ±14.4% specificity. The mean AUROC was 85.0% ±9.0% and the mean AUPRC was 88.4% ±7.9%.

With Gaussian blurring, the mean performance metrics were 83.3% ±9.2% accuracy, 83.3% ±13.7% precision, 83.3% ±13.7% recall (sensitivity), and 83.3% ±13.7% specificity. The mean AUROC was 83.3% ±9.3% and the mean AUPRC was 87.5% ±8.5%.

3.3. Effect of blurring and pixelation intensity

Interestingly, we did not observe a dramatic difference between the privacy alteration intensities depicted in Figs. 1 and 3. Table 1 shows the mean performance metrics on a separate testing set with blurring intensities created using blurring kernel sizes of 1/6th, 1/5th, 1/4th, 1/3rd, and ½ the size of the original image as well as a blurring kernel the entire size of the image. Table 2 shows the performance on a third disjoint testing with pixelation intensities created using an intermediate frame size of 96 × 96, 64 × 64, 48 × 48, 32 × 32, 16 × 16, and 8 × 8. These results indicate that the most dramatic privacy-preserving alternations can be applied with minimal to no degradation of performance across the behavioral questions we asked crowd workers.

3.4. Item-level analysis

Table 3 displays the mean absolute deviation of the mean answer for each privacy condition from the mean answer for the baseline condition. This difference provides a measure of the privacy condition's effect on annotation.

We found that pixelation resulted in smaller deviations from the unmodified video condition compared to dense optical flow and Gaussian blurring. In all but one behavioral annotation (sharing excitement), the mean deviation for pixelation was less than the other two privacy conditions. Dense optical flow, which provides maximal privacy, did not have a discernible difference from Gaussian blurring (private but less so), providing support for the use of dense optical flow in translational settings.

The annotation with the lowest deviation across all conditions was for displaying aggressive behavior. This result matches intuition, as no aggressive behavior was displayed in any of the videos we presented.

None of the 9 behaviors used for the classifiers contained a mean deviation above 0.5; the mean deviation is less than one half of the distance between one categorical ordinal variable representing symptom severity and the variable indicating one severity level higher (all questions contained 4 multiple choice options). We note that while these deviations are consistently small, the aggregation of these deviations results in higher rates of misclassification (see *Results: Effect of privacy conditions*).

4. Discussion and conclusion

We explored the potential for global image transformations to provide privacy for video subjects while preserving behavioral annotation quality. While no individual question was

drastically degraded when privacy alterations were applied, some behavioral annotations were degraded more than others. Pixelation consistently resulted in less drastic degradations than blurring and optical flow. We also found that the classifier's predictions from the crowd's annotation of the unaltered videos were strongly correlated with clinician global impressions. A slightly weaker correlation persisted even after all privacy modifications we tested, providing evidence that the classifier's output can be considered as an estimation of clinical global autism impression scores even when annotations are provided for a privacy-preserved video.

There are several limitations to the present study. While the configuration of questions we asked workers and clinicians resulted in worse performance by the classifier using the clinicians' annotations, this could have been due to over-sensitivity of the classifier rather than anything the clinicians did incorrectly. We therefore do not make any claims about the performance of clinicians as compared to crowd workers. An interesting limitation is that the definition of autism tends to shift over time with evolving DSM criteria and clinical practices [1, 33]. Because clinicians were providing annotations several years after the videos of children were recorded, it is possible that the children who did not qualify for a diagnosis at the time the videos were recorded would qualify for a diagnosis by the time the clinicians reviewed the videos.

There are several interesting avenues of future work. The annotations provided by crowd workers can potentially be used to train computer vision classifiers detecting behaviors relevant to autism detection such as emotion evocation [15,19,45,50], hand or head stimming [47], and abnormal eye contact. While humans are worse at detecting certain behavioral patterns from videos when privacy mechanisms are applied, it is possible that convolutional neural networks can more easily detect these features by learning subtle and nonlinear feature maps beyond human comprehension. Alternative privacy-preserving video alteration methods using deep learning could also be explored, such as generative adversarial networks (GANs). GANs can create privacy-preserved versions of the input video. Similarly, variational autoencoders can learn privacy-based features in the latent space. As machine learning algorithms continue to improve and relevant databases become more plentiful, the possibility of removing humans from the remote detection pipeline seems increasingly possible.

Future work should ensure that all methods work for all stakeholders. Such methods should therefore be evaluated across races, ethnicities, and other sensitive attributes to ensure fair and unbiased A.I. While some machine learning methods may help account for biased datasets, no technique matches the benefit of using balanced data. Fair and balanced healthcare A.I. initiatives must explicitly recruit participants in equal numbers across all demographics served to enable equitable services.

Acknowledgements

We thank all crowd workers who participated in the study. We also thank the clinicians who provided annotations for all study videos. The work was supported in part by funds to DPW from the National Institutes of Health (1R01EB025025-01, 1R01LM013364-01, 1R21HD091500-01, 1R01LM013083), the National Science Foundation (Award 2014232), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, Lucile Packard Foundation, Auxiliaries Endowment, the ISDB Transform Fund, the Weston Havens

Foundation, and program grants from Stanford's Human Centered Artificial Intelligence Program, Precision Health and Integrated Diagnostics Center, Beckman Center, Bio-X Center, Predictives and Diagnostics Accelerator, Spectrum, Spark Program in Translational Research, MediaX, 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. PW would like to acknowledge support from Mr. Schroeder and the Stanford Interdisciplinary Graduate Fellowship (SIGF) as the Schroeder Family Goldman Sachs Graduate Fellow.

References

[1]. Adler B Andrew, Minshawi Noha F, Erickson Craig A. Evolution of autism: from Kanner to the DSM-V. In: Handbook of early intervention for autism Spectrum disorders. New York, NY: Springer; 2014. p. 3–19.

[2]. Daniels Jena, Schwartz Jessey, Haber Nick, Voss Catalin, Kline Aaron, Fazel Azar, Washington Peter, et al. 5.13 Design and efficacy of a wearable device for social affective learning in children with autism. J Am Acad Child Adolesc Psychiatr 2017;56:S257–10.

[3]. Daniels Jena, Haber Nick, Voss Catalin, Schwartz Jessey, Tamura Serena, Fazel Azar, Kline Aaron, et al. Feasibility testing of a wearable behavioral aid for social learning in children with autism. Appl Clin Inf 2018;9(1):129.

[4]. Daniels Jena, Schwartz Jessey N, Voss Catalin, Haber Nick, Fazel Azar, Kline Aaron, Washington Peter, Feinstein Carl, Terry Winograd, Wall Dennis P. Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. NPJ Digit Med 2018;1(1):1–10. [PubMed: 31304287]

[5]. Duda Marlena, Daniels Jena, Wall Dennis P. Clinical evaluation of a novel and mobile autism risk assessment. J Autism Dev Disord 2016;46(6):1953–61. [PubMed: 26873142]

[6]. Duda M, Haber N, Daniels J, Wall DP. Crowdsourced validation of a machine-learning classification system for autism and ADHD. Transl Psychiatry 2017;7(5). e1133–e1133. [PubMed: 28509905]

[7]. Duda M, Kosmicki JA, Wall DP. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Transl Psychiatry 2015;5(4). e556–e556. [PubMed: 25918993]

[8]. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry 2016;6(2). e732–e732. [PubMed: 26859815]

[9]. Farnebäck Gunnar. Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. Berlin, Heidelberg: Springer; 2003. p. 363–70.

[10]. Fischbach Gerald D, Lord Catherine. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 2010;68(2):192–5. [PubMed: 20955926]

[11]. Geschwind Daniel H, Sowinski Janice, Lord Catherine, Iversen Portia, Shestack Jonathan, Jones Patrick, Lee Ducat, Spence Sarah J. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. Am J Hum Genet 2001;69(2):463–6. [PubMed: 11452364]

[12]. Guy William. ECDEU assessment manual for psychopharmacology. US department of health, education, and welfare, public health service, alcohol, drug abuse, and mental health administration. National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976.

[13]. Haber Nick, Voss Catalin, Wall Dennis. Making emotions transparent: google Glass helps autistic kids understand facial expressions through augmented-reaiity therapy. IEEE Spectr. 2020;57(4):46–52.

[14]. Haber Nick, Voss Catalin, Daniels Jena, Washington Peter, Fazel Azar, Kline Aaron, De Titas, Terry Winograd, Feinstein Carl, Wall Dennis P. A wearable social interaction aid for children with autism. 2020. arXiv preprint arXiv:2004.14281.

[15]. Haber Nick, Voss Catalin, Fazel Azar, Terry Winograd, Wall Dennis P. A practical approach to real-time neutral feature subtraction for facial expression recognition. In: IEEE winter conference on applications of computer vision (WACV). IEEE; 2016. p. 1–9. 2016.

[16]. Hall Dan, Huerta Michael F, McAuliffe Matthew J, Farber Gregory K. Sharing heterogeneous data: the national database for autism research. Neuroinformatics 2012;10(4):331–9. [PubMed: 22622767]

[17]. Howlin Patricia, Moore Anna. Diagnosis in autism: a survey of over 1200 patients in the UK. Autism 1997;1(2):135–62.

[18]. Hirth Matthias, Tobias Hoßfeld, Tran-Gia Phuoc. Anatomy of a crowdsourcing platform-using the example of microworkers. com. In: 2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing. IEEE; 2011. p. 322–9.

[19]. Hou Cathy, Haik Kalantarian, Peter Washington, Kaiti Dunlap, and Dennis Wall. “Development and Validation Of A Facial Emotion Classifier for Applications in the Treatment of Autism Spectrum Disorder.”.

[20]. Hyde Kayleigh K, Novack Marlena N, Lahaye Nicholas, Parlett-Pelleriti Chelsea, Anden Raymond, Dixon Dennis R, Linstead Erik. Applications of supervised machine learning in autism spectrum disorder research: a review. Rev J Autism Dev Disord. 2019;6(2):128–46.

[21]. Kalantarian Haik, Jedoui Khaled, Dunlap Kaitlyn, Schwartz Jessey, Washington Peter, Arman Husic, Tariq Qandeel, Ning Michael, Kline Aaron, Wall Dennis Paul. The performance of emotion classifiers for children with parent-reported autism: quantitative feasibility study. JMIR Ment Health 2020;7(4): e13174. [PubMed: 32234701]

[22]. Kalantarian Haik, Jedoui Khaled, Washington Peter, Tariq Qandeel, Dunlap Kaiti, Schwartz Jessey, Wall Dennis P. Labeling images with facial emotion and the potential for pediatric healthcare. Artif Intell Med 2019;98:77–86. [PubMed: 31521254]

[23]. Kalantarian Haik, Jedoui Khaled, Washington Peter, Wall Dennis P. A mobile game for automatic emotion-labeling of images. IEEE Trans Games 2018;12(2): 213–8. [PubMed: 32551410]

[24]. Kalantarian Haik, Washington Peter, Schwartz Jessey, Daniels Jena, Haber Nick, Wall Dennis P. Guess what? J Heathc Inf Res 2019;3(1):43–66.

[25]. Kalantarian Haik, Washington Peter, Schwartz Jessey, Daniels Jena, Haber Nick, Wall Dennis. A gamified mobile system for crowdsourcing video for autism research. In: IEEE international conference on healthcare informatics (ICHI). IEEE; 2018. p. 350–2. 2018.

[26]. Ke Shian-Ru, Hoang Le Uyen Thuc, Lee Yong-Jin, Hwang Jenq-Neng, Yoo Jang-Hee, Choi Kyoung-Ho. A review on video-based human activity recognition. Computers 2013;2(2):88–131.

[27]. Kline Aaron, Voss Catalin, Washington Peter, Haber Nick, Schwartz Hessey, Tariq Qandeel, Terry Winograd, Feinstein Carl, Wall Dennis P. Superpower glass. GetMobile: Mobile Comput Commun 2019;23(2):35–8.

[28]. Kogan Michael D, Vladutiu Catherine J, Schieve Laura A, Ghandour Reem M, Blumberg Stephen J, Zablotsky Benjamin, Perrin James M, et al. The prevalence of parent-reported autism spectrum disorder among US children. Pediatrics 2018; 142:6.

[29]. 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(2). e514–e514. [PubMed: 25710120]

[30]. Leblanc Emilie, Washington Peter, Varma Maya, Dunlap Kaitlyn, Penev Yordan, Kline Aaron, Wall Dennis P. Feature replacement methods enable reliable home video analysis for machine learning detection of autism. Sci Rep 2020;10(1):1–11. [PubMed: 31913322]

[31]. Levy Sebastien, Duda Marlena, Haber Nick, Wall Dennis P. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Mol Autism 2017;8(no. 1):1–17. [PubMed: 28070266]

[32]. Lord Catherine, Risi Susan, Lambrecht Linda, Cook Edwin H, Leventhal Bennett L, DiLavore Pamela C, Pickles Andrew, Rutter Michael. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 2000;30(3): 205–23. [PubMed: 11055457]

[33]. Mintz Mark. Evolution in the understanding of autism spectrum disorder: historical perspective. Indian J Pediatr 2017;84(1):44–52. [PubMed: 27053182]

[34]. Nag Anish, Haber Nick, Voss Catalin, Tamura Serena, Daniels Jena, Ma Jeffrey, Chiang Bryan, 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 2020;22(4):e13810. [PubMed: 32319961]

[35]. Ning Michael, Daniels Jena, Schwartz Jessey, Dunlap Kaitlyn, Washington Peter, Kalantarian Haik, Du Michael, Wall Dennis P. Identification and quantification of gaps in access to autism

resources in the United States: an infodemiological study. J Med Internet Res 2019;21:e13094. 7. [PubMed: 31293243]

[36]. Padilla-López, Ramón José, Andre Chaaraoui Alexandros, Flórez-Revuelta Francisco. Visual privacy protection methods: a survey. Expert Syst Appl 2015;42(9):4177–95.

[37]. Simons VIP Consortium. Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. Neuron 2012;73(6):1063–7. [PubMed: 22445335]

[38]. Tariq Qandeel, Daniels Jena, Nicole Schwartz Jessey, Washington Peter, Kalantarian Haik, Wall Dennis Paul. Mobile detection of autism through machine learning on home video: a development and prospective validation study. PLoS Med 2018;15:e1002705. 11. [PubMed: 30481180]

[39]. Tariq Qandeel, Scott Lanyon, Fleming Jessey, Schwartz Nicole, Dunlap Kaitlyn, Corbin Conor, Washington Peter, Kalantarian Haik, Khan Naila Z, Darmstadt Gary L, Wall Dennis Paul. 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. [PubMed: 31017583]

[40]. Turing Alan M Computing machinery and intelligence. In: Parsing the turing test. Dordrecht: Springer; 2009. p. 23–65.

[41]. Voss Catalin, Haber Nick, Wall Dennis P. The potential for machine learning–based wearables to improve socialization in teenagers and adults with autism spectrum disorder—reply. 11 JAMA Pediatr 2019;173. 1106–1106.

[42]. Voss Catalin, Haber Nick, Washington Peter, Kline Aaron, McCarthy Beth, Daniels Jena, Fazel Azar, et al. Designing a Holistic at-Home Learning Aid for Autism. 2020. arXiv preprint arXiv:2002.04263.

[43]. Voss Catalin, Schwartz Jessey, Daniels Jena, Kline Aaron, Haber Nick, Washington Peter, Tariq Qandeel, et al. 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. [PubMed: 30907929]

[44]. Voss Catalin, Washington Peter, Haber Nick, Kline Aaron, Daniels Jena, Fazel Azar, De Titas, et al. Superpower glass: delivering unobtrusive real-time social cues in wearable systems. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing; 2016. p. 1218–26. Adjunct.

[45]. Washington Peter, Kalantarian Haik, Kent Jack, Arman Husic, Kline Aaron, Leblanc Emilie, Hou Cathy, et al. Training an Emotion Detection Classifier Using Frames from a Mobile Therapeutic Game for Children with Developmental Disorders. 2020. arXiv preprint arXiv:2012.08678.

[46]. Washington Peter, Kalantarian Haik, Tariq Qandeel, Schwartz Jessey, Dunlap Kaitlyn, Chrisman Brianna, Varma Maya, et al. Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. J Med Internet Res 2019;21(5):e13668. [PubMed: 31124463]

[47]. Washington Peter, Kline Aaron, Cezmi Mutlu Onur, Leblanc Emilie, Hou Cathy, Stockham Nate, Paskov Kelley, Chrisman Brianna, Wall Dennis P. Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging. 2021. arXiv preprint arXiv:2101.03478.

[48]. Washington Peter, Leblanc Emilie, Dunlap Kaitlyn, Penev Yordan, Kline Aaron, Paskov Kelley, Sun Min Woo, et al. Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for video-based autism feature recognition. J Personalized Med 2020;10(3):86.

[49]. Washington Peter, Leblanc Emilie, Dunlap Kaitlyn, Penev Yordan, Varma Maya, Jung Jae-Yoon, Chrisman Brianna, et al. Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. In: Biocomputing 2021: proceedings of the Pacific symposium; 2020. p. 14–25.

[50]. Washington Peter, Cezmi Mutlu Onur, Leblanc Emilie, Kline Aaron, Hou Cathy, Chrisman Brianna, Stockham Nate, et al. Using Crowdsourcing to Train Facial Emotion Machine Learning Models with Ambiguous Labels. 2021. arXiv preprint arXiv:2101.03477.

[51]. Washington Peter, Park Natalie, Srivastava Parishkrita, Voss Catalin, Kline Aaron, Varma Maya, et al. Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. Biol Psychiatr: Cognit Neurosci Neuroimaging 2019;5(8).

[52]. Washington Peter, Marie Paskov Kelley, Kalantarian Haik, Stockham Nathaniel, Voss Catalin, Kline Aaron, Patnaik Ritik, et al. Feature selection and dimension reduction of social autism data. In: Pacific symposium ON biocomputing 2020; 2019. p. 707–18.

[53]. Washington Peter, Tariq Qandeel, Leblanc Emilie, Chrisman Brianna, Dunlap Kaitlyn, Kline Aaron, Kalantarian Haik, et al. Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Sci Rep 2021;11(1):1–11. [PubMed: 33414495]

[54]. Washington Peter, Tariq Qandeel, Leblanc Emilie, Chrisman Brianna, Dunlap Kaitlyn, Kline Aaron, Kalantarian Haik, et al. Crowdsourced feature tagging for scalable and privacy-preserved autism diagnosis. medRxiv; 2020.

[55]. Washington Peter, Voss Catalin, Kline Aaron, Haber Nick, Daniels Jena, Fazel Azar, De Titas, Feinstein Carl, Terry Winograd, Wall Dennis. Superpowerglass: a wearable aid for the at-home therapy of children with autism. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 1; 2017. p. 1–22. 3.

[56]. Washington Peter, Voss Catalin, Haber Nick, Tanaka Serena, Daniels Jena, Feinstein Carl, Terry Winograd, Wall Dennis. A wearable social interaction aid for children with autism. In: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems; 2016. p. 2348–54.

[57]. Washington Peter, Yeung Serena, Percha Bethany, Tatonetti Nicholas, Jan Liphardt, Wall Dennis P. Achieving trustworthy biomedical data solutions. In: Biocomputing 2021: proceedings of the pacific symposium; 2020. p. 1–13.

[58]. Wall Dennis Paul, Kosmicki J, Deluca TF, Harstad E, Alfred Fusaro Vincent. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry 2012;2:4. e100–e100. [PubMed: 22832900]

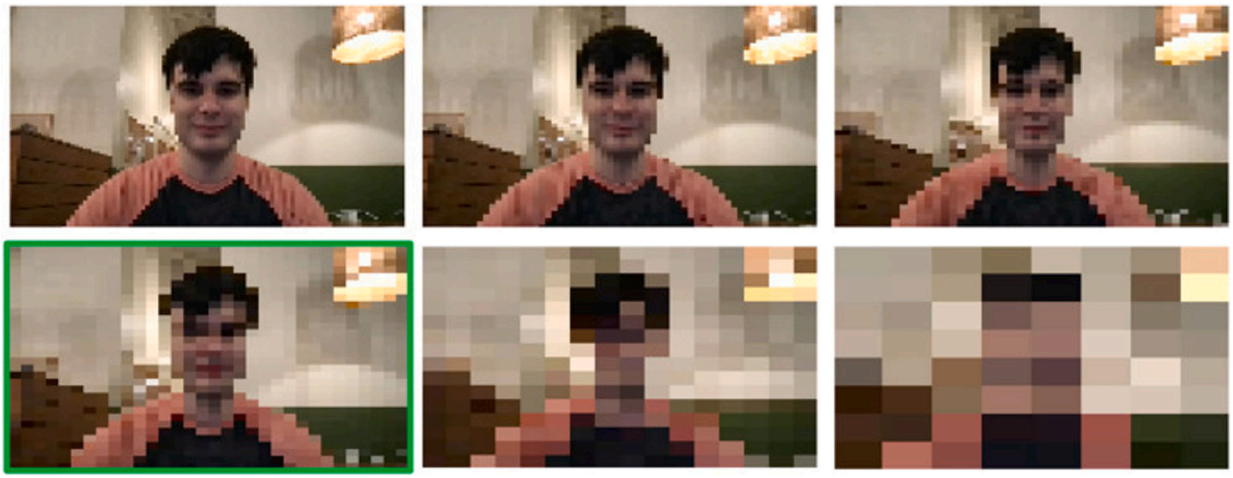

Fig. 1.

Six intensities of pixelation used in the study. The bottom left image (highlighted in green) depicts the intensity level used for the primary portion of the study. The other intensities are used for a secondary analysis comparing the effect of pixelation intensity on annotation quality.

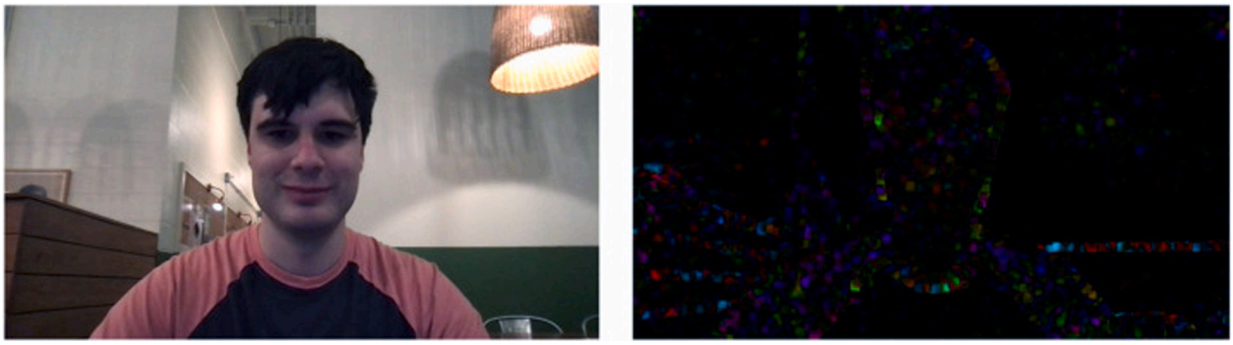

Fig. 2.
Dense optical flow was evaluated as a drastic privacy alteration as depicted here. Original frame is on the left.

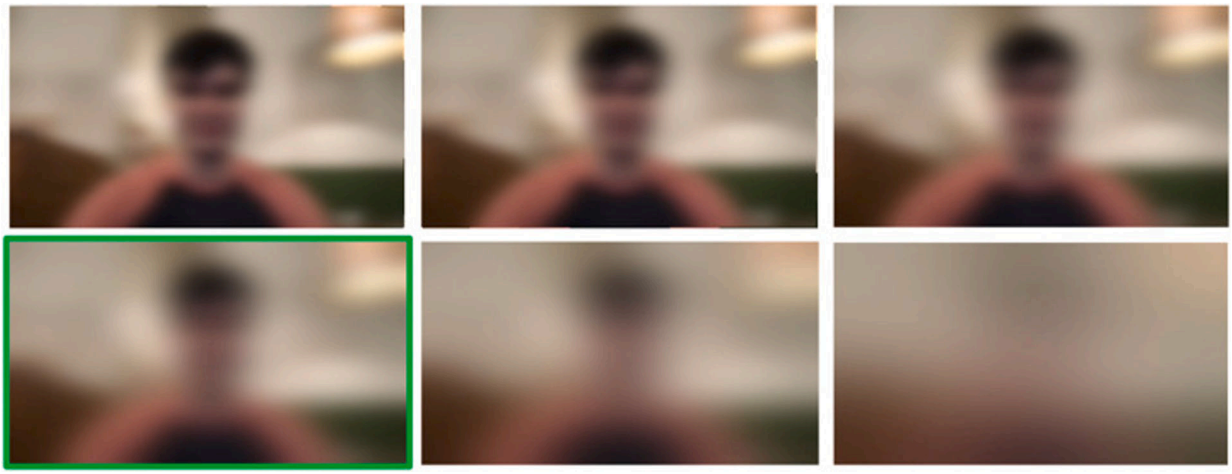

Fig. 3.

Six intensities of Gaussian blurring used in the study. The most bottom left image (highlighted in green) depicts the intensity level which was used for the primary portion of the study. The other intensities are used for a secondary analysis comparing the effect of blurring intensity on annotation quality.

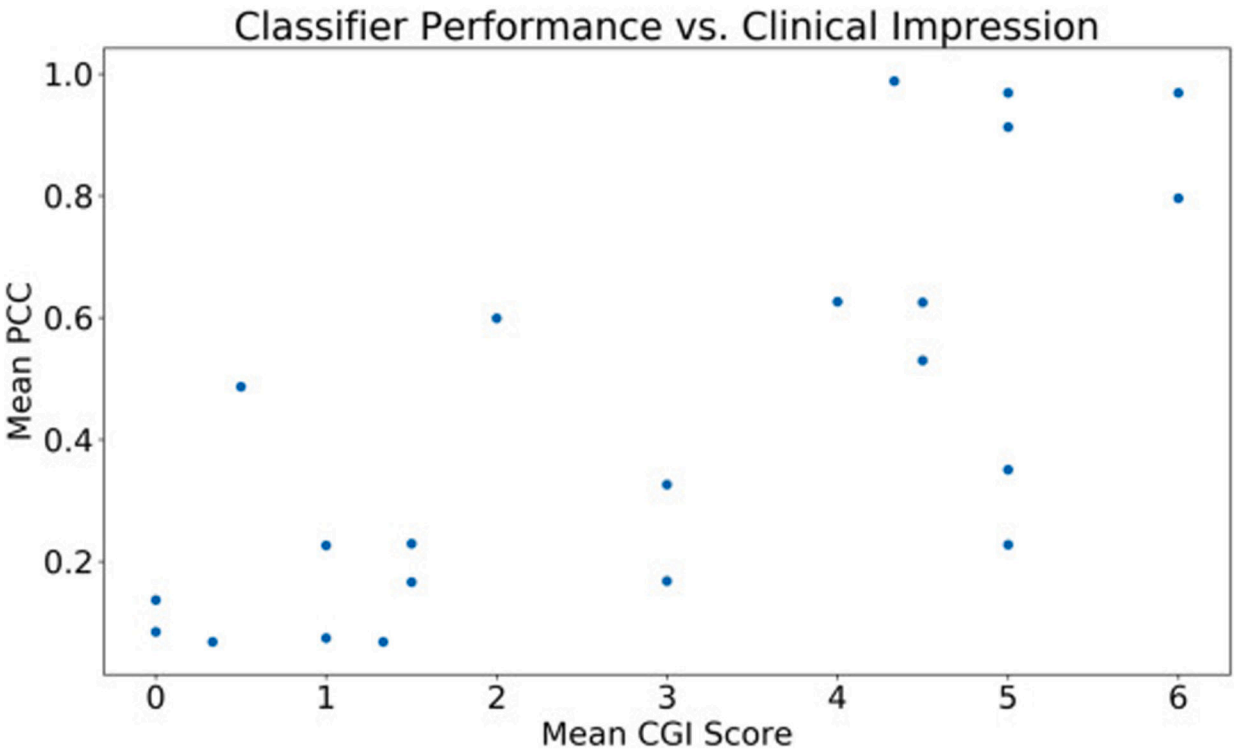

Fig. 4.

There is a clear linear correlation ($r = 0.75$, $p < 0.001$) between the mean Clinical Global Impression (CGI) score provided by professional clinicians for each video and the corresponding classifier score emitted by the logistic regression classifier with crowd inputs.

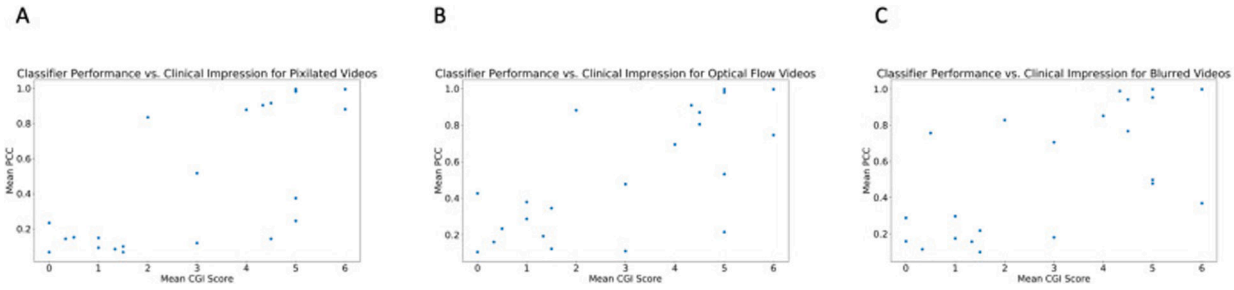

Fig. 5.

The linear correlation between the mean Clinical Global Impression (CGI) score provided by professional clinicians for each video and the corresponding classifier score emitted by the logistic regression classifier with crowd inputs is maintained with privacy-preserving video modifications. The correlation is weaker for Gaussian blurring ($r = 0.64$, $p = 0.001$ for Gaussian blurring) than for dense optical flow and pixelation ($r = 0.71$, $p = 0.0002$ for both).

Table 1

The effect of increasing levels of blurring intensity on mean performance of a separate testing set from the primary evaluation.

Blurring Kernel Size (Relative to Original)	Mean Probability of the Correct Class	Mean Accuracy	Mean Precision	Mean Recall	Mean Specificity	Mean AUROC	Mean AUPRC
1/6th	0.766 ± 0.253	92.9	87.5	100.0	85.7	73.5	70.8
1/5th	0.768 ± 0.238	78.6	75.0	85.7	71.4	77.6	83.3
1/4th	0.842 ± 0.151	85.7	85.7	85.7	85.7	63.3	77.5
1/3rd	0.670 ± 0.313	78.6	83.3	71.4	85.7	44.9	60.1
½	0.740 ± 0.261	92.9	87.5	100.0	85.7	63.3	73.3
Full Image	0.690 ± 0.316	92.9	87.5	100.0	85.7	79.6	83.7

Table 2

The effect of increasing levels of pixelation intensity on mean performance of a separate testing set from the primary evaluation.

Pixelation Intermediate Frame Size	Mean Probability of the Correct Class	Mean Accuracy	Mean Precision	Mean Recall	Mean Specificity	Mean AUROC	Mean AUPRC
96 × 96	0.809 ± 0.190	90.9	100.0	85.7	100.0	42.9	70.1
64 × 64	0.741 ± 0.230	92.3	100.0	85.7	100.0	52.4	66.9
48 × 48	0.788 ± 0.216	91.7	100.0	85.7	100.0	65.7	79.8
32 × 32	0.781 ± 0.210	75.0	83.3	71.4	80.0	45.7	67.2
16 × 16	0.814 ± 0.179	83.3	85.7	85.7	80.0	71.4	85.1
8 × 8	0.802 ± 0.189	75.5	83.3	71.4	80.0	42.9	66.1

Table 3

The mean absolute deviation for each privacy condition from the baseline condition answers for the behaviors used as inputs to the autism classifier. This difference provides a measure of the privacy condition's effect on annotation quality.

	Mean Deviation for Pixelation	Mean Deviation for Dense Optical Flow	Mean Deviation for Gaussian Blurring
Abnormal Speech	0.28	0.32	0.29
Echolalia	0.39	0.45	0.47
Repetitive or Odd Language	0.25	0.30	0.26
Expressive Language and Conversation	0.29	0.41	0.33
Eye Contact	0.29	0.37	0.33
Facial Expressiveness	0.25	0.32	0.29
Social Interaction Initiation	0.26	0.30	0.31
Shares Excitement	0.34	0.32	0.37
Aggressive Behavior	0.09	0.12	0.12