scientific reports

OPEN

Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children

Nada Kojovic[1,5]✉, Shreyasvi Natraj[1,5], Sharada Prasanna Mohanty[2], Thomas Maillart[3,4] & Marie Schaer[1]

Clinical research in autism has recently witnessed promising digital phenotyping results, mainly focused on single feature extraction, such as gaze, head turn on name-calling or visual tracking of the moving object. The main drawback of these studies is the focus on relatively isolated behaviors elicited by largely controlled prompts. We recognize that while the diagnosis process understands the indexing of the specific behaviors, ASD also comes with broad impairments that often transcend single behavioral acts. For instance, the atypical nonverbal behaviors manifest through global patterns of atypical postures and movements, fewer gestures used and often decoupled from visual contact, facial affect, speech. Here, we tested the hypothesis that a deep neural network trained on the non-verbal aspects of social interaction can effectively differentiate between children with ASD and their typically developing peers. Our model achieves an accuracy of 80.9% (F1 score: 0.818; precision: 0.784; recall: 0.854) with the prediction probability positively correlated to the overall level of symptoms of autism in social affect and repetitive and restricted behaviors domain. Provided the non-invasive and affordable nature of computer vision, our approach carries reasonable promises that a reliable machine-learning-based ASD screening may become a reality not too far in the future.

Autism spectrum disorders (ASD) are a group of lifelong neurodevelopmental disorders characterized by impairments in social communication and interactions, and the presence of restricted, repetitive patterns of interests and behaviors[1]. Despite advances in understanding the neurobiological correlates of these disorders, there is currently no reliable biomarker for autism, and the diagnosis uniquely relies on the identification of behavioral symptoms. Although ASD can be detected as early as 14 months[2] and with high certitude before two years of age[3], the latest prevalence reports reveal that more than 70% of the affected children are not diagnosed before the age of 51 months[4]. Even in the absence of a highly specialized intervention program, earlier diagnosis is associated with a significantly better outcome. Indeed, specific strategies can be deployed to optimally support the child's development during a period of enhanced brain plasticity[5]. Previous studies have demonstrated a linear relationship between age at diagnosis and cognitive gain[6,7], whereby children diagnosed before the age of two years can gain up to 20 points in intellectual quotient (IQ) on average over the first year following diagnosis, while children diagnosed after the age of four will not show any substantial cognitive gain even with adequate intervention[7]. An efficient early screening, followed by early diagnosis, is the cornerstone to timely intervention. Most currently used screening tests are questionnaire-based, performing with low to moderate accuracy[8]. Further, they are prone to recall and subjectivity bias[9]. To overcome these limitations, tools that can deliver objective and scalable quantification of behavioral atypicalities are needed, particularly for the early detection of the signs indicative of autism.

A growing number of studies focus on the objective quantification of behavioral patterns relevant for autism, using the advances in machine learning (ML) and computer vision (CV) (for a review see[10]). For instance, Hashemi and colleagues[11] developed an automated CV approach measuring the two components of an early screening test for autism[12]. By tracking facial features, they automatically measured head turn to disengage attention from an object and head turn to track a moving object visually, the behaviors that previously were scored

[1]Psychiatry Department, Faculty of Medicine, University of Geneva, 1211 Geneva, Switzerland. [2]AIcrowd Research, AIcrowd, Lausanne, Switzerland. [3]Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland. [4]Citizen Cyber Lab, University of Geneva, Geneva, Switzerland. [5]These authors contributed equally: Nada Kojovic and Shreyasvi Natraj. ✉email: Nada.Kojovic@unige.ch

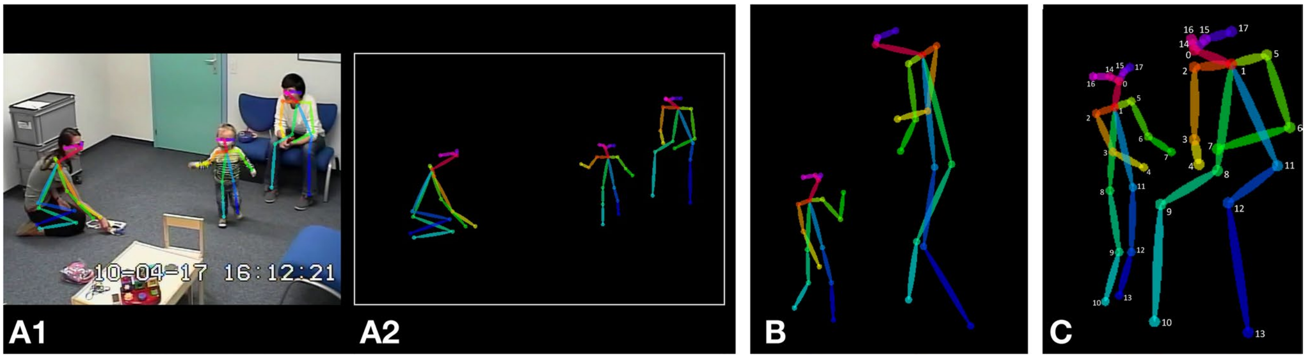

Figure 1. Example of 2D pose estimation using OpenPose on ADOS video frames: (**A1**) OpenPose keypoints overlaid a video recording from the ADOS assessment, (**A2**). OpenPose skeletal keypoints plotted over a null background, (**B**). Example of *requesting* behavior with skeletal points, (**C**). Example of *showing* behavior with numerated keypoints. Keypoint list: 0 = nose, 1 = heart, 2 = right shoulder, 3 = right elbow, 4 = right wrist, 5 = left shoulder, 6 = left elbow, 7 = left wrist, 8 = right hip, 9 = right knee, 10 = right ankle, 11 = left hip, 12 = left knee, 13 = left ankle, 14 = right eye, 15 = left eye, 16 = right ear, and 17 = left ear.

only manually. Another study using a name-calling protocol coupled with CV corroborated the well established clinical finding that toddlers with ASD respond less frequently when their name is called[13]. Additional to the automation of well established behavioral coding procedures, the use of these advanced technologies has allowed more subtle, dynamic measures of behavioral markers that would otherwise elude the standard human coding. Indeed, applying CV to the name-calling protocol revealed that, when children with ASD respond to their name, they tend to do so with an average of a 1-s delay compared to the typically developing children[13]. In other studies, the use of motion capture and CV allowed to measure the reduced complexity of emotional expression in children with ASD, especially in the eye region[14]. Additionally, the combined use of motion capture and CV have provided insights on (1) the atypical midline postural control in autism[11,15], (2) highly variable gait patterns in ASD[16] and (3) unique spatio-temporal dynamics of gestures in girls with ASD[17] that has not been highlighted in standard clinical assessments. Altogether, these studies demonstrate how computer vision and machine learning technologies can advance the understanding of autism, as they have the potential to provide precise characterizations of complex behavioral phenotypes.

The studies using ML and CV made a substantial contribution to the understanding of the disorder, offering a precise, objective measure of behavioral features that were traditionally assessed mostly qualitatively, if at all. However, there is still an important work to be done to enhance the scope and scalability of this approach. Most of the studies in this domain used fairly small samples, addressing rather specific questions focusing on one individual at time and measured behaviors elicited in controlled contexts[10]. A recent study undertook an effort to deploy a more holistic approach and, besides measuring the unique signature in the child's behavior pattern, also focused on the child's relation to immediate social context[18]. The authors used motion tracking to measure the approach and avoidance behaviors and the directedness of children's facial affect during the diagnostic assessment—the Autism Diagnosis Observation Schedule (ADOS,[19,20]). With these objective measures, the authors accounted for 30% of the standardized scores measuring the severity of autistic symptoms from only 5-min excerpts of the free play interaction with the examiner. These results are auspicious as they do not focus on an individual in isolation but are a product of a more complex effort, the dynamic measure of the child's relatedness to the social world. There is a critical need to take a more holistic stance to tackle the complex task of measuring how the child with autism interacts socially in settings close to everyday situations to advance towards a fully ecological and scalable approach.

Here, we present a machine learning algorithm to discriminate between ASD and typically developing (TD). From videos, acquired in our larger study on early development in autism, which feature social interactions between a child (with autism or TD) and an adult, we trained a deep neural network over the gold standard diagnostic assessment[19,20]. The dimensionality of the input videos was reduced applying the multi-person 2D pose estimation OpenPose technology[21] to extract skeletal keypoints for all persons present in the video (see Fig. 1). Following[22], we then applied the CNN-LSTM architecture sensible to action recognition. Our goal was to explore the potential of purely non-verbal social interactions to inform automated diagnosis class attribution. The data included in this study comprised a Training set (34 TD children and 34 from children with ASD, age range 1.2–5.1 years old), and two validation samples, namely Testing Set 1 (34 from typically developping-TD children and 34 from children with ASD, age range 1.2–5.1 years old) and Testing Set 2 ($n = 101$, uniquely children with ASD, age range 2–6.9 years old) (see Table S1). The trained model distinguished children with ASD from TD children with an accuracy exceeding 80%. These results hold potential in accelerating and automatizing autism screening approach, in a manner that is robust and only minimally influenced by video recording conditions.

Results

The final model architecture was obtained upon testing various configurations (see Fig. S2 and Supplementary section). The retained model was trained over the Training Set videos (68 ADOS videos, equally split between ASD and TD groups; see Table S1) that contained solely skeletal information on the black background, without

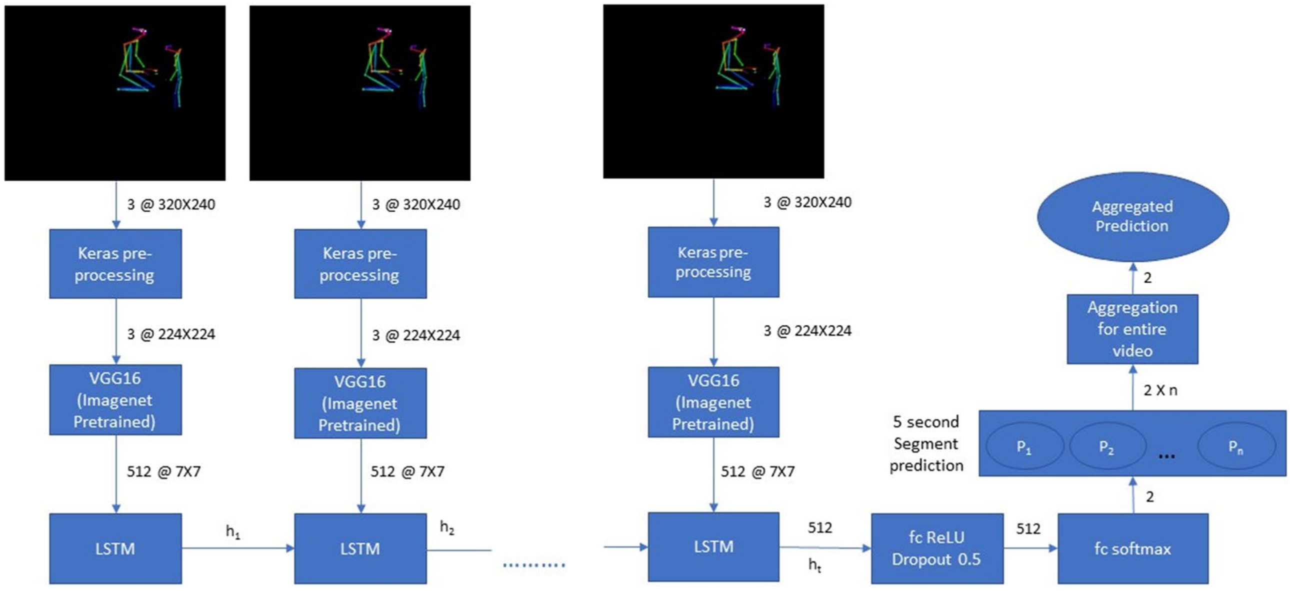

Figure 2. Neural network architecture. The pretrained Convolutional Network VGG16[22] was used to extract the characteristics of all videos split into 5s segments. The output from this feature extraction step was fed into a LSTM network operating with 512 LSTM units. Finally, the output of the LSTM was followed by 512 fully connected ReLU activated layers and a softmax layer yielding two prediction outputs for the given segment. The segment-wise classifications were aggregated for the video's duration to obtain a final prediction value (ranging from 0 to 1) that we denote "ASD probability". The video was classified as belonging to a child with ASD if the mean value of ASD probability was superior to 0.5.

Parameter	Model (80–20 split)
Accuracy	0.809
Precision (positive predicted values)	0.784
Recall (sensitivity)	0.854
Specificity	0.765
F1 Score	0.818

Table 1. Accuracy, Precision, Recall, Specificity and F1 score for Testing set 1 predictions using VGG16 LSTM trained model at 80-20 training-validation split, 100 epochs, 128 batch size.

sound (see Fig. 1). Figure 2, Figure S1, Methods and Supplementary sections detail different stages of the model training and validation. The predictions were obtained for individual 5s video segments and aggregated over the entire ADOS video for each subject from the two testing sets (see Fig. 2). We further examined the stability of the diagnosis prediction as a function of the video input length. Finally, we explored the potential of a non-binary, continuous value of ASD probability to capture meaningful clinical characteristics, examining whether standardized scores obtained from various gold-standard clinical assessments related to the ASD probability extracted from the neural network.

Prediction of autism. Our model achieved an F1 score of 0.818 and a prediction accuracy of 80.9% over a sample of 68 videos in Testing Set 1 (Table 1). The same trained model achieved a prediction accuracy of 80.2% over a Testing Set 2 comprising 101 videos from children with ASD, thus endorsing the model's stability.

Consistency of the ASD prediction over the video length. We further tested the extent to which the video length influenced our model's prediction accuracy. By varying the length of the video input in the Testing set 1, we demonstrated that an average 70% accuracy is already obtained with 10 min video segments (see Fig. 3A). As shown in Fig. 3B, the prediction consistency is also very high across the video of a single individual, even with relatively short video segments. For instance, for half of the ASD sample, our method achieves a 100% consistency in prediction based on randomly selected 10 min segments. These results strongly advocate for the feasibility of video-based automated identification of autism symptoms. Moreover, the ADOS videos used in the present study were acquired using different systems. However, the accuracy of classification showed robustness to the variability in context, thus again highlighting the potential for generalization of this type of approach (see Supplementary section and Fig. S4).

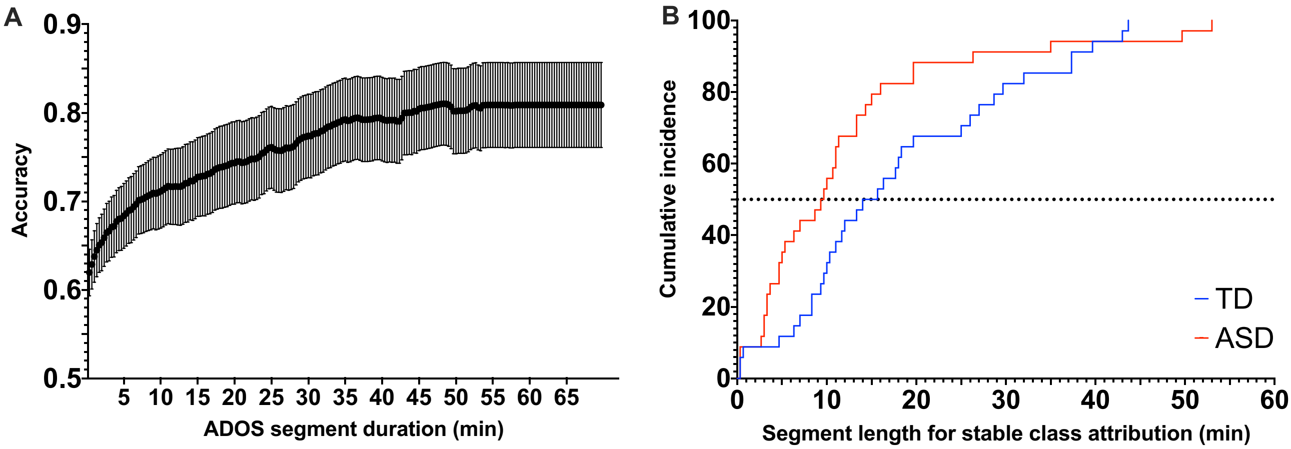

Figure 3. (**A**) Association between the prediction accuracy and the length of considered video segment. The accuracy increases with longer video segments, with the final accuracy being 81% for Testing set 1 in our sample. (**B**) Stability in the prediction as a function of the length of the considered video segment for the Testing set 1. The cumulative incidence depicts the required length of video segments that is needed to achieve 100% prediction consistency for all the segments of the same length randomly drawn from the full video of the same participant.

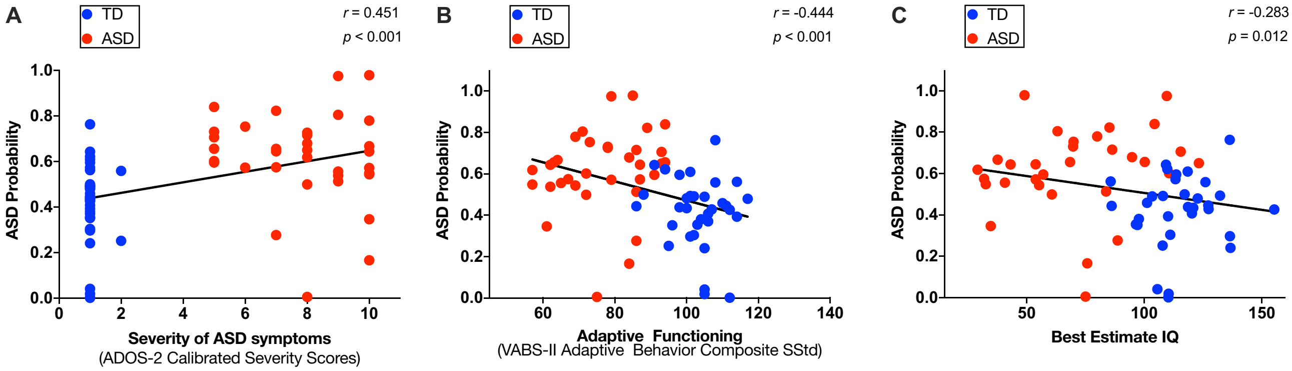

Figure 4. Scatter plots depicting the relation between predicted ASD probability and clinical measures: (**A**) Total level of severity of autistic symptoms, (**B**). Adaptive functioning, (**C**). Best estimate Intellectual Quotient. The least squares linear fit is depicted as black line and values of Spearman r coefficient and corresponding p values are shown on each panel. Ground-truth classes: ASD = red, TD = blue.

Correlation between neural network derived ASD probability and the clinical phenotype. Using validated clinical assessments ("Methods" section), we then observed that neural network derived ASD probability was positively related to the overall level of symptoms of autism ($r_s(68) = 0.451$, $p < 0.001$), (Fig. 4A), and this pattern was observed both in the domain of the social affect ($r_s(68) = 0.509$, $p < 0.001$) and in the domain of repetitive and restricted behaviors (RRB) ($r_s(68) = 0.409$, $p < 0.001$) (Fig. S5, panels A1-2). Moreover, ASD probability negatively correlated with the general adaptive functioning ($r_s(67) = -0.444$, $p < 0.001$) (Fig. 4B). Further analyses revealed that ASD probability was related to the communication ($r_s(68) = -0.386$, $p < 0.001$), socialization ($r_s(68) = -0.477$, $p < 0.001$) as well as the autonomy in daily life ($r_s(68) = -0.397$, $p < 0.001$) but not with the functioning in the motor domain ($r_s(68) = -0.186$, $p = 0.066$) (Fig. S5, panels B1-3). Finally, ASD probability showed a moderate negative correlation with cognitive functioning ($r_s(63) = -0.283$, $p = 0.012$) (Fig. 4C).

The above correlations were based on ADOS severity scores, representing a summarized measure of the degree of autistic symptoms. In addition, we were interested in understanding how the automatically derived ASD probability related to individual autistic symptoms and potentially inform us about the symptoms that were more closely related to ASD class attribution. After applying Bonferroni corrections, ASD probability was positively related with three symptoms in the communication domain of ADOS, namely gestures ($r_s(68) = 0.435$, $p < 0.001$), pointing ($r_s(68) = 0.540$, $p < 0.001$) and intonation ($r_s(68) = 0.426$, $p < 0.001$) (Fig. S6, panels A-C). In the social interaction domain of the ADOS the ASD probability was related to unusual eye contact ($r_s(68) = 0.500$, $p < 0.001$), directed facial expressions ($r_s(68) = 0.488$, $p < 0.001$), spontaneous initiation of joint attention ($r_s(68) = 0.450$, $p < 0.001$), integration of gaze and other behaviors ($r_s(68) = 0.591$, $p < 0.001$), giving ($r_s(68) = 0.438$, $p < 0.001$), showing ($r_s(68) = 0.396$, $p < 0.001$), shared enjoyment ($r_s(68) = 0.359$,

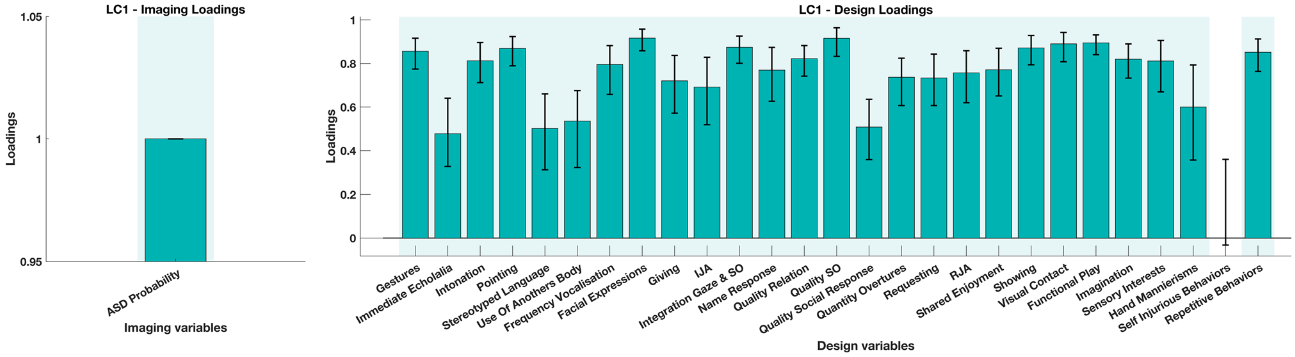

Figure 5. Loading on the latent component obtained using the Partial Least Squares Correlation (PLS-C) analysis. The cross-correlation matrix was composed on the behaviors (Y) side with 27 individual symptoms (gestures, immediate echolalia, intonation of vocalizations/verbalizations, pointing, stereotyped language, use of another's body, frequency of spontaneous vocalizations, facial expressions directed to others, giving, IJA-initiation of joint attention, integration of gaze and another behaviors during social overtures, response to name, quality of relation, quality of social response, quality of social overtures (SO), requesting, RJA-response to joint attention, shared enjoyment in interaction, showing, unusual eye contact, functional play with objects, imagination, unusual sensory interest in play material/person, hand and finger and other complex mannerisms, self-injurious behaviors, unusually repetitive interests or stereotyped behaviors) and ASD Probability on the Imaging (X) side.

$p = 0.001$), quality of social overtures ($r_s(68) = 0.484$, $p < 0.001$) (Fig. S6, panels D-K). Furthermore, ASD probability was positively related to functional play ($r_s(68) = 0.418$, $p < 0.001$) and imagination ($r_s(68) = 0.470$, $p < 0.001$)(Fig. S6, panels L-M). Finally, in the domain of repetitive behaviors and restricted interests, ASD probability was related to unusual sensory behaviors (($r_s(68) = 0.434$, $p < 0.001$) and unusually repetitive interests and stereotyped behaviors ($r_s(68) = 0.455$, $p < 0.001$) (Fig. S6, panels N-O). The symptoms that were related to the ASD probability were predominantly non verbal. While the correlational analysis seem coherent with our clinical intuition, the current level of analysis prevents us from drawing solid conclusions on what is represented in the NN.

To account for the collinearity among ADOS symptoms we deployed a multivariate analysis Partial Least Squares Correlation(PLS-C)[23,24] (see "Methods" section). This analysis is particularly suitable in the case where there is no assumption of independence among tested variables. In this manner, alongside taking into account the correlation within the symptoms, we were able to show the symptom pattern that best explained the variation in NN derived ASD probability. The PLS-C yielded a significant latent component (p = 0.001) best explaining the cross-correlation pattern between the ASD probability and the symptoms patterns. As shown on Fig. 5 the symptoms that showed higher contribution to the pattern of were predominantly non verbal (e.g. Facial Expressions, Quality of Social Overtures, Gestures, etc.). These symptoms depend heavily on the coordination of various non verbal behaviors, such as directing the facial affect (through head orientation) or using various communicative gestures. The symptoms least contributing to the pattern were among verbal symptoms including Immediate Echolalia and Stereotyped language that are especially discriminative in children with more fluid verbal language which was not the characteristics of the current sample. Intonation was the only symptom with predominantly vocal components that loaded highly on the current LC. However, in our clinical experience, the higher scores on intonation are usually followed by higher scores on many non verbal behaviors. The atypical intonation is seen in more severely affected children who usually are minimally verbal and present a lot of undirected vocalizations, bringing them simultaneously to higher scores in many other symptom domains, such as atypical social overtures, atypical social response, lack of directed facial expressions.

Discussion

Our neural network model operated on a low dimensional postural representation derived from social interaction videos between a child and an adult and robustly distinguished whether the child has autism, with a model prediction accuracy of 80.9% (precision: 0.784; recall: 0.854). We choose the model that operates on a relatively reduced set of characteristics, targeting non-verbal features of social interaction and communication, particularly relevant in very young children[19,20]. We deployed an LSTM network learning temporal dependencies of relations between skeletal keypoints in 5s interaction segments to perform the classification. Our findings' clinical validity was corroborated with positive correlations between neural network derived ASD probability and levels of autistic symptoms and negative correlations between the same measure and cognitive and everyday adaptive functioning of these children. Moreover, we showed that the accuracy of classification of around 70% was achieved based on only 10 min of filmed social interaction, opening avenues for developing scalable screening tools using smaller excerpts of videos.

The choice of the reduced dimensionality (2D estimated postural representations) in input videos was twofold: to allow a pure focus on non-verbal interaction patterns and ensure de-identification of the persons involved (skeletal points plotted against the black background). To further promote the approach's scalability, videos were not manually annotated; thus, we removed the human factor in the initial feature breakdown. Our design aligns with and expands findings from a smaller number of studies that probed automated video data classification[25,26]. Zunino et al.[25] were able to determine a specific signature of grasping gesture in individuals with autism using the raw video inputs. This approach allowed a classification accuracy of 69% using 12 examples of grasping action per individual obtained in well-controlled filming conditions. In contrast, in our study, the input videos were remarkably heterogeneous in terms of recorded behaviors. They included moments of the interaction of a child with an examiner in the presence of caregiver(s). Moreover, different examiners performed the assessments as they are acquired as a part of the ongoing longitudinal study. Finally, regarding the pure physical aspects, these assessments took place in several different rooms and were filmed with different camera systems. Nevertheless, our study's classification accuracy is superior to the one reported in the study using the controlled video of a very precise grasping action.

The clinical validity of our findings is supported by the significant correlation observed at the level of individual autistic symptoms (single items from the ADOS). The lack of ability to adequately integrate gaze to other communication means and impaired use of visual contact together with the reduced orientation of facial affect were the symptoms that were the most related to the neural network derived values of probability of ASD class attribution. Other symptoms that were strongly linked to the probability of receiving ASD class attribution comprise aberrant gesture use, unusual social overtures, repetitive patterns of behaviors and unusual sensory interests. These findings emphasize the potential of these low-dimensional social interaction videos to convey the atypicality of the non-verbal interaction patterns in young children with ASD. Indeed, out of 27 selected behaviors the 15 that were significantly related to the neural network derived ASD probability were the symptoms with a predominant non-verbal (e.g., giving, showing, spontaneous initiation of joint attention). This finding is in line with findings reported applying ADOS-like rating on home videos[27] who found that the aberrant use of visual contact was a feature that was the most determinant of ASD classification. Another study building an automated screener for autism symptoms based on annotated home videos reported that the screener capitalized on the non-verbal behaviors (e.g. eye contact, facial expressions) in younger participants while relying more on verbal features in older participants[28]. Indeed, clinically, the aberrant use of visual contact and aberrant gesture use are among the most striking and early emerging features of the disorder[19,20,29].

The major contribution of the automated identification of behaviors indicative of autism lies in enhancing the efficiency and sensitivity of the screening process. The diagnosis process is complex and delicate and is unlikely to be set on the track of automated performance before long. However, more informed, more objective and available screening is crucial to catalyze diagnosis referrals, hopefully leading to earlier intervention onset. Early interventions are of life-changing importance for individuals with autism. They improve their cognitive and adaptive functioning and reduce the severity of ASD symptoms[6,30]. In the year following the diagnosis, children who receive early intervention—and start developing language skills before the age of 3 show the most important gains as young adults[31,32].

Our results speak in favor of more objective, holistic, automatized methods as complementary tools to the ones used in clinical practice. In ASD, the availability of standardized measures of autistic symptoms was crucial in informing the clinic and the research[33]. Nevertheless, these gold-standard measures still rely on somewhat coarse descriptions of symptoms. Individual symptoms of autism are assessed on a 3 or 4 point scale[19,20,29] while phenotypical differences between the behaviors brought on the same plane can be very pronounced. The development and improvement in quantitative measures leading to a more fine-grained "digital phenotyping"[34] can be a tremendous asset in the early detection of signs of the disorder and the follow-up of its manifestation through development. Besides being more objective compared to human coding, it can allow the processing of larger quantities of the data and at the spatio-temporal resolution that is off limits to human coding. Moreover, these precise and continuous measures may uncover behavioral manifestations of the disorder that were previously not evidenced. They also may help define sub-types of the disorder to allow more precise clinical action[35]. The finding that we were able to achieve a robust accuracy of classification based on a limited set of characteristics derived from social interaction videos is very promising. This approach would further benefit from the implementation of the spatio-temporal attentional mechanism[36] to allow knowledge on the specific elements in space and time used to inform the diagnosis process in the network and improve our understanding of the manifestation of the disorder. Additionally, building on the evidence from the present study, our next goal is to perform fine-grained annotation of the behaviors along the non verbal continuum that were the most contributing to the discrimination between the two groups. Thus, the precise annotation of the incidence of communicative gestures, shared affective states, atypical social overtures would be highly beneficial to provide the insight on the content of neural network learning and provide a more sensitive measure of the disorder manifestation.

Methods

Participants.

The initial sample included sixty-eight children with autism (2.80 ± 0.92 years) and 68 typically developing children (2.55 ± 0.97 years) who were equally distributed to compose the Training and Testing set (Testing Set 1), matched for diagnosis, age, gender and ADOS module (see Table S1). To validate the robustness of our classification method we included an additional testing sample comprising 101 videos from children with ASD (3.46 ± 1.16 years) that we denote Testing set 2. All data used in this study were acquired as a part of a larger study on early development in autism. Detailed information about cohort recruitment was given elsewhere[37–39]. The current study and protocols were approved by the Ethics Committee of the Faculty of Medicine of the Geneva University, Switzerland. The methods used in the present study were performed in accordance with the relevant guidelines and regulations of the Geneva University. For all children included in this study, informed written consent was obtained from a parent and/or legal guardian. Children with ASD were included based on a clinical diagnosis according to DSM-5 criteria[1], and the diagnosis was further corroborated using the gold standard diagnostic assessments (see Clinical Measures subsection and Supplementary section).

Parameter	Training set	Testing set 1	Testing set 2
Number of videos	68	68	101
Total duration (in h)	47.782	48.661	50.043
Videos per class	34	34	101 (ASD)
Average video length	42.16 min	42.68 min	41.965 min
Average number of 5 s segments per video	505.92	512.16	503.574

Table 2. Representation of video characteristics included in the Training and Testing set 1 as well as Testing set 2 used to validate the robustness of neural network derived classification.

Typically developing (TD) children were screened for the presence of any known neurological or psychiatric illness and ASD in any first-degree relative of the child.

Clinical measures. A direct measure of autistic symptoms was obtained using the Autism Diagnostic Observation Schedule-Generic ADOS-G[19], or a more recent version Autism Diagnostic Observation Schedule-2nd edition (ADOS-2)[20]. Cognitive functioning was assessed using various assessments depending on the children's age and their ability to attend demanding cognitive tasks. We defined the Best Estimate Intellectual Quotient[39,40] that combines the most representative cognitive functioning measures for each child. Adaptive functioning was assessed using the Vineland Adaptive Behavior Scales, second edition (VABS-II;[41]) (see Supplementary section for a detailed characterization of clinical measures).

Video data. To probe the diagnosis classification using machine learning on social interaction videos, we used filmed ADOS assessment acquired in the context of our study. Practical reasons drove this choice, ADOS being the most frequent video-based assessment in our study (systematically administered in all children included in our study). Moreover, ADOS provides a standardized and rich context to elicit and measure behaviors indicative of autism across broad developmental and age ranges[19]. Its latest version (ADOS-2) encompasses five modules covering the age from 12 months to adulthood and various language levels ranging from no expressive use of words to fluent complex language. To best fit the younger participants' developmental needs, Modules Toddler 1 and 2 are conducted while moving around a room using a variety of attractive toys, while Modules 3 and 4 happen mostly at a table and involve more discussion with lesser use of objects. In this work, we focused uniquely on the Modules Toddler, 1 and 2, as these require fewer language abilities and are more sensitive to non-verbal aspects of social communication and interaction that we target using machine learning. The clinical findings on the prevalence of non verbal-symptoms in younger children[19,20,29] drove our choice to focus uniquely on non-verbal aspects of communication and interaction.

Pose estimation. To purely focus on social interaction and essentially its non-verbal aspects, we extracted skeletal information on people present in ADOS videos using deep learning based multi-person 2D pose estimator-OpenPose[21]. OpenPose estimates keypoints of persons detected on the image/video independently for each frame. It uses a simple 2D camera input not requiring any external markers or sensors, thus allowing the retrospective analysis of videos. It also is immune to variations in resolutions and setting that images and videos might present. For the OpenPose output, we opted for the COCO model providing 2D coordinates of 18 keypoints (13 body and 5 facial keypoints; see Fig. 1). The ordering of the keypoints is constant across persons and frames. Consistent with our focus on the non-verbal features of interaction during the semi-standardized ADOS assessment, we removed the background from all the videos and preserved only skeletal information for further analysis. To obtain feature vectors invariant to a rigid body and affine transformations and to increase the generalizability of our approach, we based our calculation on image output and not on raw keypoints coordinates (Fig. 1)[42]. Of note, the informed written consent for publication of identifying information/images in an online open-access publication was obtained from both adults and parent and/or legal guardian of child featuring in the illustration on Fig. 1 and Fig. S1.

Building the neural network. The OpenPose processed videos were down-sampled from 696 × 554 to 320 × 240 pixels and split into segments of 5s (see Table 2). To estimate the training and validation loss we used a categorical cross entropy loss function using a rmsprop optimizer. We found that the 5-s video segments were optimal for model training and resulted in less validation loss compared to longer segments (10 s or 15 s) (see Fig. S3). We opted for a CNN LSTM architecture for our model as it previously showed a good performance in video-based action classification[43]. We used a VGG16 convolutional neural network[44], pretrained on the ImageNet[45] dataset to extract high dimensional features from individual frames of the 5 s video clips.The VGG16 is a 16 layers convolutional neural network that works with a 224 × 224 pixel 3 channel(RGB) input frame extracted from the video segment. The resolution is then decreased along the each convolution and pooling layer as 64 @ 112 × 112, 128 @ 56 × 56, 256 @ 28 × 28, 512 @ 14 × 14 and 512 @ 7 × 7 after the last convolution or pooling stage which has 512 feature maps. The high dimensional features extracted are flattened and input to a 512 hidden layered 0.5 dropout LSTM at a batch size of 128[46] followed by fully connected dense layers with ReLU activation, 0.5 dropout and a softmax classification layer giving an 2 dimensional output (corresponding to the two classes, ASD and TD).

The input training data of 68 ADOS videos were split in the ratio of 80-20, where the model used 80% of data for training and 20% of data was used for validation. We then analyze the model's training and validation loss to avoid overfitting and perform hyperparameter tuning. The training and validation loss over a varied number of epochs is shown on Figure S3. The least validation loss model was deployed to predict over 5-s segment of the videos from the two testing sets (Testing Set 1 and Testing Set 2). We average the predictions for all the video segments to obtain a final prediction value denoted as "ASD probability". We trained the neural network model over 5-s video segments at 128 batch size, 100 epochs and 80-20 training-validation split and used the trained model to make predictions over Testing Set 1 and Testing set 2 to check the accuracy of the prediction results across different testing sets.

The model training and validation was performed at University of Geneva High Performance Computing cluster, Baobab (Nvidia RTX 2080Ti, Nvidia Titan X and Nvidia Quadro P1000 GPUs).

Statistical analyses. The obtained measure of ASD probability derived from the neural network was further put in relation to standardized behavioral phenotype measures in children included in the Testing 1 sample. We calculated Spearman rank correlations with measures of severity of symptoms of autism (Total CSS, SA CSS and RRB CSS), adaptive (VABS-II Total and scores across four subdomains) and cognitive functioning (BEIQ) (Supplementary section). Furthermore, in order to obtain more fine-grained insight into the relation of ASD probability across the entire video and the specific symptoms of autism, this measure was correlated with raw scores on a selected set of 27 items that repeat across the three modules we used (for more details, please refer to Supplementary section). Results were considered significant at $p < 0.05$. The significance level was adjusted using Bonferroni correction for multiple comparisons. Thus the results concerning the two ADOS subdomains results were considered significant at $p < 0.025$ and four VABS-II subdomains at $p < 0.013$. For the analyses involving the 27 individual items of ADOS (for a full list please refer to Table S2), the results were considered significant at $p < 0.002$. Considering the collinearity between our ADOS measures we employed the Partial Least Squares Correlation (PLS-C) analysis[23,24], to model the relation with the ASD probability. This method refers to the optimal least squares fit to the part of the correlation matrix. A cross-correlation matrix (R) is computed between the NN derived ASD Probability (denoted as X or Imaging variable) and 27 individual ADOS symptoms (denoted as Y or Design variables) $\mathrm{R} = \mathrm{Y}^T\mathrm{X}$. The R is then decomposed using the Singular Value Decomposition $\mathrm{R} = \mathrm{U}\Delta\mathrm{V}^T$ yielding two singular vectors U and V (also denoted as saliences). The singular vector U represents the symptom pattern that best characterizes the R, while the singular vector V represent the dependent variable pattern that best characterizes the cross-correlation matrix R. In order to test the pattern generalizability we performed the permutation testing by generating 1000 permutation samples yielding a sampling distribution of the singular values under the null hypothesis. The stability of the obtained saliences U and V is obtained by creating 1000 bootstrap samples (sampling with replacement the observations in X and Y). To derive the latent variable pairs (LV) reflecting the maximal covariance patterns between the ASD probability on one side and a pattern of individual symptoms of ASD on the other side the singular vectors U and V were projected into the original Behavior (Y) and Imaging (X) matrices respectively. Thus obtained patterns by definition would be independent of alternative LV ASD probability—symptom pairings.

For a comparison of the Training and Testing samples with regards to clinical assessments (ADOS, VABS-II and BEIQ), we used Student t-tests and Mann-Whitney tests when measures did not follow a normal distribution according to the D'agostino & Pearson test (See Table S1). Statistical analyses were conducted using GraphPad Prism v.8, (https://www.graphpad.com/scientific-software/prism/) and SPSS v.25 (https://www.ibm.com/analytics/spss-statistics-software).

Relation of video length to prediction accuracy. Our final aim was to apprehend the length of video segments required for stable class attribution, thus probing the approach's scalability. To this end, we employed a sliding window approach, starting with a length of 20 s and then stepwise increasing the window length by 20 s until window length matched video duration. In each window, ASD probability values are averaged over the containing 5-s segments for each of 68 videos in the Testing set 1 (Fig. 3A). Using this method, we also test the prediction consistency for videos of a single individual class by identifying the sliding window length required for stable class attribution (Fig. 3B).

Received: 8 March 2021; Accepted: 9 July 2021
Published online: 23 July 2021

References

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)* (American Psychiatric Pub, New York, 2013).
2. Landa, R. J., Gross, A. L., Stuart, E. A. & Faherty, A. Developmental trajectories in children with and without autism spectrum disorders: the first 3 years. *Child Dev.* **84**, 429–442. https://doi.org/10.1111/j.1467-8624.2012.01870.x (2013).
3. Ozonoff, S. *et al.* Diagnostic stability in young children at risk for autism spectrum disorder: a baby siblings research consortium study. *J. Child Psychol. Psychiatry* **56**, 988–998. https://doi.org/10.1111/jcpp.12421 (2015).
4. Maenner, M. J. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2016. *MMWR Surveill. Summ.*https://doi.org/10.15585/mmwr.ss6904a1 *(2020).*
5. Dawson, G. Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Dev. Psychopathol.* **20**, 775–803. https://doi.org/10.1017/S0954579408000370 (2008).

6. Dawson, G. *et al.* Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics* **125**, e17-23. https://doi.org/10.1542/peds.2009-0958 (2010).
7. Robain, F., Franchini, M., Kojovic, N., Wood de Wilde, H. & Schaer, M. Predictors of treatment outcome in preschoolers with autism spectrum disorder: an observational study in the Greater Geneva Area, Switzerland. *J. Autism Dev. Disord.* **50**, 3815–3830. https://doi.org/10.1007/s10803-020-04430-6 (2020).
8. Yuen, T., Penner, M., Carter, M. T., Szatmari, P. & Ungar, W. J. Assessing the accuracy of the modified checklist for autism in toddlers: a systematic review and meta-analysis. *Dev. Med. Child Neurol.* **60**, 1093–1100. https://doi.org/10.1111/dmcn.13964 (2018).
9. Taylor, C. M., Vehorn, A., Noble, H., Weitlauf, A. S. & Warren, Z. E. Brief report: can metrics of reporting bias enhance early autism screening measures?. *J. Autism Dev. Disord.* **44**, 2375–2380. https://doi.org/10.1007/s10803-014-2099-5 (2014).
10. de Belen, R. A. J., Bednarz, T., Sowmya, A. & Del Favero, D. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Transl. Psychiatry* **10**, 1–20. https://doi.org/10.1038/s41398-020-01015-w (2020) (**Number: 1 Publisher: Nature Publishing Group**).
11. Hashemi, J. *et al.* Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. *Autism Res. Treat.* **2014**, e935686. https://doi.org/10.1155/2014/935686 (2014).
12. Bryson, S. E., Zwaigenbaum, L., McDermott, C., Rombough, V. & Brian, J. The autism observation scale for infants: scale development and reliability data. *J. Autism Dev. Disord.* **38**, 731–738. https://doi.org/10.1007/s10803-007-0440-y (2008).
13. Campbell, K. *et al.* Computer vision analysis captures atypical attention in toddlers with autism. *Autism Int. J. Res. Pract.*https://doi.org/10.1177/1362361318766247 *(2018).*
14. Guha, T., Yang, Z., Grossman, R. B. & Narayanan, S. S. A computational study of expressive facial dynamics in children with autism. *IEEE Trans. Affect. Comput.* **9**, 14–20. https://doi.org/10.1109/TAFFC.2016.2578316 (2018).
15. Dawson, G. *et al.* Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder. *Sci. Rep.* **8**, 17008. https://doi.org/10.1038/s41598-018-35215-8 (2018).
16. Rinehart, N. J. *et al.* Gait function in newly diagnosed children with autism: cerebellar and basal ganglia related motor disorder. *Dev. Med. Child Neurol.* **48**, 819–824. https://doi.org/10.1017/S0012162206001769 (2006).
17. Rynkiewicz, A. *et al.* An investigation of the 'female camouflage effect' in autism using a computerized ADOS-2 and a test of sex/gender differences. *Mol. Autism*https://doi.org/10.1186/s13229-016-0073-0 *(2016).*
18. Budman, I. *et al.* Quantifying the social symptoms of autism using motion capture. *Sci. Rep.* **9**, 1–8. https://doi.org/10.1038/s41598-019-44180-9 (2019).
19. Lord, C. *et al.* The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30**, 205–223 (2000).
20. Lord, C. *et al. Autism Diagnostic Observation Schedule: ADOS-2* (Western Psychological Services, Los Angeles, Calif., 2012) (**OCLC: 851410387**).
21. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008 [cs] (2019)
22. Orozco, C. I., Xamena, E., Buemi, M. E. & Berlles, J. J. Human Action Recognition in Videos using a Robust CNN LSTM Approach (Reconocimiento de Acciones Humanas en Videos usando una Red Neuronal CNN LSTM Robusta) (2021).
23. Krishnan, A., Williams, L. J., McIntosh, A. R. & Abdi, H. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. *NeuroImage* **56**, 455–475. https://doi.org/10.1016/j.neuroimage.2010.07.034 (2011).
24. McIntosh, A. R. & Lobaugh, N. J. Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage* **23**, S250–S263. https://doi.org/10.1111/j.1467-8624.2012.01870.x8 (2004).
25. Zunino, A. *et al.* Video Gesture Analysis for Autism Spectrum Disorder Detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 3421–3426, https://doi.org/10.1109/ICPR.2018.8545095 (2018). ISSN: 1051-4651.
26. Li, J. *et al.* Classifying ASD children with LSTM based on raw videos. *Neurocomputing* **390**, 226–238. https://doi.org/10.1111/j.1467-8624.2012.01870.x9 (2020).
27. Tariq, Q. *et al.* Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: development and validation study. *J. Med. Internet Res.* **21**, e13822. https://doi.org/10.2196/13822 (2019).
28. Abbas, H., Garberson, F., Glover, E. & Wall, D. P. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* **25**, 1000–1007. https://doi.org/10.1111/jcpp.124211 (2018) (**Publisher: Oxford Academic**).
29. Lord, C. *et al.* Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J. Autism Dev. Disord.* **19**, 185–212 (1989).
30. Estes, A. *et al.* Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. *J. Neurodev. Disord.* **7**, 24. https://doi.org/10.1186/s11689-015-9117-6 (2015).
31. Anderson, D. K., Liang, J. W. & Lord, C. Predicting young adult outcome among more and less cognitively able individuals with autism spectrum disorders. *J. Child Psychol. Psychiatry* **55**, 485–494. https://doi.org/10.1111/jcpp.124213 (2014).
32. Pickles, A., Anderson, D. K. & Lord, C. Heterogeneity and plasticity in the development of language: a 17-year follow-up of children referred early for possible autism. *J. Child Psychol. Psychiatry* **55**, 1354–1362. https://doi.org/10.1111/jcpp.124214 (2014).
33. Lord, C. *et al.* A multi-site study of the clinical diagnosis of different autism spectrum disorders. *Arch. Gen. Psychiatry* **69**, 306–313. https://doi.org/10.1001/archgenpsychiatry.2011.148 (2012).
34. Dawson, G. & Sapiro, G. Potential for digital behavioral measurement tools to transform the detection and diagnosis of autism spectrum disorder. *JAMA Pediatr.* **173**, 305–306. https://doi.org/10.1001/jamapediatrics.2018.5269 (2019).
35. Hsin, H. *et al.* Transforming psychiatry into data-driven medicine with digital measurement tools. *NPJ Digit. Med.* **1**, 1–4. https://doi.org/10.1038/s41746-018-0046-0 (2018) (**Number: 1 Publisher: Nature Publishing Group**).
36. Sharma, S., Kiros, R. & Salakhutdinov, R. Action Recognition using Visual Attention. arXiv:1511.04119 [cs] (2016)
37. Franchini, M. *et al.* The effect of emotional intensity on responses to joint attention in preschoolers with an autism spectrum disorder. *Res. Autism Spectr. Disord.* **35**, 13–24. https://doi.org/10.1016/j.rasd.2016.11.010 (2017).
38. Franchini, M. *et al.* Early adaptive functioning trajectories in preschoolers with autism spectrum disorders. *J. Pediatr. Psychol.* **43**, 800–813. https://doi.org/10.1093/jpepsy/jsy024 (2018).
39. Kojovic, N., Ben Hadid, L., Franchini, M. & Schaer, M. Sensory processing issues and their association with social difficulties in children with autism spectrum disorders. *J. Clin. Med.* **8**, 1508. https://doi.org/10.3390/jcm8101508 (2019).
40. Howlin, P., Savage, S., Moss, P., Tempier, A. & Rutter, M. Cognitive and language skills in adults with autism: a 40-year follow-up. *J. Child Psychol. Psychiatry* **55**, 49–58. https://doi.org/10.1017/S09545794080003703 (2014).
41. Sparrow, S. S., Balla, D. & Cicchetti, D. V. *Vineland II: Vineland Adaptative Behavior Scales: Survey Forms Manual: a Revision of Hte Vineland Social Maturity Scale by Edgar A. Doll* (Pearson, Boca Raton, 2005).
42. N, B. R., Subramanian, A., Ravichandran, K. & Venkateswaran, N. Exploring Techniques to Improve Activity Recognition using Human Pose Skeletons. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 165–172. https://doi.org/10.1109/WACVW50321.2020.9096918 (2020).
43. Karpathy, A. *et al.* Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223 (2014). ISSN: 1063-6919.
44. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] (2015)

45. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848 (2009). ISSN: 1063-6919.
46. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).

Acknowledgements

We express our utmost gratitude to all families that took part in this study. We thank our colleagues for their precious contribution to data collection. Authors express a special gratitude to Ms Kenza Latrèche who helped with manual annotation of recording settings.

Author contributions

S.P.M., N.K., T.M., M.S. planned the experimental setup; N.K. and M.S conducted the experiments and recordings. N.K. and M.S recruited families to the study; N.K. conducted the preparation of the videos and applied OpenPose; S.N. conducted pre-processing of videos for the neural network and developed the neural network system; N.K. and M.S. performed statistical analysis; N.K. and S.N. wrote the initial draft of the manuscript, which was edited and reviewed by all authors.

Funding

Funding for this study was provided by National Centre of Competence in Research (NCCR) Synapsy, financed by the Swiss National Science Foundation (SNF, Grant Number: 51NF40_185897), by SNF Grants to M.S. (#163859 and #190084), the UNIGE COINF2018 equipment grant, by the SDG Solution Space (https://www.fablabs.io/labs/sdgsolutionspace) and by the Fondation Pôle Autisme (https://www.pole-autisme.ch).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information is available for this paper at https://doi.org/10.1038/s41598-021-94378-z.

Correspondence and requests for materials should be addressed to N.K.

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

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