2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII) | 979-8-3503-2743-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACII59096.2023.10388157

On Scalable and Interpretable Autism Detection from Social Interaction Behavior

William Saakyan[1], Matthias Norden[1], Lola Herrmann[2], Simon Kirsch[3], Muyu Lin[2],
Simon Guendelman[2], Isabel Dziobek[2], and Hanna Drimalla[1*]

[1]Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Germany
[2]Institute of Psychology, Humboldt University of Berlin
[3]Department of Psychiatry and Psychotherapy, Medical Center-University of Freiburg

Abstract—**Autism Spectrum Condition (ASC) is characterized by social interaction difficulties that can be challenging to assess objectively in the diagnostic process. In this paper, we evaluate the capability of using videos of a standardized social interaction to differentiate non-verbal behaviors of individuals with and without ASC. We collected a large video dataset consisting of 164 participants with ASC (n = 83) and neurotypical individuals (n = 81) who completed the computer-based Simulated Interaction Task (SIT) in different studies including lab and home settings. To classify individuals with and without ASC, we trained uni- and multimodal machine learning models based on different modalities such as facial expressions, gaze behavior, head pose and voice features. Our results indicate that a multimodal late fusion approach achieved the highest accuracy (74%). In the unimodal setting, classification based on facial expressions (accuracy 73%) and voice features (accuracy 70%) were most effective. An explainability analysis of the most relevant features for the facial expression model indicated that features from all emotional parts as well as from both the speaking and listening part of the interaction were informative. Based on our results, we developed a scalable online version of the SIT to collect diverse data on a large scale for the development of machine learning models that can differentiate between different clinical conditions. Our study highlights the potential of machine learning on videos of standardized social interactions in supporting clinical diagnosis and the objective and effective measurement of differences in social interaction behavior.**

Index Terms—**ASC, Classification, ML, Multimodal, Explainable, Autism**

I. Introduction

Facial expressions, voice characteristics, and other non-verbal cues play a pivotal role in human communication, allowing individuals to convey emotions and intentions without the use of words. The ability to accurately interpret and respond to these cues is essential for effective social interaction [1]. Impairments in social communication and behavior are characteristic for Autism Spectrum Conditions (ASC) as defined by the Diagnostic and Statistical Manual of Mental Disorders [2]. These impairments include challenges in initiating or maintaining conversations, avoiding eye contact, exhibiting unexpected emotional responses, and engaging in behaviors that are perceived as socially inappropriate [3]. These symptoms can significantly impact the daily lives of individuals with ASC, leading to difficulties in forming and maintaining friendships, developing romantic relationships, and succeeding in educational and employment settings.

Traditional assessments of social interaction difficulties within the diagnostics of ASC primarily rely on questionnaires and the subjective evaluations of experts [2], [3]. Experts usually need to be trained to conduct complex autism related tests for diagnosis like the Autism Diagnostic Observation Schedule or the Autism Diagnostic Interview - Revised [4]. Consequently, patients seeking for support often have to deal with long waiting times or cannot find suitable therapists [5]. Especially diagnosing adults with ASC poses a challenge due to their capabilities to compensate for their behaviors [6], which is even more present in females [7] and may lead to missed cases and misdiagnosis [8]. Moreover, despite the common view of autism to be predominantly a male disorder, more recent studies suggest that females are underrepresented in ASC studies [9] and should be more integrated [10].

Beyond that, the expression of ASC symptoms is highly heterogeneous among affected individuals [11]. Some patients might exhibit difficulties with eye contact, while others may struggle with demonstrating appropriate facial expressions in response to social cues. Consequently, also social skills interventions like cognitive behavioral therapy or social stories may not be equally effective for all individuals with ASC [12].

Not only diagnosing ASC but also characterizing the associated social interaction difficulties of different patients with ASC could hence support the development of more targeted interventions and strategies that can address their unique challenges.

Automated analysis of non-verbal social cues (e.g. facial expressions and voice) can provide a solution to the limitations of subjective assessments, offering more scalable, objective and reliable evaluations of social interaction characteristics linked to autism. In this study, we aim to evaluate the capability of using videos of standardized social interactions to differentiate the non-verbal behavior of autistic and neurotypical adults in clinical and home settings. The contributions of this

This study was funded by the Federal Ministry of Education and Research Germany (BMBF; 01IS20046), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC-2049 – 390688087). The BMBF and the DFG had no involvement in the study design, the collection, analysis and interpretation of data, the writing of the report and the decision to submit the article for publication.

* Corresponding author, drimalla@techfak.uni-bielefeld.de

979-8-3503-2743-4/23/$31.00 ©2023 IEEE

work are as follows: First, we collect a large video dataset capturing social behavior of adults with or without ASC using a recent standardized interaction paradigm, in clinical and home settings. While many studies focus on children and/or male participants [9], our dataset contains a remarkable high proportion of females. Despite the challenge of replicating machine learning results on larger datasets and across studies [13], we successfully replicate findings of [14]. Moreover, we enhanced the classification performance of individuals with ASC and neurotypical (NT) individuals by additional feature extraction of non-verbal interaction behavior from the collected videos and applying a late fusion approach. We evaluate which modalities are most informative for the detection of ASC and apply a machine learning explainability method to inform on the importance of individual features. Finally, we present an updated version of the Simulated Interaction Task (SIT) that will improve gaze tracking and is also better suited for outside-the-clinic settings. We aim to contribute to a more scalable, objective and interpretable assessment of specific social interaction difficulties in individuals with ASC to ultimately make diagnostics more accessible and reliable.

II. Related Work

A. Video-based Autism Detection

Video-based autism detection algorithms have shown promising results in recent years [15]. Many of these are meant to support early diagnosis and focus on the recognition of stereotyped behaviors like rocking, pacing or hand flapping which can be characteristic for ASC [16]–[18]. In studies conducted by Abbas et al. [19] and Tariq et al. [20], videos of children taken at home were assessed for ASC-related behaviors by both untrained and experienced evaluators. The ratings were then used in combination with other diagnostic features to develop a predictive model for ASC. While the results improved over traditional questionnaire-based assessments, these methods still require manual annotations of the videos, which could include evaluator biases [21] and are not scalable. In their recent paper, Ali et al. [22] collected and annotated a video dataset for action recognition tasks of children with ASC in an uncontrolled environment. They proposed a multi-modality fusion deep learning network and were able to recognize autistic behaviors with an accuracy of 86.04%. Hashemi et al. [23] used videos of children to identify behavioral signs of ASC based on the Autism Observation Scale for Infants. They assessed head motion and body pose and demonstrated how automated measurements can aid clinical diagnosis.

While these systems are promising for supporting diagnostics of ASC based on stereotypical behaviors, only a few works focus on characteristics related to social interactions despite their relevance for daily life.

Using their "Multimodal Dyadic Behaviour Dataset" that contains 121 children's social and communicative behavior captured in videos during a clinical session, Rehg et al. [24] developed algorithms to detect specific ASC behaviors and predict child engagement. Similarly, Billing et al. [25] recently published their "DREAM" dataset which covers more than 300 hours of therapy sessions of children interacting with a social robot or a therapist. These datasets are valuable for supporting the development of video-based ASC diagnosis algorithms and characterizing the social interaction behavior. Still, many of these are specifically focusing on children and recorded in highly controlled clinical sessions.

Georgescu et al. [26] evaluated videos of naturalistic and complex interactions between adults with ASC or without ASC and a non-autistic/neurotypical interaction partner. The study included 29 participants with and 29 participants without ASC. While they could classify adults with ASC from neurotypically developing adults using non-verbal intrapersonal motion synchrony with an accuracy of 75.9%, they left evaluating other potentially social interaction relevant features like facial expressions and gaze.

According to a recent review on computer vision approaches in ASC research from 2009 to 2019 [15], these approaches can be useful for the automated quantification of behavioral markers but that standardized large-scale benchmark datasets, especially capturing settings outside the clinics, are still missing, which makes a holistic evaluation difficult. Nearly all studies focused on children with ASC in order to support early diagnosis [15] despite the increasing need and relevance of methods for detecting ASC in adults [5], [27].

B. Simulated Interaction Task

The Simulated Interaction Task (SIT) [14], [28] was proposed as a computer-based tool for automatically assessing social interaction deficits and has shown to reliably elicit social interaction behavior. The authors performed a study with 37 adults with ASC and 43 healthy controls and detected individuals with ASC with an accuracy of 73%, sensitivity of 67%, and specificity of 79%, using facial expression and vocal features extracted from the videos collected with the SIT. Reduced smiling and a higher voice fundamental frequency were found to be characteristic for ASC while gaze behavior was not significantly different between groups. The video dataset in this study was gathered in a highly controlled setting and focused on patients with high-functioning ASC. It's not clear, if the results are generalizable for settings outside the clinics. Further, machine learning analyses were limited to facial expressions, gaze behavior and vocal characteristics while the head pose was not evaluated despite its utility in previous works [25], [29].

In this study, we replicate and extend the results obtained in [28] and [14]. We apply the SIT in three different studies in clinical and home settings, gathering a large video data set of individuals with ASC and healthy controls in standardized social interactions. We include additional social interaction relevant features in a multimodal classification task and provide insights into digital biomarkers of ASC using explainability methods.

III. Methods

In this section, we outline the data acquisition process and the machine learning techniques employed to distinguish patients diagnosed with ASC from neurotypical individuals based on social interaction behavior.

A. Data Collection

In order to obtain video data of social interaction behavior in individuals with ASC and neurotypical individuals, we collected and combined three datasets from two university labs (SIT$_{Uni1}$ and SIT$_{Uni2}$) and various home settings (SIT$_{Home}$). All studies have been approved by the respective ethical committees (approval numbers 20-1144_3 and 2021-20) and participants received monetary compensation.

1) Participants & Recruitment: The lab studies are part of a collaborative research project (register number DRKS00017817) taking place at the Humboldt University of Berlin and the Medical Center-University of Freiburg[1]. Participants were included based on a diagnosis of Asperger syndrome, high-functioning autism or atypical autism with deficits in social communication and social interaction. Diagnoses were provided by clinicians using the ICD-10 criteria [3]. Further relevant inclusion criteria were an age between 18 and 65 years, none or stable psychopharmacotherapy, IQ higher or equal to 80, fluency in German and a relevant psychosocial impairment lower or equal to a score of 60 as measured with the Global Assessment of Functioning [2]. Participants with psychiatric comorbidities of schizophrenia, psychosis, acute maniac episode within bipolar disorder, acute severe depression or acute suicidality were not included. Neurotypical participants were required to have no history of psychiatric conditions in their medical records and to not have been taking psychotropic medication within the last 3 months.

Participants of the home study were recruited via different psychotherapy outpatient clinics and therapy groups on the one hand, and via participant recruitment systems of universities, online advertisements and postings on the other hand. We included participants who met the following criteria: fluency in German and an age range between 18 and 65. Additionally, participants with glasses were recommended to use contact-lenses during the SIT, to ensure more precise tracking of eye movements. In this study, participants were asked to report any clinical diagnoses they had received from a psychologist, psychiatrist, or doctor. Participants were included when reporting a diagnosis of ASC, Social Anxiety Disorder (SAD), Depression or no psychopathology at all (NT group). Diagnoses of ASC were checked through medical records during recruitment. As we are interested in the specific social interaction behavior related to ASC, we excluded participants that reported comorbidities with conditions that are potentially relevant to social interaction (i.e., SAD and Depression) for our analysis.

TABLE I
Collected datasets

Dataset Name	**Study Setting**	**Participants**			**Labels**	
		$\sum$	***Excluded***	***Left***	***ASC***	***NT***
SIT$_{Home}$	Home	97	27	70	26	44
SIT$_{Uni1}$	Lab	61	2	59	29	30
SIT$_{Uni2}$	Lab	35	0	35	28	7
Total		193	29	164	83	81

Table I presents the number of participants and exclusions for the respective studies. Further exclusions were based on video and feature extraction quality described below. Demographic information of the final participants whose data have been analysed for this work can be found in Table II. Overall, our dataset includes 88 male, 74 female, and 2 diverse participants, ranging in age from 18 to 63 years.

2) Procedure: Participants of the lab studies were informed on the study procedure via telephone. On the lab visit they signed a written consent and performed the SIT as one part of the trial procedure. The SIT was conducted in one of two rooms with stable lighting and recording conditions.

Participants of the home study were informed on the study procedure via e-mail. They were informed that the study comprises a virtual dialogue with a pre-recorded actress and to ensure constant lighting conditions with no disturbing backgrounds in their rooms. They were linked to an online questionnaire platform where they had to provide a digital consent and answer questions on demographic and diagnostic information. Following the questionnaires, the participants were then instructed to install and start the SIT program on their computers during which they were video recorded. Finally, the data collected with the SIT were uploaded to a university server by the participants.

3) SIT: The SIT [14], [28] incorporates a 7-minute long standardized video conversation with a pre-recorded actress. Participants perform the SIT in front of a screen while being video-recorded. Before the conversation starts, participants are asked to position their head within a certain area on the screen in order to adjust for different camera setups. During the conversation, the actress introduces herself, outlines the conversation, and then interviews the participant on food preferences. The first part is a "neutral" part on dinner preparations, followed by a "joy" part on liked, and a "disgust" part on disliked food. Each of the three conversation topics is divided into an "actress speaking" part during which the actress explains her own preferences and a "participant speaking" part during which the participant should respond to the respective topic while the actress shows empathic listening behaviors such as smiling and nodding.

B. Preprocessing

Since the SIT is designed for anyone to use with their own computers and webcams, the video quality in our study will naturally differ between participants. In the collected data, most of the recordings had 25-30 FPS. To ensure homogeneity across the different video recordings used in our study, we

[1] Detailed information on the trial can be found on https://doi.org/10.1186/s13063-021-05205-9.

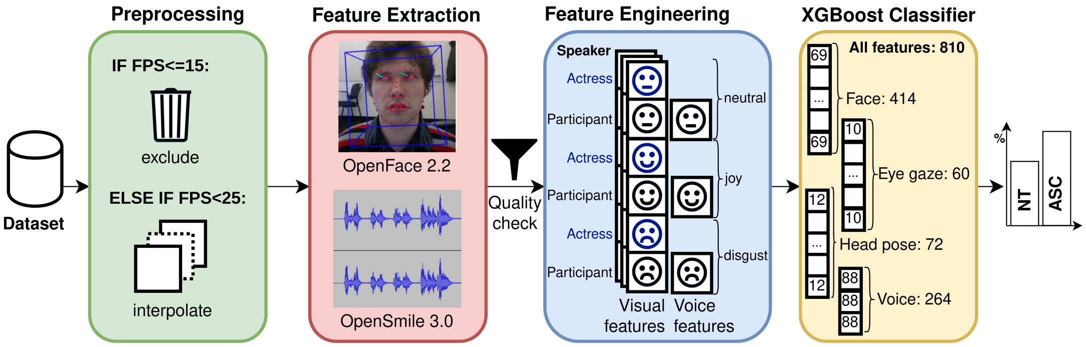

Fig. 1. The complete pipeline. In the preprocessing step, the collected dataset is filtered, and videos are interpolated if necessary. Next, relevant features are extracted, and after verifying their correctness and confidence, statistical functionals are calculated based on the interaction parts for each modality. The XGBoost classifier is trained on the features, depending on the setting (unimodal or multimodal), to predict whether a participant has ASC. For the multimodal setting, features from all modalities are concatenated into a single vector.

TABLE II
DEMOGRAPHICS: GENDER (M/F/D) AND AGE (MEDIAN, RANGE)

Dataset	ASC			ASC Age		NT		NT Age	
Name	M	F	D	Median	Range	M	F	Median	Range
SIT$_{Home}$	13	13	0	36	18-57	23	21	27	18-58
SIT$_{Uni1}$	16	13	0	28	18-53	12	18	32	19-49
SIT$_{Uni2}$	19	7	2	35	20-63	5	2	32	21-59
Total	48	33	2	35	18-63	40	41	29	18-59

interpolated the videos with FPS higher than 15 but lower than 25 to have 30 FPS. We excluded recordings with an FPS of 15 or lower, as increasing the FPS by more than two times resulted in unnatural appearing videos. Ultimately, we obtained a dataset consisting of videos recordings captured at a minimum frame rate of 25 FPS.

C. Feature extraction and engineering

In this section, we describe the process of extracting and engineering relevant features from the video and audio data to train our machine learning models for ASC classification.

1) Visual features: Using the open source toolkit OpenFace 2.2 [30] we extracted facial expression, head pose and eye gaze.

After extracting visual features with OpenFace, we left out frames with a facial landmark detection confidence under 75% and where the face detection success was 0 (binary value). If we had to leave out more than 10% of the frames, we did not include the related participants in our analysis.

Using the timestamps of the pre-recorded actress video, we split each participant's video into three interaction sections: neutral, joy, and disgust. We further divided each of these parts based on the speaker (either the actress or the participant), resulting in a total of six interaction segments. For each part, statistical functionals were computed.

Based on the Facial Action Coding System, facial action units (AUs) represent 44 specific facial muscle actions, e.g., AU12 representing lip corner puller. OpenFace can identify a subset of AUs which are frequently applied in **facial expression** analysis. OpenFace processes each video frame to determine the presence (binary) and intensity (on a scale of 0 to 5) of 18 AUs, except for AU28, which only provides presence information. We did not apply user normalization. This means that the estimates of AU features are based on a single frame without calibration to a participant, as our focus was on capturing individual differences. For each of the six parts we determined the mean, variance, and median of each Action Unit intensity, along with the mean for each binary value. Since the frame count for the same interaction part can vary among different participants due to excluded frames or small variations of FPS, the mean of the binary value can be interpreted as the frequency of an AU presence, normalized by the frame count.

Eye gaze modality refers to the analysis of a participant's gaze direction and movement, which can provide information about attention and engagement during an interaction. The eye gaze features were calculated based on the x and y angles in radians of a gaze vector extracted by OpenFace. First, we calculated gaze movement velocity and acceleration, saccade amplitude, and fixation duration. Next, for each of these features mean and standard deviation were calculated. To prevent potential gender bias, we excluded the mean and standard deviation of gaze angle y, as male participants might be taller than female participants on average and included only gaze angle x. To calculate the duration of fixation, we identified the time intervals (in seconds), during which the change in a participant's gaze angle was smaller than a specific threshold. We used a threshold of 9 degrees, calculated using the rounded average degree error from [30].

Head pose modality analyzes the orientation and movement of a participant's head, also offering valuable information about their engagement and attention during interactions. For the head pose modality, the calculations were performed using rotation vectors given by OpenFace (in radians) around the X (pitch), Y (yaw), and Z (roll) axes, with the camera positioned as the origin. These are computed based on a 3D representation of facial landmarks. We computed the mean and standard deviation of multiple features, such as roll, yaw, movement and stability durations, and the velocity and acceleration of head movement. Similar to the eye gaze modality, we excluded the mean and standard deviation of pitch to eliminate potential gender bias. To calculate movement and stability of the head pose as durations in seconds, we extracted all durations where rotation values were higher (movement) or lower (stability) than a set threshold. We used a threshold of 3 degrees, calculated using the rounded average degree error from [30].

2) Audio features: For the **voice** modality, we employed the python library OpenSmile 3.0 [31] and extracted the eGeMAPSv02 [32] feature set. Unlike for the visual features, we calculated functionals exclusively for the three interaction parts in which it was the participant's turn to speak.

The complete pipeline, along with an overview of the feature sizes is illustrated in Figure 1.

D. Machine Learning Model and Training

The final representation is generated by concatenating the extracted features from each of the six interaction parts into a single vector for each participant. This vector includes facial expressions, eye gaze, head pose, and voice modalities (with three parts specifically for the voice modality) from all the interaction parts. First, we trained models on unimodal data, i.e., only features from the corresponding modality were used.

Given that our data is tabular in nature, we opted to employ XGBoost[2] for data classification, as it has been demonstrated that it performs exceptionally well on this type of data [33]. XGBoost is an efficient and scalable gradient boosting algorithm designed for various machine learning tasks, including classification and regression, and is known for its speed and performance. We conducted hyperparameter tuning using a nested cross-validation (CV) approach, whereby we employed a participant-based Leave-One-Out Cross-Validation [34] in the outer loop and a 5-fold CV in the inner loop to tune hyperparameters. However, we observed that it did not yield any improvements compared to the model using default parameter values. The hyperparameter values varied significantly across different splits, which may be attributed to the limited size of the dataset, preventing accurate estimation of the optimal hyperparameters. For more detailed information on the specific settings used for hyperparameter tuning, please refer to the corresponding GitHub repository[3]. Therefore, we utilized a model with default values and without any fine-tuning. At the time of training, the default boosting algorithm - *gbtree*, was used with a learning rate of 0.3 and a default tree depth set to 6. In the multimodal setting, we used an early and a late fusion approach. For early fusion, all features from each modality were concatenated into a single feature vector. For the late fusion approach we combined the probability outputs from the XGBoost models trained on the unimodal tasks. This involved a total of 8 values, consisting of two values per class for each of the four modalities. To capture non-linear relationships and interactions, we applied polynomial features with a degree of 2. Given the relatively small number of values, we opted for a logistic regression model with the *liblinear* solver and *L2* norm for the classification of the data. Furthermore, we implemented participant-based Leave-One-Out Cross-Validation [34] for all analyses presented in this paper.

TABLE III
CLASSIFICATION RESULTS

Modality	**Accuracy**	**ASC Precision**	**ASC Recall**
Multimodal (late)	**74**	**73**	**76**
Multimodal (early)	66	68	63
Facial Expression	73	**73**	75
Voice	70	71	66
Eye gaze	55	56	55
Head pose	68	68	70

E. Model Interpretation

We investigated which features had the most significant impact on the model's decision and in which part they were most relevant. To accomplish this, we utilized the SHapley Additive exPlanations (SHAP) library [35] to provide the contribution of various facial expression features to the final prediction. In addition to calculating feature contributions, the SHAP framework provides various visualizations to visually describe feature contributions. It helps to interpret the output of a machine learning model by determining the importance of each feature to the model's predictions.

IV. RESULTS

Using a total of 164 participants' videos, we trained a machine learning model for autism classification based on non-verbal social interaction features. To assess the performance of models, we computed accuracy, precision and recall of the detection of participants with ASC. The corresponding values are available in Table III. The late fusion approach integrating all modalities yielded the highest performance, while sing the early fusion approach resulted in slightly inferior performance compared to the unimodal models. In terms of individual modality performance, the accuracy of the voice modality was 70%, following the facial expression modality, which achieved the highest accuracy of 73% in the unimodal setting and outperformed the other modalities. Both precision and recall values for ASC prediction were similarly high.

Overall, the trained models achieved high performance across all modalities, except for the eye gaze modality, where the performance was marginally above chance level. Some modalities performed better than others, as shown by the ROC-curves and AUC-values in Figure 2.

To evaluate which features influenced the performance of our best-performing modality, the results of utilizing SHAP library are presented in Figure 3. Feature names are formatted as "(interaction part: neutral, joy, disgust)-(speaker: actress, participant)-(AU: binary **c** or intensity **r**)-(name of the functionals)". On the x-axis each dot represents a SHAP value for the participant. A value of 0 indicates that the model disregards the value of that particular feature. A higher SHAP value signifies that the value of the given feature increases the

[2] https://xgboost.readthedocs.io/en/stable/python/index.html

[3] https://github.com/mbp-lab/acii23_sit_autisim_detection

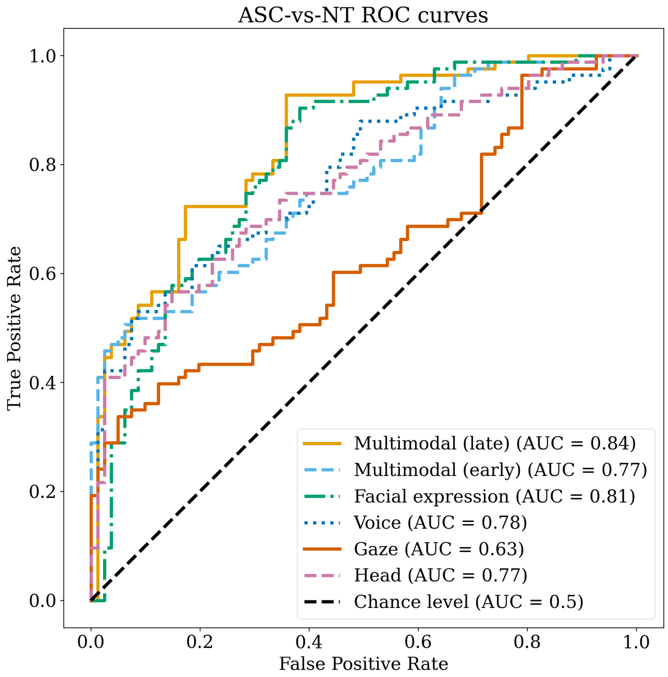

Fig. 2. ROC curves for each modality.

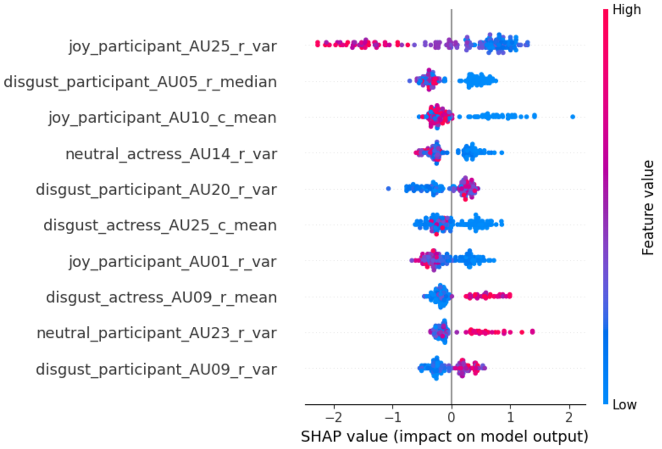

Fig. 3. The distribution of importance for facial expression features and their influence on predictions.

model's output. The value of the corresponding feature for that participant is represented by the color, with red indicating high values and blue - low values. The larger positive output means a higher impact towards predicting participants having ASC.

As observed in Table II, gender distribution among neurotypical participants is balanced; however, this is not the case for participants with ASC, where male participants outnumber females by 55%. To explore the potential presence of gender bias, we plotted the distribution of correct predictions by gender across five different modalities in Figure 4. It displays the percentage of True Positive (correctly predicting ASC condition) and True Negative (correctly predicting non-ASC condition) predictions. The percentage of correct predictions

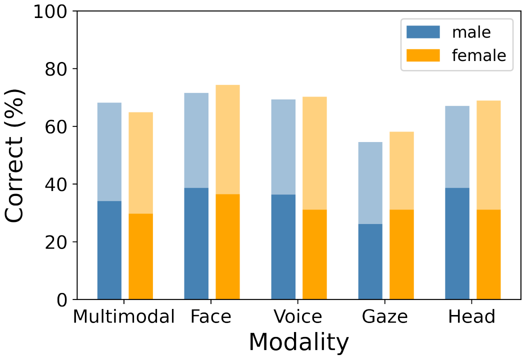

Fig. 4. The stacked bar chart illustrates the distribution of correct male and female predictions across five different modalities. The dark color represents the percentage of True Positive ASC predictions, while the lighter color indicates True Negative predictions.

for each modality and different genders are similar as well as the rates of True Positive and True Negatives.

V. Online Version of SIT

Based on the experiences of conducting the SIT in a home setting, we developed an online version of the SIT to make data collection more scalable and accessible. In this version, it is possible to participate without installing any software using only a browser and webcam. Additionally, we implemented a gaze calibration part as described in [36] at the beginning of the test to enable more accurate gaze analyses. A demonstration version of the current prototype without storing any data can be tested via https://www.simulatedinteraction.com.

VI. Discussion

In this paper, we examined the predictability of the diagnosis of ASC based on facial expression, gaze behavior and voice patterns and head pose in 83 participants with ASC and 81 neurotypical participants who took part in three different studies with lab and home settings. Our findings suggest that it is feasible to classify participants with and without ASC based on extracted features from videos capturing non-verbal interaction behavior. Further, we were able to replicate the results from Drimalla et al. [14] using a dataset more than twice the size. This achievement is particularly noteworthy considering the challenges associated with small sample sizes in high-dimensional datasets, which can lead to biased machine learning performance estimates [37] as well as the challenge of replicating results on larger and more balanced datasets [13].

Similar to [14], the combination of all modalities result in a improved performance compared to unimodal approaches. Also, both facial expression as well as vocal patterns were most informative for the detection of autism spectrum condition in the unimodal setting. The head pose representation was also proven to be informative.

The eye gaze modality exhibits the poorest performance, which is close to the chance level, similar to the work of Drimalla et al. [14]. As previous work [38] suggests eye gaze can provide valuable features, we examined the eye gaze features post-hoc. Significant changes in participant's gaze angle visible in the video, were not accurately reflected in the feature space, exhibiting little to no variations. Enhancing the eye gaze feature extractor may improve the classification task's success.

Aiming for a transparent classification method, we examined with an explainability method, how the different features of the most predictive modality, Facial Expression, contributed to the detection of ASC. The features' importance values indicated that the most relevant features originate from all parts of the conversation, including neutral, joy and disgust parts. Similarly, AUs from parts, where mainly the participants were actively speaking as well as parts were they were listening, seem to convey information. Further XAI methods [39] may shed light on if particular groups of features were especially relevant or gain information on particular individuals.

The remarkable high representation of females in our dataset allowed us to investigate the potential influence of gender-related features, such as fundamental frequency [40], on the classification task. We observed similar results across all modalities for male and female participants, speaking against a strong gender bias of one specific modality.

Further, we ensured that the classification works similarly well for different genders. Due to the small number of diverse participants, we could only investigate this for female and male participants. There was no evidence for a gender bias of the classifier, as both were equally well classified. Nevertheless, we cannot rule out that the ground truth diagnosis might be gender biased, as gender differences in social interaction and communication for individuals with ASC are likely not captured by pre-existing diagnostic instruments and may potentially result in the under-recognition of autism in females [41].

Predicting a dimensional diagnostic value instead of diagnostic categories, may enable more targeted and promising medical or psychotherapeutic treatments. Going one step further and predicting dimensional interaction traits in participants with different conditions as well as in a neurotypical population, may improve general understanding of social interaction impairments beyond classical diagnostic categories.

The newly developed online version of the SIT can now be easily integrated in diagnostic procedures and conducted for patients with many different conditions. Further, its easy set-up now allows to conduct large-scale online studies with the SIT to collect data with even greater diversity in terms of conditions, cultural background, socio-economic status, etc., which can improve the performance and generalizability of the findings. A follow-up online study of this kind with a data protection compliant data storage concept has already been approved by the ethics committee.

VII. Conclusion

The contributions of this paper include the collection of a large diverse dataset of social interaction behaviors of individuals with and without ASC, implementing uni- and multimodal classifiers for automatic detection of ASC and developing an online version of the SIT. The classification based on facial expressions and voice features was most effective, outperforming the multimodal approach while classification results based on eye gaze were worst. For the future, we aim to focus on both method development and conceptual development. More complex machine learning models, an improved representation of the non-verbal behavior including more precise gaze tracking and more sophisticated fusion approaches may benefit the classification. Conceptually, we aim to dive deeper into a more complex analysis of interaction traits, moving beyond the simple classification of diagnoses or conditions. Lastly, we will continue collecting a diverse dataset of social interaction behavior from individuals with different clinical conditions as well as neurotypical individuals through collaborations with university outpatient clinics and large-scale online studies.

Ethical Impact Statement

With our work we aim to contribute to a more accurate, objective and scalable assessment of social interaction behavior in ASC. Specifically, we evaluated which information extracted from videos of social interactions can be informative for detecting and describing ASC. Although we aimed for a larger and more representative sample, our dataset could still inherit existing biases and stereotypes, particularly because the diagnoses of ASC are partially based on subjective interpretations of clinicians. Directly applying algorithms which are trained on potentially biased datasets for diagnostic purposes could circularly reinforce the existing biases and lead to inaccurate results in different cohorts. Importantly, the systems presented in this work are not meant to substitute clinician expertise within the diagnostic process but to assist by providing objective assessments of social interaction behavior. Our algorithms were developed using a specific interaction paradigm and hence, should not be taken as a general measure to detect ASC behaviors, especially in other contexts. We emphasized the limitations of the proposed methods and invite for careful consideration when applying such algorithms for decision making.

References

[1] R. A. Hinde and R. A. Hinde, *Non-Verbal Communication*. Cambridge University Press, 1972.

[2] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-5.*, 5th ed. Arlington, VA: American Psychiatric Association, 2013.

[3] W. H. Organization, "International statistical classification of diseases and related health problems (10th ed.)," 2016.

[4] S. Bölte and F. Poustka, "[Psychodiagnostic instruments for the assessment of autism spectrum disorders]," *Zeitschrift Fur Kinder- Und Jugendpsychiatrie Und Psychotherapie*, vol. 33, no. 1, pp. 5–14, Jan. 2005.

[5] "Autism Statistics, January to December 2022," https://digital.nhs.uk/data-and-information/publications/statistical/autism-statistics/january-to-december-2022.

[6] G. A. McQuaid, N. R. Lee, and G. L. Wallace, "Camouflaging in autism spectrum disorder: Examining the roles of sex, gender identity, and diagnostic timing," *Autism: The International Journal of Research and Practice*, vol. 26, no. 2, pp. 552–559, Feb. 2022.

[7] M. Kirkovski, P. G. Enticott, and P. B. Fitzgerald, "A Review of the Role of Female Gender in Autism Spectrum Disorders," *Journal of Autism and Developmental Disorders*, vol. 43, no. 11, pp. 2584–2603, Nov. 2013.

[8] L. Fusar-Poli, N. Brondino, P. Politi, and E. Aguglia, "Missed diagnoses and misdiagnoses of adults with autism spectrum disorder," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 272, no. 2, pp. 187–198, Mar. 2022.

[9] A. M. D'Mello, I. R. Frosch, C. E. Li, A. L. Cardinaux, and J. D. Gabrieli, "Exclusion of females in autism research: Empirical evidence for a "leaky" recruitment-to-research pipeline," *Autism Research*, vol. 15, no. 10, pp. 1929–1940, 2022.

[10] L. Barnard-Brak, D. Richman, and M. H. Almekdash, "How many girls are we missing in ASD? An examination from a clinic- and community-based sample," *Advances in Autism*, vol. 5, no. 3, pp. 214–224, Jan. 2019.

[11] W. Jones and A. Klin, "Heterogeneity and Homogeneity Across the Autism Spectrum: The Role of Development," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 48, no. 5, pp. 471–473, May 2009.

[12] M. D. Lerner and S. W. White, "Moderators and Mediators of Treatments for Youth With Autism Spectrum Disorders," in *Moderators and Mediators of Youth Treatment Outcomes*, M. Maric, P. J. Prins, and T. H. Ollendick, Eds. Oxford University Press, Aug. 2015, p. 0.

[13] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, May 2015.

[14] H. Drimalla, T. Scheffer, N. Landwehr, I. Baskow, S. Roepke, B. Behnia, and I. Dziobek, "Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (SIT)," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, Feb. 2020.

[15] R. A. J. de Belen, T. Bednarz, A. Sowmya, and D. Del Favero, "Computer vision in autism spectrum disorder research: A systematic review of published studies from 2009 to 2019," *Translational Psychiatry*, vol. 10, no. 1, pp. 1–20, Sep. 2020.

[16] S. S. Rajagopalan, "Computational behaviour modelling for autism diagnosis," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: Association for Computing Machinery, Dec. 2013, pp. 361–364.

[17] S. S. Rajagopalan and R. Goecke, "Detecting self-stimulatory behaviours for autism diagnosis," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 1470–1474.

[18] S. S. Rajagopalan, A. Dhall, and R. Goecke, "Self-Stimulatory Behaviours in the Wild for Autism Diagnosis," in *2013 IEEE International Conference on Computer Vision Workshops*. Sydney, Australia: IEEE, Dec. 2013, pp. 755–761.

[19] H. Abbas, F. Garberson, E. Glover, and D. P. Wall, "Machine learning approach for early detection of autism by combining questionnaire and home video screening," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 1000–1007, Aug. 2018.

[20] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall, "Mobile detection of autism through machine learning on home video: A development and prospective validation study," *PLOS Medicine*, vol. 15, no. 11, p. e1002705, Nov. 2018.

[21] W. Saakyan, O. Hakobyan, and H. Drimalla, "Assessment of representational bias in emotion datasets," https://github.com/mbp-lab/caip2021_bias_emotions/tree/v1.0.0, 2021.

[22] A. Ali, F. F. Negin, F. F. Bremond, and S. Thümmler, "Video-based Behavior Understanding of Children for Objective Diagnosis of Autism," in *VISAPP 2022 - 17th International Conference on Computer Vision Theory and Applications*, Feb. 2022.

[23] J. Hashemi, M. Tepper, T. Vallin Spina, A. Esler, V. Morellas, N. Papanikolopoulos, H. Egger, G. Dawson, and G. Sapiro, "Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants," *Autism Research and Treatment*, vol. 2014, p. 935686, 2014.

[24] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, H. Rao, J. C. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye, "Decoding Children's Social Behavior," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3414–3421.

[25] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. H. Garcia, P. G. Esteban, H. Liu, V. Nair, S. Matu, A. Mazel, M. Selescu, E. Senft, S. Thill, B. Vanderborght, D. Vernon, and T. Ziemke, "The DREAM Dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy," *PLOS ONE*, vol. 15, no. 8, p. e0236939, Aug. 2020.

[26] A. L. Georgescu, J. C. Koehler, J. Weiske, K. Vogeley, N. Koutsouleris, and C. Falter-Wagner, "Machine Learning to Study Social Interaction Difficulties in ASD," *Frontiers in Robotics and AI*, vol. 6, 2019.

[27] D. G. M. Murphy, J. Beecham, M. Craig, and C. Ecker, "Autism in adults. New biologicial findings and their translational implications to the cost of clinical services," *Brain Research*, vol. 1380, pp. 22–33, Mar. 2011.

[28] H. Drimalla, N. Landwehr, I. Baskow, B. Behnia, S. Roepke, I. Dziobek, and T. Scheffer, "Detecting Autism by Analyzing a Simulated Social Interaction," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Eds. Cham: Springer International Publishing, 2019, pp. 193–208.

[29] C. Song, J. Li, and G. Ouyang, "Early Diagnosis of ASD based on Facial Expression Recognition and Head Pose Estimation," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2022, pp. 1248–1253.

[30] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 59–66.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462.

[32] M. Yang, J. Konan, D. Bick, A. Kumar, S. Watanabe, and B. Raj, "Improving Speech Enhancement through Fine-Grained Speech Characteristics," Jul. 2022.

[33] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022.

[34] G. Webb, C. Sammut, C. Perlich, T. Horváth, S. Wrobel, K. Korb, W. Noble, C. Leslie, M. Lagoudakis, N. Quadrianto, W. Buntine, L. Getoor, G. Namata, J. Jin, J.-A. Ting, S. Vijayakumar, S. Schaal, and L. De Raedt, "Leave-One-Out Cross-Validation," Jan. 2010.

[35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[36] K. Semmelmann and S. Weigelt, "Online webcam-based eye tracking in cognitive science: A first look," *Behavior Research Methods*, vol. 50, no. 2, pp. 451–465, Apr. 2018.

[37] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS One*, vol. 14, no. 11, p. e0224365, 2019.

[38] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives of General Psychiatry*, vol. 59, no. 9, pp. 809–816, Sep. 2002.

[39] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy (Basel, Switzerland)*, vol. 23, no. 1, p. 18, Dec. 2020.

[40] M. P. Gelfer and V. A. Mikos, "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels," *Journal of Voice: Official Journal of the Voice Foundation*, vol. 19, no. 4, pp. 544–554, Dec. 2005.

[41] H. Wood-Downie, B. Wong, H. Kovshoff, S. Cortese, and J. A. Hadwin, "Research Review: A systematic review and meta-analysis of sex/gender differences in social interaction and communication in autistic and nonautistic children and adolescents," *Journal of Child Psychology and Psychiatry*, vol. 62, no. 8, pp. 922–936, Aug. 2021.