Deep Neural Network Reveals the World of Autism From a First-Person Perspective

Mindi Ruan, Paula J. Webster, Xin Li, and Shuo Wang

People with autism spectrum disorder (ASD) show atypical attention to social stimuli and aberrant gaze when viewing images of the physical world. However, it is unknown how they perceive the world from a first-person perspective. In this study, we used machine learning to classify photos taken in three different categories (people, indoors, and outdoors) as either having been taken by individuals with ASD or by peers without ASD. Our classifier effectively discriminated photos from all three categories, but was particularly successful at classifying photos of people with >80% accuracy. Importantly, visualization of our model revealed critical features that led to successful discrimination and showed that our model adopted a strategy similar to that of ASD experts. Furthermore, for the first time we showed that photos taken by individuals with ASD contained less salient objects, especially in the central visual field. Notably, our model outperformed classification of these photos by ASD experts. Together, we demonstrate an effective and novel method that is capable of discerning photos taken by individuals with ASD and revealing aberrant visual attention in ASD from a unique first-person perspective. Our method may in turn provide an objective measure for evaluations of individuals with ASD. ***Autism Res*** *2020, 00: 1–10.* © 2020 International Society for Autism Research and Wiley Periodicals LLC

Lay Summary: People with autism spectrum disorder (ASD) demonstrate atypical visual attention to social stimuli. However, it remains largely unclear how they perceive the world from a first-person perspective. In this study, we employed a deep learning approach to analyze a unique dataset of photos taken by people with and without ASD. Our computer modeling was not only able to discern which photos were taken by individuals with ASD, outperforming ASD experts, but importantly, it revealed critical features that led to successful discrimination, revealing aspects of atypical visual attention in ASD from their first-person perspective.

Keywords: autism spectrum disorder; deep neural network; photos; faces; attention; saliency; artificial intelligence

Introduction

The ability to attend to what is important in the environment is one of the most fundamental cognitive functions in humans. However, people with autism spectrum disorder (ASD) show profound impairments in attention, especially to social stimuli such as human faces and social scenes [Wang & Adolphs, 2017; Wang et al., 2015; Wang et al., 2014]. Prior studies have documented that individuals without ASD spend significantly more time than peers with ASD looking at the eyes when viewing human faces presented in movie clips [Klin, Jones, Schultz, Volkmar, & Cohen, 2002] or photographs [Pelphrey et al., 2002]. When comparing social versus nonsocial stimuli, people with ASD show reduced attention to human faces and to other social stimuli such as the human voice and hand gestures; however, they pay more attention to nonsocial objects [Dawson, Webb, & McPartland, 2005; Sasson, Elison, Turner-Brown, Dichter, & Bodfish, 2011], notably including gadgets, devices, vehicles, electronics, and other objects of idiosyncratic "special interest" [Kanner, 1943].

In order to better understand the atypical social behavior in ASD, there is an increasing trend to employ more natural and ecologically valid stimuli (e.g., complex scenes taken with a natural background) [Ames & Fletcher-Watson, 2010; Birmingham, Cerf, & Adolphs, 2011; Byrge, Dubois, Tyszka, Adolphs, & Kennedy, 2015; Chawarska, Macari, & Shic, 2013; Rice, Moriuchi, Jones, & Klin, 2012; Santos et al., 2012; Shic, Bradshaw, Klin, Scassellati, & Chawarska, 2011; Wang et al., 2015] and to test participants in a more natural setting as opposed to a

From the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, West Virginia (M.R., X.L.); Department of Chemical and Biomedical Engineering and Rockefeller Neuroscience Institute, West Virginia University, Morgantown, West Virginia (P.J.W., S.W.)

Received May 14, 2020; accepted for publication July 27, 2020

Address for correspondence and reprints: Xin Li, Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506. E-mail: xin.li@mail.wvu.edu

Shuo Wang, Department of Chemical and Biomedical Engineering and Rockefeller Neuroscience Institute, West Virginia University, Morgantown, WV 26506. E-mail: wangshuo45@gmail.com

Published online 00 Month 2020 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/aur.2376

© 2020 International Society for Autism Research and Wiley Periodicals LLC

restricted clinical setting in which core ASD behaviors may not be seen. Tasks presenting faces in a naturalistic setting demonstrate that people with ASD have reduced attention to faces and specifically to the eye region [Freeth, Chapman, Ropar, & Mitchell, 2010; Klin et al., 2002; Norbury et al., 2009; Riby, Hancock, Jones, & Hanley, 2013; Riby & Hancock, 2009]. In particular, recent studies have directly tested people with ASD during natural interactions with lab personnel [Marinoiu, Zanfir, Olaru, & Sminchisescu, 2018; Rehg et al., 2013; Wang et al., 2016]. For example, we asked ASD participants and controls to take photos in natural social settings, and showed that while those with ASD take more photos of people than controls, those photos are more often not front-facing and/or are taken from odd perspectives indicating a lack of social engagement [Wang et al., 2016].

Recent studies have been employing machine learning techniques to quantitatively characterize atypical behavior in ASD. In addition to its benefits of improving and streamlining ASD diagnosing [Duda, Kosmicki, & Wall, 2014; Tariq et al., 2018; Wall, Kosmicki, DeLuca, Harstad, & Fusaro, 2012], machine learning can reveal critical features of atypical behavior in ASD from various domains of data, including eye movement [Jiang & Zhao, 2017; Wang et al., 2015], scoring of Autism Diagnostic Interview-Revised (ADI-R) [Wall et al., 2012], scoring of Autism Diagnostic Observation Schedule (ADOS) [Duda et al., 2014], and home videos [Tariq et al., 2018]. Therefore, machine learning can provide an effective and crucial way for us to identify and understand the factors that contribute to atypical behavior in ASD.

The current study explored the feasibility of using a deep neural network (DNN) to discern between photos taken by participants with ASD versus those taken by controls. These photos provide a unique first-person perspective of what is visually salient to the photographer and reflect their social interactions with their environment not seen in other studies in which those with ASD are asked to view stock photos. Therefore, by classifying these photos, we were able to reveal aspects of aberrant visual attention in individuals with ASD. Indeed, our deep neural network could effectively classify whether a photo was taken by an individual with ASD or by a peer without ASD. Surprisingly, our machine-based approach consistently outperformed human-based classification ratings conducted by ASD experts. Notably, our explainable artificial intelligence (XAI) technique revealed the critical features that support the classification. Together, we showed that photos taken from a first-person perspective by those with ASD can aid in understanding their unique visual perspective of the world and that deep neural networks may provide an efficient and objective method to aid in the analysis of visual attention deficits in ASD.

Methods and Materials

Participants

All participants were from our previous report [Wang et al., 2016]. Briefly, 16 high-functioning participants with ASD (12 male) and 21 controls (18 male) were recruited (Supplementary Table S1). All ASD participants met DSM-5/ICD-10 diagnostic criteria for ASD, and all met the cutoff scores for ASD on the ADOS-2 revised scoring system for Module 4 [Hus & Lord, 2014], and the ADI-R [LeCouteur, Rutter, & Lord, 1989; Lord et al., 1994] or Social Communication Questionnaire (SCQ) [Rutter et al., 2003]. The ASD group had a full-scale IQ (FSIQ) of 111.6 ± 12.2 (mean ± SD, from the Wechsler Abbreviated Scale of Intelligence-II), a mean age of 29.7 ± 11.2 years, and a mean Autism Quotient (AQ) of 29.7 ± 8.07. Controls had a comparable FSIQ of 111.0 ± 9.90 (*t*-test, $P = 0.92$, although IQ was only available on a subset) and a comparable mean age of 33.0 ± 9.31 years (*t*-test, $P = 0.33$). The groups were also matched for gender, race, and education (Supplementary Table S1). Participants provided written informed consent according to protocols approved by the institutional review board of the California Institute of Technology (Caltech), and all methods were carried out in accordance with the approved guidelines.

Task

Participants were provided with a camera and instructed to take photos of anything they wanted, such as objects, rooms, scenery, or people, and they could take as many photos as they wished. There were three blocked conditions (in counterbalanced order between participants):

1. People Block. Photographing for this block took place in the rooms and hallway of a Caltech laboratory. Participants were instructed to primarily take photos of two lab members, who were fully aware of the experiment and thus were prepared to pose or be expressive. Some participants with ASD were also instructed to take self-portraits. Participants were free to set up the space however they liked (e.g., they could move around the room or interact with the objects in the room) and they could also ask the two lab members to move or pose to their instruction.
2. Indoor Block. Photographing took place in the same indoor environment and participants were instructed to walk around the lab and feel free to enter lab spaces to photograph objects.
3. Outdoor Block. Photographing took place on the Caltech campus outside of the building. Participants were instructed to walk anywhere on campus if they wished and take photos of any objects or people of their own choosing.

During each condition, participants were asked to take at least ten photos. Five participants with ASD completed two sessions of the experiment and we have pooled photos from both sessions for each participant for analysis.

Rating by ASD Experts

ASD experts skilled in ASD evaluations using the ADOS rated all photos independently on a 9 point scale (1 = the photo was definitely taken by a person *with* ASD to 9 = the photo was definitely taken by a person *without* ASD; see [Wang et al., 2016] for detailed instructions for photo ratings). Photos, especially those taken by participants with ASD of themselves, were excluded from rating if an ASD expert could recognize the identity of the participant so that all raters remained blinded to which group (ASD or control) the participant belonged. All photos were shown in randomized order within each of the three photo task conditions (People, Indoors, Outdoors). Notably, these professional raters were highly consistent in their ratings [Wang et al., 2016].

Classification

We have adopted a strategy of fine-tuning a pretrained VGG16 convolutional neural network (CNN) on ImageNet to discriminate photos taken by participants with ASD from those taken by controls (Fig. 1A). This CNN is capable of recognizing a large class of objects [Simonyan & Zisserman, 2014], including faces, indoor objects, and outdoor objects, and is thus suitable for our photo stimuli. The CNN consisted of a feature extraction section (13 convolutional layers) and a classification section (three fully connected [FC] layers). The feature extraction section was consistent with the typical architecture of a CNN. We applied a 3×3 filter with 1-pixel padding and 1-pixel stride to each convolutional layer, which followed by a Batch Normalization (BatchNorm) and Rectified Linear Unit (ReLU) operation. Some of the convolutional layers were followed by five 2×2 max-pool operations with a stride of 2. There were three FC layers in each classification section: the first two had 4096 channels each, and the third performed a two-way ASD classification and thus contained two channels. Each FC layer was followed by a ReLU and 50% dropout to avoid overfitting. A nonlinear Softmax operation was applied to the final output of VGG16 network to make the binary classification prediction. It is worth noting that classification was performed and fine-tuned for each task condition separately. Our analysis was carried out in three main steps:

1. Data preparation. There was a total of 1672 photos in our dataset (Supplementary Table S1). The People Block contained 490 photos from participants with ASD and 217 photos from controls. The Indoor Block contained 265 photos from participants with ASD and 229 photos from controls. The Outdoor Block contained 229 photos from participants with ASD and 242 photos from controls. Photos from all participants were pooled for training and testing. To augment the dataset, before training, the input images were randomly cropped to a random size and then rescaled to 224×224 RGB images, meanwhile the images had a 0.5 probability to be horizontally flipped. In each training/testing run (separately for each task condition), the dataset was randomly split into three parts: 60% served as the training set, 20% served as the validation set, and 20% served as the test set. We repeated the procedure 10 times with different random splits of training and testing data. Additional different splits of training and testing data were also tested and we derived qualitatively the same performance (see Results).
2. Training the DNN. Our VGG16 network ran on the deep learning framework of PyTorch [Paszke et al., 2017; Subramanian, 2018]. To improve model performance with our small dataset, we have applied Transfer Learning to our model. For the feature extraction part, we loaded the pretrained weights on ImageNet and froze the convolutional layers to prevent their weights from updating during training. With a better feature extraction, our dataset was mainly used to train the FC layers to improve its ability of classification. Training was performed by the Stochastic Gradient Descent (SGD) optimizer with the base learning rate of 10^{-3}.
3. Permutation test. To further confirm the results, statistical significance was estimated by permutation test. There were ten runs, and in each run, photo labels were randomly shuffled and the training/testing procedure was repeated. *P*-values were calculated by comparing the observed accuracy to that from the permutation test.

Receiver Operating Characteristic Curves

We used receiver operating characteristic (ROC) curves to evaluate and compare classification performance. We constructed two kinds of ROC curves, respectively calculated from professional rating scores and our model prediction output, as a comparison. For the rating scores, three independent ASD experts familiar with the clinical presentation of ASD and research reliable on the ADOS-2 were asked to evaluate and score every photo from our dataset. The score represents how confident they are that the photo was taken by a participant with ASD or a control (scores from 1 to 9; 1 = confident the photo was taken by a participant with ASD, 9 = confident the photo

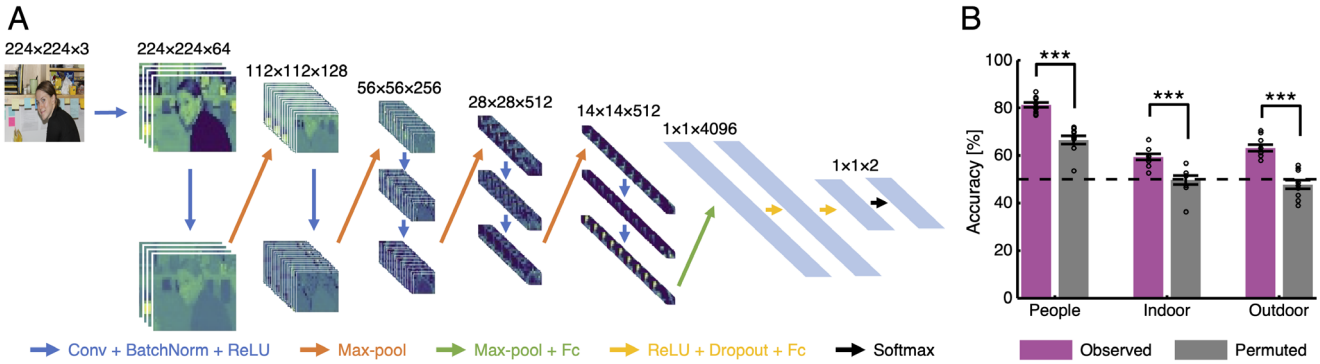

Figure 1. Model architecture and performance. (A) Model architecture. The input was a fixed-size 224 × 224 RGB image. The image was passed through a stack of convolutional layers, where the filters were used with a very small receptive field of 3 × 3. The convolution stride was fixed to 1 pixel; the spatial padding of convolutional layer input was such that the spatial resolution was preserved after convolution (i.e., the padding was 1-pixel for 3 × 3 convolutional layers). Spatial pooling was carried out by five max-pooling layers, which followed some of the convolutional layers. Max-pooling was performed over a 2 × 2 pixel window, with stride 2. Three fully connected (FC) layers followed a stack of convolutional layers: the first two had 4096 channels each, the third performed a two-way ASD classification and thus contained two channels. The final layer was the Softmax layer. All hidden layers were equipped with the rectification (ReLU) nonlinearity. (B) Model prediction accuracy. Our model could differentiate photos taken by those with ASD from those taken by controls in all conditions. Error bars denote ±SEM across runs and circles show individual values. Asterisks indicate significant difference in prediction accuracy between observed (magenta) versus permuted (gray) runs using unpaired *t*-test: ***: $P < 0.001$.

was taken by a control participant). Thus, curves of human performance were constructed based on the experts' average scores for each photo.

For our model, following the last layer, the Softmax operation output the probability distribution of the predictions as positive (a participant with ASD took the photo) or negative (a control participant took the photo). The prediction probabilities were used to construct the ROC curve. To reduce bias in the dataset, we tested our model 10 times with ten different splits of training and testing data; but every shuffle of the training, validation, and test datasets was based on the same split ratio. The area under the curve (AUC) of the ROC was calculated by integrating the area under the ROC curve (trapezoid rule). Note that since ROC is a probability curve, AUC indicates the degree or measure of separability (i.e., tells how well the model is capable of distinguishing between classes such as ASD versus control).

Saliency Analysis

To detect salient objects in the photos, we applied the most recent saliency detection algorithm to extract the saliency map from an input photo [Hou et al., 2017]. This algorithm applies short connections to the skip-layer structures within a Holistically-Nested Edge Detector (HED-SC). By taking full advantage of multilevel and multiscale features extracted from fully convolutional neural networks, HED-SC can offer fine-granularity representations at each layer leading to state-of-the-art saliency detection performance. We calculated the average saliency values for two regions. The central region consisted of a rectangle located at the image center and sized by 1/3 width × 1/3 height of the image; and the peripheral region consisted of the rest of the image.

Results

Model Performance

We designed a DNN that effectively discriminated photos that were taken by participants with ASD from photos taken by controls (Fig. 1A). The model reached an accuracy of 81.3% ± 3.34% (mean ± SD across runs) for the People Block, 59.4% ± 3.95% for the Indoor Block, and 63.2% ± 4.30% for the Outdoor Block (Fig. 1B). The DNN performance was above chance for all conditions, including the People Block (permuted: 66.5% ± 5.49%; unpaired two-tailed *t*-test between observed versus permuted performance: $t(18) = 7.27$, $P = 9.33 \times 10^{-7}$), the Indoor Block (permuted: 49.7% ± 5.93%; $t(18) = 4.30$, $P = 4.32 \times 10^{-4}$), as well as the Outdoor Block (permuted: 47.8% ± 5.73%; $t(18) = 6.79$, $P = 2.34 \times 10^{-6}$; Fig. 1B), demonstrating that the VGG model could be applied to successfully discriminate all categories of photos as having been taken by people with ASD, but photos of people were the most discriminative. Notably, our model performance still held when we excluded all self-portraits from the People Block (80.8% ± 3.13%; $P = 1.13 \times 10^{-6}$), suggesting that our classification was not simply driven by self-portraits from participants with ASD (note that only participants with ASD took self-portraits). Our results were also consistent with our prior published report from ASD experts experienced in ADOS administration demonstrating that photos from the People Block

are most discriminative between participants with ASD and controls. It is worth noting that ASD experts did not successfully discriminate photos from the Outdoor Block as having been taken by those with versus without ASD [Wang et al., 2016].

We conducted the following control analyses to confirm our model performance. (1) We employed a tenfold cross-validation that derived 83.7% ± 3.87% accuracy for the People Block, 64.6% ± 6.60% accuracy for the Indoor Block, and 59.3% ± 5.92% accuracy for the Outdoor Block. (2) We further tested model bias by randomly assigning a label to each photo (i.e., each photo had a 50% probability to be labeled as ASD or control). We derived a chance performance (i.e., 50% accuracy) for all three conditions: we derived 47.3% ± 14.3% accuracy for the People Block (paired *t*-test against 50%: $P = 0.57$), 50.8% ± 6.58% accuracy for the Indoor Block ($P = 0.70$), and 47.9% ± 6.57% accuracy for the Outdoor Block ($P = 0.33$), suggesting that our model was not biased toward reporting one category of photos. (3) Consistent with (2), we derived similar results (People Block: accuracy = 77.0% ± 5.58%, AUC = 0.85 ± 0.052; Indoor Block: accuracy = 63.3% ± 3.50%, AUC = 0.68 ± 0.052; Outdoor Block: accuracy = 64.2% ± 5.24%, AUC = 0.71 ± 0.048) when we used an equal number of photos from participants with ASD and controls, suggesting that the results could not simply be attributed to more photos having been taken by participants with ASD. (4) Although blurred photos are a major characteristic indicating ASD [Wang et al., 2016] (also see below in Fig. 2), when we excluded all blurred photos from the analysis, we still derived similar results (People Block: accuracy = 81.3% ± 4.50%, AUC = 0.91 ± 0.030; Indoor Block: accuracy = 64.7% ± .73%, AUC = 0.69 ± 0.060; Outdoor Block: accuracy = 61.0% ± 5.77%, AUC = 0.68 ± 0.062). (5) We conducted a leave-one-participant-out analysis by training the classifier with photos from all but one participant and testing on the remaining participant. We found an above-chance performance (People Block: accuracy = 61.7% ± 27.9% [mean ± SD across participants/validations], AUC = 0.72; Indoor Block: accuracy = 55.8% ± 25.0%, AUC = 0.59; Outdoor Block: accuracy = 59.7% ± 23.2%, AUC = 0.60), suggesting that photos from different participants within a group shared similar features and our model could generalize to predict whether a new photo was taken by a control participant or a participant in the ASD group. The People Block still showed the best performance, likely because the way that participants from each group composed the photos was more consistent. Note that AUC could only be assessed if we pooled photos from different participants because there was only one label for each participant; and we derived an accuracy of 67.0% for the People Block, an accuracy of 56.8% for the Indoor Block, and an accuracy of 58.6% for the Outdoor Block, if we pooled all photos to assess prediction performance. We next explored the factors that led to correct classification.

Model Explanation Through Layer-Wise Relevance Propagation

To provide an explanation of our model's output in the domain of its input, we applied layer-wise relevance propagation (LRP) to our trained classifier. LRP can use the network weights created by the forward-pass to propagate the output back through the network up until the

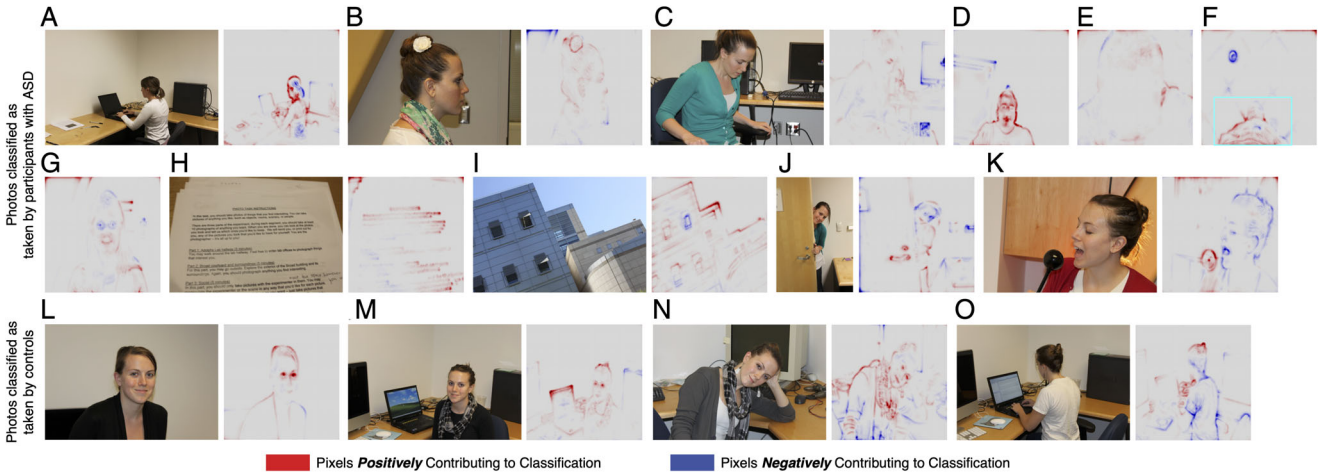

Figure 2. Layer-wise Relevance Propogation (LRP) explanation. (A–K) Photos that were classified as having been taken by participants with ASD. (L–O) Photos that were classified as having been taken by controls. Red pixels positively contributed to the classification whereas blue pixels negatively contributed to the classification. (J, K) Photos that were incorrectly classified as having been taken by participants with ASD. (N, O) Photos that were incorrectly classified as having been taken by controls. The cyan box in (F) illustrates where the subject of the photo was located.

original input image. The explanation given by LRP is a heatmap of which pixels in the original image contribute to the final output (Fig. 2; red pixels *positively* contributed to the classification whereas blue pixels *negatively* contributed to the classification). In the People Block, we found that classification of a photo as having been taken by participants with ASD was driven by the following factors: (1) photos had a view of the subject's back (Fig. 2A) or side (Fig. 2B); (2) subjects in the photos did not pose or look at the camera (Fig. 2A–C); (3) subjects in the photos were not expressive (Fig. 2B,C); (4) photos had an odd visual perspective (Fig. 2D–F; the cyan box in Fig. 2F denotes where the subject was located); and (5) photos were blurred (Fig. 2G; also note that the eyes in this photo looked at the camera and negatively contributed to classifying this photo as taken by participants with ASD [as shown in blue]). In contrast, photos with rich facial expressions and a regular angle of view (Fig. 2L,M) would lead to our classifier identifying them as having been taken by controls. Similarly, in the Indoor Block and Outdoor Block, photos that were blurred (Fig. 2H) or slanted (Fig. 2I) were classified as having been taken by participants with ASD. These features derived from machine learning were consistent with intuition from ASD experts from our prior study [Wang et al., 2016], suggesting that both approaches (human ratings and DNN) adopted a similar strategy in discriminating the photos.

Notably, in the photos that were mistakenly classified as being taken by participants with ASD (Fig. 2J,K), front view of the face (Fig. 2J) and facial expressions (Fig. 2K) still contributed to classification of the photos as having been taken by controls (i.e., negatively contributed to classifying the photos as from participants with ASD). On the other hand, in the photos that were mistakenly classified as being taken by controls (Fig. 2N,O), front view of the face and rich facial expressions (Fig. 2N) led to such classification whereas the view of the subject's back (Fig. 2O) still contributed to classification of the photos as having been taken by participants with ASD (i.e., negatively contributed to classifying the photos as from controls). Therefore, these results suggest that our classifier adopted a consistent strategy in classifying the photos and the incorrectly classified photos might be driven by other factors.

Saliency Analysis of Photos

Since participants with ASD demonstrate atypical visual saliency [Wang et al., 2015], we employed a saliency model to detect salient objects in the photos (see Methods). We found that in the People Block (Fig. 3A–C), photos taken by participants with ASD had similar saliency values compared to those taken by controls for both the central region (ASD: 149 ± 59.8 [mean ± SD], controls: 145 ± 65.2; two-tailed unpaired *t*-test: $t(705)$ =

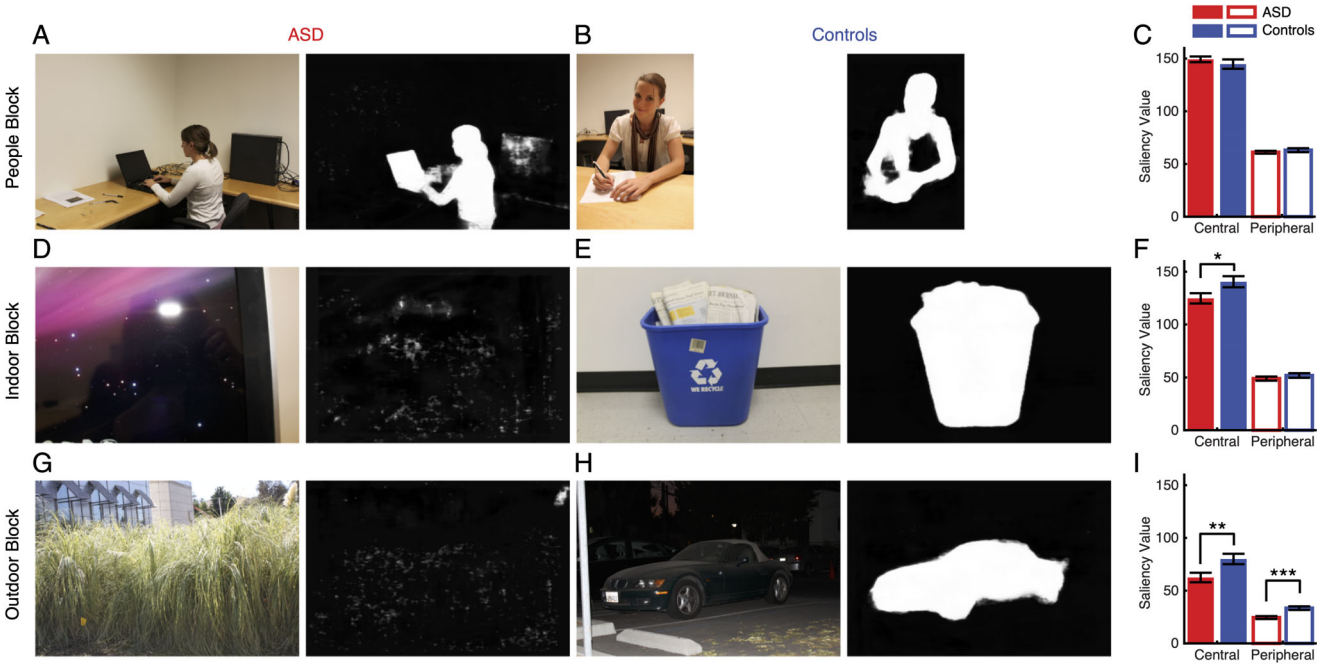

Figure 3. Saliency analysis. (A–C) People Block. (D–F) Indoor Block. (G–I) Outdoor Block. (A, B, D, E, G, H) Example photos (left) with detected saliency maps (right). (A, D, G) Photos taken by participants with ASD. (B, E, H) Photos taken by controls. (C, F, I) Average saliency value. Error bars denote ±SEM across photos. Solid bars denote the central region and open bars denote the peripheral region. Asterisks indicate significant difference between ASD and controls using two-tailed unpaired *t*-test: *: $P < 0.05$, **: $P < 0.01$, and ***: $P < 0.001$. Red: ASD. Blue: controls.

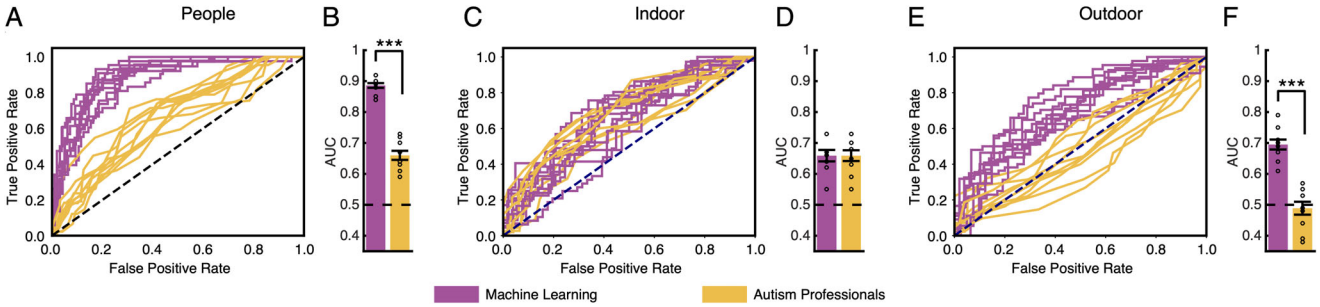

Figure 4. Our classifier model outperformed human ASD experts. (A, B) People Block. (C, D) Indoor Block. (E, F) Outdoor Block. (A, C, E) ROC curves. (B, D, F) Area under the ROC curve (AUC values). Error bars denote ±SEM across runs and circles show individual values. Asterisks indicate significant difference in performance between machine learning (purple) and ASD professionals (yellow) using paired *t*-test: ***: $P < 0.001$.

0.90, $P = 0.37$, effect size in Hedges' g (standardized mean difference): $g = 0.073$, permutation $P = 0.37$) and the peripheral region (ASD: 61.1 ± 29.4, controls: 63.1 ± 27.6; $t(705) = 0.82$, $P = 0.41$, $g = 0.067$, permutation $P = 0.37$). In the Indoor Block (Fig. 3D–F), photos taken by participants with ASD were less salient in the central region (ASD: 125 ± 79.2, controls: 141 ± 79.5; $t(492) = 2.20$, $P = 0.028$, $g = 0.20$, permutation $P = 0.036$) but not in the peripheral region (ASD: 48.9 ± 31.7, controls: 51.9 ± 33.1; $t(492) = 1.00$, $P = 0.32$, $g = 0.091$, permutation $P = 0.32$). Notably, in the Outdoor Block (Fig. 3G–I), photos taken by participants with ASD were less salient in both the central region (ASD: 62.5 ± 68.2, controls: 80.0 ± 76.5; $t(469) = 2.62$, $P = 0.009$, $g = 0.24$, permutation $P = 0.008$) and the peripheral region (ASD: 24.5 ± 20.8, controls: 33.4 ± 26.5; $t(469) = 4.05$, $P = 5.87 \times 10^{-5}$, $g = 0.37$, permutation $P < 0.001$). As expected, the central region of the photos was more salient than the peripheral region (all Ps $< 10^{-14}$). Together, our results suggest that photos taken by participants with ASD contained less salient objects compared to photos taken by controls, especially in the central region of the visual field. Notably, this new finding was not revealed in our previous human evaluations [Wang et al., 2016].

Comparison With Human Performance

Lastly, we compared our DNN model with human ASD experts who scored each photo based on the degree to which they thought it had been taken by a participant with ASD or by a control participant (Fig. 4). For the People Block, the AUC value for the model's performance was 0.89 ± 0.024 (mean ± SD across runs) and human performance was 0.66 ± 0.046 (Fig. 4A,B; paired *t*-test; $t(9) = 16.8$, $P = 4.15 \times 10^{-8}$, $g = 5.85$, permutation $P < 0.001$; note that in each run, the test data were the same for the current machine learning analysis and for ratings by ASD experts). For the Indoor Block, the AUC value for model performance was 0.66 ± 0.059 and human performance was 0.66 ± 0.055 (Fig. 4C,D; $t(9) = 0$, $P = 1.0$, $g = 1.87 \times 10^{-15}$, permutation $P = 0.99$). For the Outdoor Block, the AUC value for model performance was 0.70 ± 0.051 and human performance was 0.49 ± 0.066 (Fig. 4E,F; $t(9) = 7.12$, $P = 5.57 \times 10^{-5}$, $g = 3.34$, permutation $P < 0.001$). Together, this result suggests that our classifier could generally outperform professional ASD experts and might be used as an effective method to better understand visual attention in ASD and to possibly diagnose individuals with ASD.

Discussion

In this study, we developed a machine learning algorithm that can effectively discriminate photos taken by participants with ASD from photos taken by controls. Despite the relatively small size of our training data, our VGG-based algorithm has shown consistent discriminating performance. Importantly, our analysis revealed critical features that led to such successful discrimination and showed that photos taken by participants with ASD contained less salient objects, especially in the central visual field. Notably, our machine learning based ASD classification even outperformed classification by human ASD experts. Together, our findings can provide deeper insight into aberrant visual attention in ASD from a unique first-person perspective, which may in turn serve as a useful objective diagnostic tool for ASD.

Our findings also suggest that photos taken of people were the most relevant to discern between photos taken by ASD participants from those taken by controls, consistent with general social deficits in ASD [Wang & Adolphs, 2017]. Given that participants needed to communicate with the person being photographed when the photos in the People Block were taken, the photos from the People Block not only showed how participants with ASD perceived other people, but also reflected the degree to which they communicated with others. Therefore, our results also reflect deficits in social communication and

interaction typical of ASD. It is worth noting that the experimenters being photographed had abundant experience working with participants with ASD, so they were very comfortable around individuals with ASD. Although the experimenters were not blinded to which participants had ASD, they were instructed to respond to participants with ASD and controls similarly.

It is also interesting to note that participants with ASD tended to take more photos of people than did controls. This observation has not been exploited by the current machine learning algorithm, but suggests that participants with ASD may prefer to use a camera as a surrogate interface for interacting with other people. Furthermore, ASD is highly heterogeneous at the biological and behavioral levels [Happe, Ronald, & Plomin, 2006]; and all participants involved in this study were high-functioning since our photographing task required cognitive abilities to use the camera as well as basic social communication skills to interact with others when they took photos of the other people. Therefore, our results may not apply to all individuals across the autism spectrum. However, machine learning has immense potential to enhance diagnostic and intervention research in the behavioral sciences, and may be especially useful considering the heterogeneous nature of ASD [Bone et al., 2015].

The computational framework developed in this study can be readily extended to future studies that investigate other complex human social behavior and/or other neurological conditions. There have been several studies that have used photos to reveal predictive markers of psychiatric and/or neurological disorders. For example, Instagram photos have been found to reveal predictive markers of depression [Reece & Danforth, 2017]. Along this line of research, deep learning has great potential to support the use of first-person perspective photos as predictive markers of visual attention deficits for other neurological disorders such as stroke and traumatic brain injury. Our present results further demonstrate that deep learning is more accurate at discerning whether a photo represented an ASD perspective and is more efficient than human-based photo classification which requires highly skilled ASD experts and is very time intensive. Therefore, research investigating visual attention and diagnostic methods may benefit from the addition of first-person photos and DNNs to analyze those photos.

It is worth noting that there is a subtle difference when we compared the discriminating performance between ASD experts and machine learning. While machine learning used the actual photos as the training data, ASD experts did not have any training on this particular photo discrimination task nor did they receive any feedback about their performance. Rather, ASD experts had to use their knowledge or intuition to make their judgments. A future study will be needed to compare machine learning with human raters (even non-ASD experts) who are similarly trained on the photo discrimination task. Another possible extension of this study is to take first-person perspective videos. Despite some effort of machine learning based ASD analysis using home videos [Tariq et al., 2018], there is still no database of first-person perspective videos taken by individuals with ASD. We hypothesize that video clips taken by ASD patients may contain more useful and discriminative information than static photos (e.g., motion-related saliency information is often easier to characterize from a sequence of images rather than from a single image). We leave such extension of this work a possibility in a future study.

More broadly, in line with our present results, there have been efforts to collect data from the first-person perspective of participants using head-mounted cameras or head-mounted eye trackers [Bambach, Crandall, & Yu, 2015; Borjon et al., 2018; Franchak, Kretch, Soska, & Adolph, 2011]. These egocentric-view data have revealed valuable information about how infants perceive faces over their first year of life [Jayaraman, Fausey, & Smith, 2015], how social attention is coordinated between infants and parents [Yu & Smith, 2013], how people navigate in a cluttered environment [Franchak & Adolph, 2010], and how a brain lesion patient looks at faces of other people during conversations [Spezio, Huang, Castelli, & Adolphs, 2007]. A clear future direction will be to apply head-mounted cameras or head-mounted eye trackers to record egocentric views from people with ASD during their real interactions with other people and the environment. These videos will provide the most direct data about how people with ASD perceive the world from their first-person perspective. Furthermore, continuous recordings from head-mounted cameras or head-mounted eye trackers can generate massive amounts of data (notably in comparison with static photos like those used in our present study). Therefore, deep learning, which has already shown promise to discriminate ASD from controls using eye movement data with natural scene images [Jiang & Zhao, 2017; Xie et al., 2019], will make an important contribution to the analysis and interpretation of such egocentric-view data.

Acknowledgments

This research was supported by an NSF CAREER Award (1945230), ORAU Ralph E. Powe Junior Faculty Enhancement Award, West Virginia University (WVU), WVU PSCoR Program, and the Dana Foundation (to Shuo Wang), and an NSF Grant (OAC-1839909) and the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5) (to Xin Li). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Mindi Ruan, Xin Li, and Shuo Wang designed research. Mindi Ruan performed research and analyzed data. Mindi Ruan, Paula J. Webster, Xin Li, and Shuo Wang wrote the article. All authors discussed the results and contributed toward the manuscript.

REFERENCES

Ames, C., & Fletcher-Watson, S. (2010). A review of methods in the study of attention in autism. Developmental Review, 30, 52–73.

Bambach, S., Crandall, D. J., & Yu, C. (2015). Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. (pp. 351–354).

Birmingham, E., Cerf, M., & Adolphs, R. (2011). Comparing social attention in autism and amygdala lesions: Effects of stimulus and task condition. Social Neuroscience, 6, 420–435.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. Journal of Autism and Developmental Disorders, 45, 1121–1136.

Borjon, J. I., Schroer, S. E., Bambach, S., Slone, L. K., Abney, D. H., et al. (2018). A view of their own: Capturing the egocentric view of infants and toddlers with head-mounted cameras. JoVE, 140, e58445. https://doi.org/10.3791/58445

Byrge, L., Dubois, J., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. The Journal of Neuroscience, 35, 5837–5850.

Chawarska, K., Macari, S., & Shic, F. (2013). Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. Biological Psychiatry, 74, 195–203.

Dawson, G., Webb, S. J., & McPartland, J. (2005). Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. Developmental Neuropsychology, 27, 403–424.

Duda, M., Kosmicki, J. A., & Wall, D. P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Translational Psychiatry, 4, e424.

Franchak, J. M., & Adolph, K. E. (2010). Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults. Vision Research, 50, 2766–2774.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. Child Development, 82, 1738–1750.

Freeth, M., Chapman, P., Ropar, D., & Mitchell, P. (2010). Do gaze cues in complex scenes capture and direct the attention of high functioning adolescents with ASD? Evidence from eye-tracking. Journal of Autism and Developmental Disorders, 40, 534–547.

Happe, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. Nature Neuroscience, 9, 1218–1220.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. (2017). Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (pp. 3203–3212).

Hus, V., & Lord, C. (2014). The autism diagnostic observation schedule, module 4: Revised algorithm and standardized severity scores. Journal of Autism and Developmental Disorders, 44, 1996–2012.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. PLoS One, 10, e0123780.

Jiang, M., & Zhao, Q. (2017). Proceedings of the IEEE International Conference on Computer Vision. (pp. 3267–3276).

Kanner, L. (1943). Autistic disturbances of affective contact. The Nervous Child, 2, 217–250.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Archives of General Psychiatry, 59, 809–816.

LeCouteur, A., Rutter, M., & Lord, C. (1989). Autism diagnostic interview: A standardized investigator-based instrument. Journal of Autism and Developmental Disorders, 19, 363–387.

Lord, C., Rutter, M., & Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of Autism and Developmental Disorders, 24, 659–685.

Marinoiu, E., Zanfir, M., Olaru, V., & Sminchisescu, C. (2018). Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (pp. 2158–2167).

Norbury, C. F., Brock, J., Cragg, L., Einav, S., Griffiths, H., & Nation, K. (2009). Eye-movement patterns are associated with communicative competence in autistic spectrum disorders. Journal of Child Psychology and Psychiatry, 50, 834–842.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., et al. (2017). Automatic differentiation in pytorch. Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA. Retrieved from https://openreview.net/forum?id=BJJsrmfCZ

Pelphrey, K., Sasson, N., Reznick, J. S., Paul, G., Goldman, B., & Piven, J. (2002). Visual scanning of faces in autism. Journal of Autism and Developmental Disorders, 32, 249–261.

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. EPJ Data Science, 6, 15.

Rehg, J. M., Abowd, G. D., Rozga, A., Romero, M., Clements, M. A., et al. (2013). IEEE Conference on Computer Vision and Pattern Recognition. (pp. 3414–3421).

Riby, D., Hancock, P., Jones, N., & Hanley, M. (2013). Spontaneous and cued gaze-following in autism and Williams syndrome. Journal of Neurodevelopmental Disorders, 5, 13.

Riby, D., & Hancock, P. J. B. (2009). Looking at movies and cartoons: Eye-tracking evidence from Williams syndrome and autism. Journal of Intellectual Disability Research, 53, 169–181.

Rice, K., Moriuchi, J. M., Jones, W., & Klin, A. (2012). Parsing heterogeneity in autism spectrum disorders: Visual scanning of dynamic

social scenes in school-aged children. Journal of the American Academy of Child & Adolescent Psychiatry, 51, 238–248.
Rutter, M., Bailey, A., Berument, S., Lord, C., & Pickles, A. (2003). The social communication questionnaire. Los Angeles, CA: Western Psychological Services.
Santos, A., Chaminade, T., Da Fonseca, D., Silva, C., Rosset, D., & Deruelle, C. (2012). Just another social scene: Evidence for decreased attention to negative social scenes in high-functioning autism. Journal of Autism and Developmental Disorders, 42, 1790–1798.
Sasson, N. J., Elison, J. T., Turner-Brown, L. M., Dichter, G. S., & Bodfish, J. W. (2011). Brief report: Circumscribed attention in young children with autism. Journal of Autism and Developmental Disorders, 41, 242–247.
Shic, F., Bradshaw, J., Klin, A., Scassellati, B., & Chawarska, K. (2011). Limited activity monitoring in toddlers with autism spectrum disorder. Brain Research, 1380, 246–254.
Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv: 1409.1556.
Spezio, M. L., Huang, P.-Y. S., Castelli, F., & Adolphs, R. (2007). Amygdala damage impairs eye contact during conversations with real people. The Journal of Neuroscience, 27, 3994–3997.
Subramanian, V. (2018). Deep learning with PyTorch: A practical approach to building neural network models using PyTorch. Birmingham, England: Packt Publishing Ltd.
Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., & Wall, D. P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Medicine, 15, e1002705.
Wall, D. P., Kosmicki, J., DeLuca, T. F., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. Translational Psychiatry, 2, e100.
Wang, S., & Adolphs, R. (2017). Social saliency. In Q. Zhao (Ed.), Computational and cognitive neuroscience of vision (pp. 171–193). Singapore: Springer.
Wang, S., Fan, S., Chen, B., Hakimi, S., Paul, L. K., Zhao, Q., & Adolphs, R. (2016). Revealing the world of autism through the lens of a camera. Current Biology, 26, R909–R910.
Wang, S., Jiang, M., Duchesne Xavier, M., Laugeson Elizabeth, A., Kennedy Daniel, P., et al. (2015). Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron, 88, 604–616.
Wang, S., Xu, J., Jiang, M., Zhao, Q., Hurlemann, R., & Adolphs, R. (2014). Autism spectrum disorder, but not amygdala lesions, impairs social attention in visual search. Neuropsychologia, 63, 259–274.
Xie, J., Wang, L., Webster, P., Yao, Y., Sun, J., et al. (2019). *A two-stream end-to-end deep learning network for recognizing atypical visual attention in autism spectrum disorder*. arXiv preprint arXiv:1911.11393.
Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. PLoS One, 8, e79659.

Supporting Information

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

Table S1. Participant characterization. ASD was evaluated using a DSM-5 diagnosis, the Autism Diagnostic Observation Schedule-2 (ADOS-2) [Lord et al., 1989], Autism Diagnostic Interview-Revised (ADI-R) [Lord, Rutter, & Couteur, 1994], and Social Communication Questionnaire (SCQ) [Rutter, Bailey, Berument, Lord, & Pickles, 2003]. The ADOS-2 was scored according to the latest algorithm and calibrated severity scores (CSSs) were derived for exploratory correlation analyses [Hus & Lord, 2014]. The ADOS-2 is a structured interaction with an experimenter that is videotaped and scored by trained clinical staff, yielding scores on several scales. The ADOS-2 revised algorithm cutoff scores indicating an ASD diagnosis are 6 for Social Affect and 8 for Social Affect plus Restricted and Repetitive Behavior. Calibrated severity scores for each domain range from 1 to 10, with 10 indicating greatest severity. ADOS item scores were not available for three ASD participants, so we were unable to utilize the revised scoring system; however, original ADOS-2 algorithm scores for these three participants are as follows: A4: Communication = 4, Reciprocal Social Interaction (RSI) = 9, Imagination/Creativity (IC) = 1, Stereotyped Behaviors & Restricted Interests (SBRI) = 1; A6: Communication = 4, RSI = 5, IC = 0, SBRI = 1; A10: Communication = 6, RSI = 11, IC = 1, SBRI = 0.