IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 January 2023, accepted 25 March 2023, date of publication 20 April 2023, date of current version 22 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3269027

RESEARCH ARTICLE

Computer Vision-Based Assessment of Autistic Children: Analyzing Interactions, Emotions, Human Pose, and Life Skills

VARUN GANJIGUNTE PRAKASH[1], MANU KOHLI[1], SWATI KOHLI[1], A. P. PRATHOSH[2], TANU WADHERA[3], DIPTANSHU DAS[4], DEBASIS PANIGRAHI[5], AND JOHN VIJAY SAGAR KOMMU[6]

[1]CogniAble, Gurugram, Haryana 122022, India
[2]Department of Electrical Communication Engineering, Signal Processing Building West, Indian Institute of Science, Bengaluru 560012, India
[3]Department of Electronics and Communication Engineering, Indian Institute of Information Technology Una (IIITU), Una, Himachal Pradesh 177209, India
[4]Institute of NeuroDevelopment, Kolkata, West Bengal 700005, India
[5]Jagannath Hospital, Bhubaneswar, Odisha 751007, India
[6]Department of Child and Adolescent Psychiatry, National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, Karnataka 560029, India

Corresponding author: Varun Ganjigunte Prakash (varungp@cogniable.tech)

This work was supported in part by the Biotechnology Industry Research Assistance Council (BIRAC), India, under Grant BIRAC/FITT0528/BIG-13/18; and in part by Social Alpha, India.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Maulana Azad Medical College New Delhi, the All India Institute of Medical Sciences Jodhpur, the National Institute of Mental Health and Neuro-Sciences (NIMHANS) Bangalore, the Vardhman Mahavir Medical College New Delhi, Latur Medical College, and the Central Institute of Psychiatry in Ranchi.

ABSTRACT In this paper, the proposed work implements and tests the computer vision applications to perform the skill and emotion assessment of children with Autism Spectrum Disorder (ASD) by extracting various bio-behaviors, human activities, child-therapist interactions, and joint pose estimations from the recorded videos of interactive single- or two-person play-based intervention sessions. A comprehensive data set of 300 videos is amassed from ASD children engaged in social interaction, and three novel deep learning-based vision models are developed, which are explained as follows: (i) activity comprehension to analyze child-play partner interactions (activity comprehension model); (ii) an automatic joint attention recognition framework using head and hand pose; and (iii) emotion and facial expression recognition. The proposed models are also tested on children's real-world, 68 unseen videos captured from the clinic, and public datasets. The activity comprehension model has an overall accuracy of 72.32%, the joint attention recognition models have an accuracy of 97% for follow eye gaze and 93.4% for hand pointing, and the facial expression recognition model has an overall accuracy of 95.1%. The proposed models could extract behaviors of interest, events of activities, emotions, and social skills from free-play and intervention session videos of long duration and provide temporal plots for session monitoring and assessment, thus empowering clinicians with insightful data useful in diagnosis, assessment, treatment formulation, and monitoring ASD children with limited supervision.

INDEX TERMS Autism spectrum disorder, activity comprehension, facial expressions, joint attention, ASD screening, applied behavior analysis.

I. INTRODUCTION

Children with Autism Spectrum Disorder (ASD) typically exhibit biobehavioral patterns such as repetitive behavior, difficulty in establishing friends, poor social communication abilities, and limited understanding and expression of emotions [1]. Traditional diagnostic methods such as blood tests, genetic testing, and brain imaging have limited success in establishing diagnoses, severity degree, and skill assessments of ASD children.

The associate editor coordinating the review of this manuscript and approving it for publication was Dian Tjondronegoro.

Behavioral methods are the gold standard for diagnosing ASD in children, which entails the physician documenting the patient's medical history, interviewing the parents, and manually observing the children's behavior. These observations are recorded as detailed in the instruction guidelines for diagnostic rating scale instruments such as the Autism Diagnostic Observation Schedule (ADOS) and the Childhood Autism Rating Scale (CARS-2) [2], [3]. The rating scale usually suggests a child's skills in social engagement, joint attention, emotional expressions, instruction following, play and life skills, imitation abilities, and visual attention. The diagnosis is established if the observation scores cumulatively exceed a predetermined threshold. After diagnosis, a functional assessment is conducted utilizing the instruments such as VBMAPP [4] to build a personalized intervention program that can improve the necessary skills of ASD children for their school and societal inclusion. The functional assessment includes detailed observations and measurements of children's skills in various domains such as independent play, social communication, self-stimulatory behavior, joint attention, imitation, and understanding of emotions through facial expressions [5], and other necessary skills of ASD children [6].

However, there are various limitations when using conventional diagnostic and functional assessment methods. Firstly, the interpretative coding of a child's behavior observed is manual and time-consuming. Secondly, a clinician's observations may not always be reliable or valid due to differences in professional training, experience, available resources, and cultural backgrounds. Thirdly, there is a huge demand-supply mismatch between the number of professionals available to treat nearly 2% of newborn children diagnosed with ASD [1]. These challenges are exaggerated in Low and Middle-Income Countries (LMICs) [1], [7], [8] where there is a severe shortage of clinicians and poor infrastructure to manage ASD conditions. Therefore, new technological methods for rapid and automatic data collection and analysis can enhance clinician capacity and improve quality, affordability, and accessibility in ASD detection and assessments.

Technology has demonstrated significant benefits by employing Machine Learning (ML) and Deep Learning (DL) for early diagnosis and functional assessments of ASD [9], [10]. ML has uncovered essential and minimal features [11], [12] of ASD diagnostic instruments such as the Autism Diagnostic Observation Schedule (ADOS) [2], and the Autism Diagnostic Interview-Revised (ADI-R) [13], thereby accelerating the diagnosis procedure without compromising accuracy [14], [15]. ML and DL methods can analyze an unprecedented quantity of multimodal and multidimensional clinical data from videos, images, texts, voice messages, and sensors due to the rapid evolution of technology and digitization [9]. The analysis can suggest patterns, aid in the development of clinical decision support systems to diagnose ASD or developmental delays, and provide suggestions for treatment and personalization, enhancing the clinician's capacity. Earlier studies on ASD screening developed a multimodel approach with video annotation performed by humans [12], [39], [45], however, very little work has been done on the automatic extraction and classification of human actions from untrimmed videos for ASD detection [46]. The state-of-the-art ML and DL methods have improved quality, outcomes, and access to ASD screening, diagnosis [12], and assessments [39]. Researchers have trained supervised learning ML models on multimodal data to develop ASD screening and diagnosis [39] solutions with moderate to high psychometric outcomes in minimal time, ensuring their internal validity. These solutions have focused on detecting children with ASD and ODD [12] on cross-cultural datasets.

In the past decade, computer vision-based behavior imaging and facial analysis have shown promising results in assisting clinicians with the diagnosis of multiple medical conditions including ASD [16], [17], [18]. Moreover, computer vision-based methods can offer an accurate, low-cost, and non-invasive alternative compared to traditional labor-intensive manual assessments and invasive methods such as electroencephalogram (EEG) [19].

Even though computer vision has demonstrated many promising solutions, its application in assessing behavior, play, imitation and life skills, posture, and gait analysis to assess the joint attention of ASD children has not yet been explored [20], [21], [22]. In addition to these, there are no large-scale efforts to develop facial expression recognition models or detect joint attention skills of young children from real-time videos. Therefore, we address these issues by developing novel computer vision models to extract and classify the joint attention skills, facial expressions, and life skills from untrimmed videos of ASD children and assist the clinician in diagnosing ASD or establishing the functional assessment for ASD children.

(i) To assess children's joint attention skills automatically, we developed computer vision models by analyzing postural changes in response to instruction or stimuli given by the clinician.
(ii) To recognize nine emotional expressions, namely anger, disgust, fear, happiness, sadness, surprise, laughter, crying, and neutral for children aged 1 to 5, we developed the Facial Expression Recognition (FER) model by gathering extensive facial images from diverse ethical and cultural backgrounds.
(iii) To perform an automatic functional assessment of children from their intervention video sessions, their engagement duration, and frequency with clinicians, parents, or play partners on ten life skill activities, namely run, sit, stand, engagement, instruction engagement, hit or fight someone, watch someone, hold an object or oblique toys, walk, and answer the phone, are assessed.

The paper is organized as follows: Section II briefly describes state-of-the-art computer vision methods used in ASD management. In Section III, we provide the details of the study procedure, and Sub-section III-A provides a detailed description of the problem and answers the questions

raised in developing video-driven assessments. Section IV describes the data collection procedure and the technological methodology to realize the study aims. Section V provides a detailed evaluation and results of our models on real-time videos; in Section VI, we discuss the results interpretations, practical implications, limitations, and future directions; and in Section VII we provide the conclusion.

II. LITERATURE REVIEW

This section discusses relevant studies, state-of-the-art computer vision methods, implementation challenges, and improvement options. Sub-section II-A summarizes the current state of ASD assessment and intervention methods and new studies incorporating ML and DL models into ASD diagnosis and therapy. Sub-section II-B discusses state-of-the-art computer vision methods deployed for Human Pose Estimation. Sub-section II-C discusses the importance of joint attention skills and data-driven assessment techniques. Sub-section II-D describes the significance of facial expression recognition in ASD assessment and treatment planning. Finally, Sub-section II-E summarizes the state-of-the-art human action recognition methods and their applications in assessment and treatment formulation for ASD.

A. ASD TREATMENT

Most evidence-based ASD intervention methods can enhance the child's ability, especially in the first three years [23]. However, the demand for professionally trained therapists has outpaced the supply; consequently, clinicians' availability and cost-effectiveness are crucial for promoting treatment accessibility. Cognitive Behavioral Therapy (CBT) is a behavioral intervention that can help individuals with ASD to achieve their goals and change their lifestyles [21], [24], [25].

Applied behavioral analysis (ABA) is a gold-standard intervention widely used to assist ASD individuals with behavioral and communication challenges by promoting desirable social behaviors [26] such as overcoming food intolerance, improving intelligence quotient (IQ), social communication, and teaching play and life skills using principles of reinforcement [27]. A higher quality of life for ASD children can be foreseen through early diagnosis followed by evidence-based treatment methods [20], [28]. An accurate diagnostic and functional evaluation is essential to evaluate the child's area of strength for customizing an intervention program to the child's unique needs. ADOS [2], ADI-R [13], The Modified Checklist for Autism in Toddlers, Revised, with Follow-Up (M-CHAT-R/F) [29], and The Childhood Autism Rating Scale-2 (CARS-2) [30] are a few widely used gold-standard ASD diagnostic and screening instruments developed in Western countries [31], [32], [33]. Therefore, the outcomes of these assessments have limited efficacy when employed in LMICs due to a lack of training and cultural disparities [34], [35], [36], [37].

Artificial intelligence (AI) technology especially ML and DL can address these limitations due to its unique facets such as increased processing power of computer hardware and multimodal data availability, thereby leading to faster ASD diagnosis [38]. Recently, the clinical study of multi-modular ML-based ASD diagnosis based on questionnaires and home videos has demonstrated a sensitivity of 90% towards ASD detection [39]. Some of the other improvements that have been witnessed with the application of AI are: (i) Detection of ASD at an early age, (ii) Reduction in the number of assessment items as a result of implementing the feature reduction method, (iii) Effective classification between different ASD, Typically Developing (TD), and other neurodevelopmental disorders, (iv) Automatic feature extraction of bio-behaviors from multimodal data [9], [40].

Due to the availability of multimodal data from diverse bio-behavioral sources, such as videos including ASD behavioral features [12], [41], audio [42], facial expressions [43], and Electronic Health Record (EHR) data [44], DL applications trained on unstructured data have accelerated the detection and management of ASD and can be implemented at the point of care [9], [12], [41], [45], [46], [47]. The feasibility of the therapeutic intervention and prognosis leveraging AI has shown reasonable success [48] for ASD and other neurodevelopmental disorders. Furthermore, individualized socially assistive robotic intervention and automation based on engagement analysis has aided in the development of a low-cost, robot-based therapeutic framework for ASD children [49].

However, most studies focus on one of the seven key data categories, such as stereotyped behaviors, eye gaze, facial expressions, postural analyses, motor movements, auditory, and electronic health records [9] adopting ML and DL techniques with Graphical Processing Units (GPU) and high processing cloud capabilities [9]. To the best of our knowledge, this is the first study to employ computer vision to extract data from various bio-behaviors, including play, engagement, facial expression, and joint-attention abilities.

B. HUMAN POSE ESTIMATION

Computer-vision-based Human Pose Estimation (HPE) methods including conventional and instance-based pose estimation models to novel deep network architectures can detect human body poses in 2D or 3D space by regressing skeletal joint angles or critical points using a single view or several view cameras with monocular or depth modalities [50], [51], [52], [53], [54], [55]. In addition, developing computer vision applications for specific task measurements involves accurate measurements of both human body joints and their parts. Head pose estimation involves the prediction of head orientation and assessing human attention and head pose. Similarly, hand detection and tracking provide a fine-grained estimation of hand posture for regressing skeletal finger points and gesture recognition tasks [50]. However, most pose estimation algorithms are designed for adults or pedestrians, and few

solutions have focussed on special needs children or pediatric healthcare [56], [57]. Several significant constraints prevail in the development and deployment of the HPE methods for various child-specific problems in managing ASD conditions as follows: (i) data security, privacy, and ethical challenges, (ii) expensive data collection, (iii) manual data annotation process, and (iv) camera calibration and setup, and (v) single and targeted solutions to a specific problem [52]. The current methods for human pose estimation are designed to track specific movements and activities and may not be able to capture a broad variety of child behaviors or activities. For instance, head pose estimation may be ineffective for monitoring social engagement or other nonverbal indicators. Human pose estimation methods are not always accurate, particularly when monitoring the movements of children, who have smaller, more rapid movements. Additionally, children are more likely to make sudden, unpredictable movements that can be difficult to accurately monitor. To circumvent these problems, our goal is to develop dedicated models for tracking hands and heads that work well for adults as well as toddlers.

C. JOINT ATTENTION

Joint attention (JA) is a social communication method of engaging one's attention with another person using objects and gestures. Limited JA skills are one of the earliest indicators of ASD which JA necessitates capturing, sustaining, and transferring attention and fostering the growth of essential social abilities, such as engaging with others and understanding their perspective [14]. Similarly, few works implemented a DL classification model for evaluating joint attention in individuals with ASD by utilizing short video clips of joint attention initiation. The system evaluates joint attention aids in the early detection and intervention of ASD. Similarly, a vision-based joint attention detection system for ASD using eye-tracking technology showed good accuracy in detecting joint attention among non-ASD adults. An automated tool called RJAfinder has been developed that quantifies responding to joint attention behaviors in ASD using eye-tracking data. RJAfinder can compare RJA events among ASD children, typically developing children, and adults and finds fewer RJA events that ASD children display than the other two groups. Cazzato et al. [58] have examined how robot-assisted therapy affected the social interactions of children with ASD and used expensive depth cameras to aid in non-invasive JA evaluation. Few studies have used eye-tracking technology to investigate eye gaze accuracy, fixation, eye transition, and eye movement during technology-aided JA assessments [15], [59], [60], [61] methods.

D. FACIAL EXPRESSION RECOGNITION AND EYE CONTACT

The intensity and frequency of eye contact and facial expressions can facilitate verbal and non-verbal communication between individuals. Maintaining eye contact can be distressing for some ASD individuals leading to social anxiety. Subsequently, the capacity to imitate and comprehend facial emotions is crucial for the social functioning of any individual. ASD children have difficulty understanding and responding to nonverbal cues and recognizing and comprehending facial expressions and emotions. Carpenter et al. [43] have extracted positive, neutral, and other facial landmarks from a database of 3D facial expressions utilizing a trained computer vision model and have discovered that children with ASD have more neutral facial expressions, which corresponds to the fact that facial expression imitation is an essential indicator of social interaction skills. Zhao et al. [62] have implemented a DL model to recognize facial expressions by utilizing multiple databases while training it with the facial expressions of sixteen Chinese children. The experimental results of Zhao et al. have shown that the ASD group's average imitation expression is found to be less than 60%, a significant deterministic threshold for ASD.

Alvari et al. [63] have examined facial expressions using the Facial Action Coding System (FACS) and extracted the intricate dynamics of ASD and TD children's social smiles from home recordings. The findings of Alvari et al. have suggested that ASD children exhibit less happiness than TD children in their first years, confirming that ASD children have difficulty distinguishing faces and take a long time to comprehend facial expressions. Deep learning-based facial expression recognition (FER) has been explored in numerous architectures such as convolutional neural networks, deep belief networks, autoencoders, generative adversarial networks, and ensembles of networks. These architectures performed the best on a variety of benchmark datasets as they focused on the two important issues of overfitting and expression-unrelated variations.

E. ACTIVITY RECOGNITION

Activity recognition identifies significant events of interest in vast video datasets [64], [65], [66], [67]. Earlier techniques employed human posture traits [68], feature descriptors [69], and dense trajectories [70] based on the appearance from camera movement. However, ML and computer vision (CV) have improved various aspects of human visual perception to find clinically meaningful patterns from the images and videos and classify activities of interest to diagnose and functionally assess ASD children [12], [71], [72], [73]. However, one of the obstacles associated with applying CV in ASD detection and its management is the high labor cost and downtime associated with the manual annotation of video. Furthermore, due to the computationally intensive description and monitoring of motion data from the real-time feed, activity monitoring can result in low generalizability because of potential tracking failure in non-neural network-based systems. Hence, we propose a novel DL model to address these limitations by training the model on a limited publicly available dataset of action classes relevant to ASD diagnosis and assessments.

Researchers have developed neural network architectures such as two-stream Inflated 3D [74] and C3D [75], [76], which incorporate optical flow and RGB image properties for capturing person features and movements. In addition to the standard bag-of-visual-words method [77], these architectures have strengthened the activity recognition framework. In terms of capturing the temporal patterns and simplifying parameter learning for underlying architectures, 3D convolutional layers are superior to 2D ones [74]. However, these methods work only for short, trimmed videos and do not perform well with longer untrimmed videos containing simultaneous multi-person actions. Temporal action localization-based methods have solved this limitation by slicing a long-duration video into manageable time segments. A single-stage end-to-end network suggests action intervals and also classifies potential actions [76], [78].

In contrast, two-stage methods use distinct neural networks for making suggestions on the occurrence of significant human movements and then classifying those activities [79], [80]. However, these methods perform poorly with videos similar in actions or outside the training data distribution. Also, activity recognition systems must be resistant to occlusions and capable of handling crowded scenarios with multiple people and actions. Several contemporary paradigms simultaneously investigate spatial-temporal characteristics to identify and mark the location of activities. Tublet generation [81], human- and object-centric learning [78], [82], skeleton-based method [83], and graph convolution networks [84] have enabled the incorporation of various human and environmental features for accurate activity recognition. As part of this paper, we aim to develop a general video activity classifier to detect multiple actions of interest in natural videos accurately.

III. STUDY PROCEDURE

1) The children are recruited from SM Learning Skills Academy Pvt. Ltd., India, a special needs clinic, and the National Institute of Mental Health and Neuro-Sciences (NIMHANS), Bangalore, India. All participants' consent is recorded. The children who have already been diagnosed with ASD participate in play-based interactions.
2) The study objectives and details of data capture are explained to potential participants' parents or caregivers, and any doubts regarding the same are clarified. Finally, their consent for data usage is recorded.
3) Video recordings of interactive ABA therapy sessions between a child and the therapist are recorded.

The principles of ABA have demonstrated its utility for promoting learning and behavior change in children with ASD. In the following sections, we describe the general concerns with conventional ABA sessions and outline our primary goals for enhancing three areas of ABA intervention namely general interaction and life skills, emotion recognition, and joint attention analysis.

A. PROBLEM FORMULATION

This study focuses on the following areas,

1) **Activity Comprehension:** Developmental concerns raised by parents are the first reporting point for performing diagnosis or age-level skill assessment. Clinicians engage children in play-based sessions, provide them with various stimuli, and ask them to undertake various activities to test their age-appropriate skills while diagnosing or assessing their functional skills for ASD. These evaluations also include monitoring the children engaged in independent play in an unsupervised session. During post-ASD diagnosis, the children are engaged in intervention sessions where they are taught various motor and life skills and activities of daily living, including child interaction with the parent, toys, or clinicians during the session. These diagnostic, functional assessment, and intervention sessions are manually analyzed by clinicians, where behaviors of interest and actions of interest are analyzed, spending significant man hours to establish the diagnosis and functional assessment. In order to circumvent manual observation and treatment monitoring, there is a need to adopt automatic treatment monitoring and analysis. Therefore, we propose to use computer vision in the video-recorded session to capture children's activity performance and skill levels. With the help of computer vision, massive engagement video data can be utilized to train a system to identify children's engagement. As part of the general engagement, we measure ten activities as listed in Table 1.
2) **Facial expressions:** The facial expressions of children with ASD can significantly differentiate from TD children in response to various external stimuli. Further, children's motivation during assessments and the ABA intervention session system can be tracked by analyzing children's emotions during and after the intervention. The emotion and facial expressions can give insights into a precursor of maladaptive behaviors or unusual responses to visual or auditory stimuli and investigate the potential causes of these responses. In addition, there are limited datasets of ASD children; therefore, we augment existing datasets by collecting publicly available facial expressions of children and adolescents of diverse ethnicity, culture, and geographic location, thereby conforming to fairness and decreasing bias. Also, we intend to identify nine facial expressions in children using a deep learning model.
3) **Joint attention:** The JA assessment involves a human observer who records the child's responses and the observation data leading to longer wait times, requiring additional therapists, and burdening clinicians' work. Now, it is possible to automate this process and provide an assessment report by developing algorithms to process RGB videos recorded on any camera without requiring expensive sensors attached to children or human observers in treatment scenes.

The JA assessment report comprises marked timestamps at which the child's responses to stimuli are observed, enabling them to conduct their intervention plan effectively.

The study procedure of model development, real-world testing, and performance evaluation is illustrated in Fig. 1.

TABLE 1. Ten activities of interest predicted by the model.

Activity classes	Description
Run	A person running
Sit	A person sitting
Stand	A person standing
Engagement	A person talking to someone
Instruction Engagement	A person listening to someone
Hit/Fight someone	A person hitting/fighting someone
Watch someone	A person watching someone
Hold object/Oblique toys	A person holding something
Walk	A person walking
Answer phone	A person answering phone

IV. METHODS

A. DATA COLLECTION AND PROCESSING

The video data of 300 children with ASD diagnosis of varied age groups (1-5 years) are collected. The average duration of collected videos is 20 minutes. The model outcomes are blind-tested on 68 videos (see Section V-A).

The following prerequisites are completed before data collection at each site.

1) Researchers have finalized a set of behavioral markers exhibited prominently by ASD children such as poor eye contact, motor, and joint attention, imitation skills, absence of play, and self-stimulatory behavior to perform diagnosis or functional assessments by referring to literature and expert discussions.
2) We developed a web application used by expert annotators to traverse temporally across a video and perform frame-by-frame annotations on preselected multi-level video clinical landmarks. The annotators performed the annotations by capturing eye contact, joint attention, imitation, and activities of interest, and their respective metadata as features. For instance, depending on the therapist-child interaction, the annotator would add metadata such as ''yes'' or ''no'' for the 'Joint Attention' and 'Social Behavior' categories for the follow gaze, finger pointing, and eye contact classes, and additional information if responses are spontaneous or elicited. Similarly, the annotators mark all the activities of interest (see Table 1) for both the child and the play partner in the videos. Each video's annotations are saved in an XML file containing the child's age, gender, ground-truth diagnosis, audio or video annotations, metadata, and timestamps.
3) Annotators examine publicly available videos collected from YouTube and Vimeo for good resolution of at least 240p before annotation.

B. ACTIVITY COMPREHENSION MODEL

In this section, we provide an overview of our Activity Comprehension model, and technological methods of implementation are explained in the subsequent subsections. We define the ABA enrollment and assessment pipeline as shown in Fig. 2. The flow diagram also shows the steps involved in the real-world deployment of the Activity Comprehension model. These video sessions capture ABA intervention sessions in which the child is engaged with the therapist in a play-based scenario (see Section III). The video is then inferred into our Activity Comprehension model to provide predictions. Fig. 3 shows the plots of temporal events of interest (activities) and an automatic pie chart generation where each point on the child/play partner interaction plot belongs to one of the activity classes (see Table 1) at a given time step.

Spatio-temporal action recognition methods [82], [83], [85], [86], [87] have been developed to train actions simultaneously with spatial and temporal annotations, i.e., where the action is and what the action is about in a given image. However, these methods require dense annotations at the frame level, which is difficult and time-consuming in a clinical setting. Therefore, we utilize open-sourced spatiotemporal transformer approach [88] to analyze person-person interaction. We also aim to understand child-play partner interactions efficiently and thereby provide actionable insights from the ABA video assessments. We propose our Activity Comprehension model approach in Fig. 4 which comprises child-play partner activity prediction and activity separation, described in the following sections.

1) ACTIVITY PREDICTION

The Activity Comprehension model is trained on the AVA dataset [64] which contains 235 training movie videos, 64 validation videos, and 131 test movie videos (see Table 2) from which 10 out of 80 action categories are utilized for our study. The Activity Comprehension model localizes people's actions in both space and time and recognizes actions with a novel asynchronous interaction aggregation method [88]. The action recognition system uses the object detection model [89] for localizing people. However, the object detection model misses detecting several people/children in the dense crowd and occlusions cases. Hence, we utilized Yolo-v5 [90] for real-time person detection in videos with significant improvements over Yolo-v3 [89]. The input to the Activity Comprehension model is an ABA video (approximately 10 minutes in length) containing scenes of engagement/activity between a child and their play partner and the model predicts the activities (see Table 1) of both the child and the play partner temporally. We record and store the detected bounding boxes, model predictions, and time intervals of the child and the play partner. However, it should be noted that the model does not explicitly make distinctions concerning the people in the scene as children or play partners. We only know the number of people in the scene

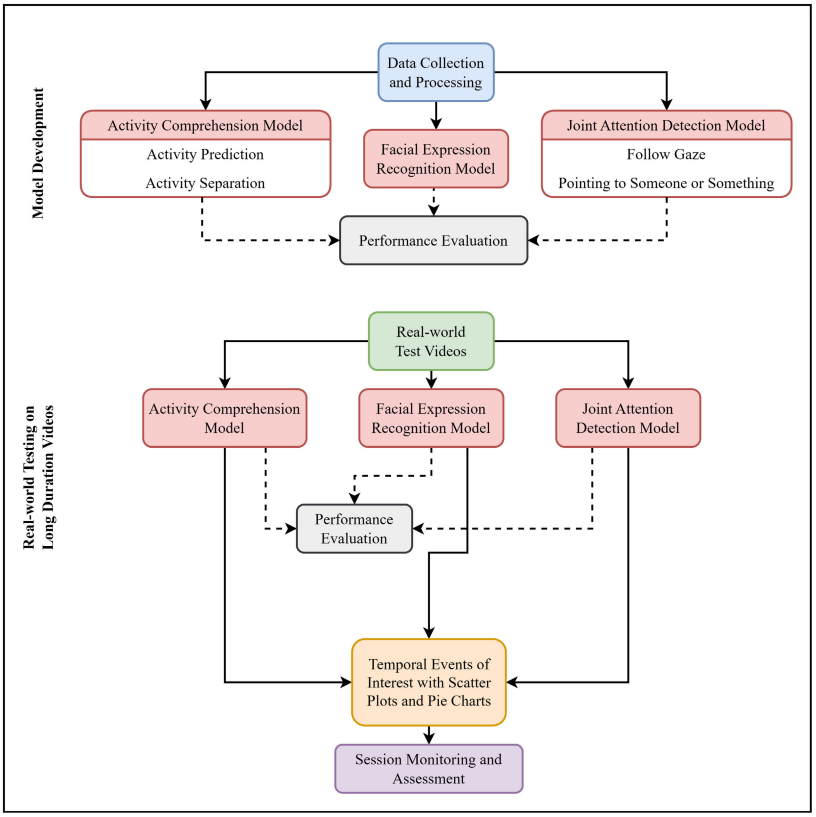

FIGURE 1. An overview of study procedures for model development, real-world testing, and performance evaluation.

(by counting the number of person bounding boxes) and their associated activities. Hence, there is a need to segregate the activities of the child and play partner to make logical conclusions for the assessment by implementing the activity separation method which is explained in the next section.

2) ACTIVITY SEPARATION

To distinguish a child from a play partner, we develop a child detector model. Knowing the child's location is necessary because we infer its actions from the features extracted from subsequent frames. The child detector model is trained using 3027 annotated images of children (see Table 2) by fine-tuning weights from Faster R-CNN with Resnet-50 (v1) [91] to obtain a mAP@0.5IOU score of 0.94. As a result, a child is distinguished from a play partner, and child features are extracted. Table 3 summarizes the hyperparameters used and the results of the detector model evaluation. The child detection model predicts the child's location in each video frame and produces a bounding box for the child. We store the detected bounding boxes along with the time interval.

We perform Intersection over Union (IoU) of detected boxes on the person detection boxes (accumulated child and play partner bounding boxes from subsection IV-B1) and child detection box (from subsection IV-B2). The IoU of two detection boxes is defined as the ratio of the overlapped area to the area of the union of two bounding boxes (Equation 1).

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{1}$$

We compute the IoUs between each pair of axis-aligned bounding boxes which are one IoU between the child detector bounding box and one of the bounding boxes from the person detection bounding boxes. We select the IoU score of ≥ 0.75 as a good threshold for locating the child and perform the following two checks to ensure that the located bounding box is correct out of several other bounding boxes obtained from the activity prediction model. At first, the center coordinates pixel values of both the bounding boxes, having a good IoU match are calculated. Then we check whether the center coordinate lies in either of the two quadrants of the image (distinguished as left and right with the center axis) and compare the quadrants in which the center coordinate lies and record it. Next, we calculate the Euclidean distance d_c between the center coordinates of a good IoU match (Equation 2). The lesser the distance score between two center coordinates, the higher the chance that the person's bounding box encompasses the child. From the experiments, we choose a distance threshold of 20 pixels and then select the person bounding box with a good IoU match that agrees with the quadrant rule and the distance measure. We now know the child's location and, therefore, it is easy to recover

TABLE 2. Data details of models for computer vision-based assessment.

		Training			Testing	
Model Name	Sub-block	Training	Validation /test	Diagnosis	Real world test	Diagnosis
Activity Comprehension	Activity prediction	235 videos	64/131 videos	Unknown	21 from clinic and 27 publicly available videos	Known and Unknown
	Child detector	2538 frames	220/269 frames			
Facial expression recognition	-	300 ASD videos or 51037 frames	3000/2000 frames	Unknown	10 videos	Unknown
Joint Attention	Follow Gaze -					
	Head detection	231719 frames	11089/6302 frames	Unknown	10 videos	Unknown
	Child detector	2807 frames	215/252 frames			
	Head pose estimation	127967 frames	10428/25483 frames			
	Pointing to someone or something	18724 frames	964/4681			

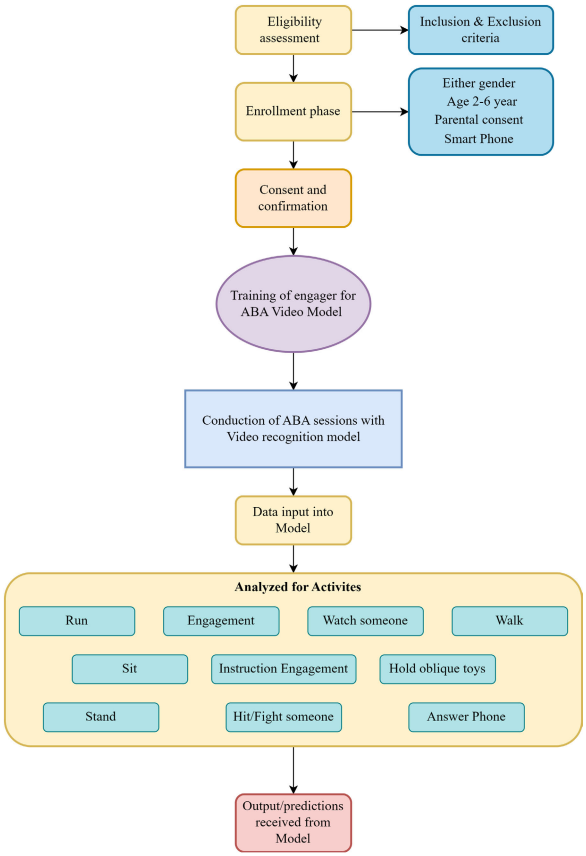

FIGURE 2. Flow diagram of ABA enrollment and assessment using activity comprehension model.

the predictions and time intervals from the person detection list obtained from the action prediction model. Similarly, we recover activity predictions and time intervals from all non-child detection boxes belonging to the play partner. With these results, we can explain an automatic ABA activity assessment in detail in the next section.

TABLE 3. Child detector model hyperparameters and results.

Metric	Name/Value
Similarity calculation	IOU
Regularizer	L2
Initial learning rate	003
Momentum optimizer value	0.90
mAP@.50IOU	0.9458
Box Classifier Loss/classification loss	0.0945
Box Classifier Loss/localization loss	0.07822
RPN Loss/localization loss	0.04173
RPN Loss/objectness loss	0.115
Total Loss	0.03141
AR@100 (Detection Boxes Recall)	0.704

The distance between the two center coordinates

$$(c_{x_1}, c_{y_1}), (c_{x_2}, c_{y_2}) = d_c = \sqrt{(c_{x_1} - c_{x_2})^2 + (c_{y_1} - c_{y_2})^2} \tag{2}$$

where (c_{x_1}, c_{y_1}) and (c_{x_2}, c_{y_2}) are the center coordinates of two bounding boxes.

3) ABA VIDEO ASSESSMENT

We collected 21 videos from Clinic (No. 3.05/30th Institutional Ethics Meeting of Behavioral Sciences Division, National Institute of Mental Health and Neuro-Sciences (NIMHANS), Bangalore, India - Approved on 26/06/2021), and 27 publicly available ABA videos for testing our ABA Activity Comprehension model on one NVIDIA V100 GPU, and predictions of videos are in the form of output plots as illustrated in Fig. 4. The ABA Activity Comprehension

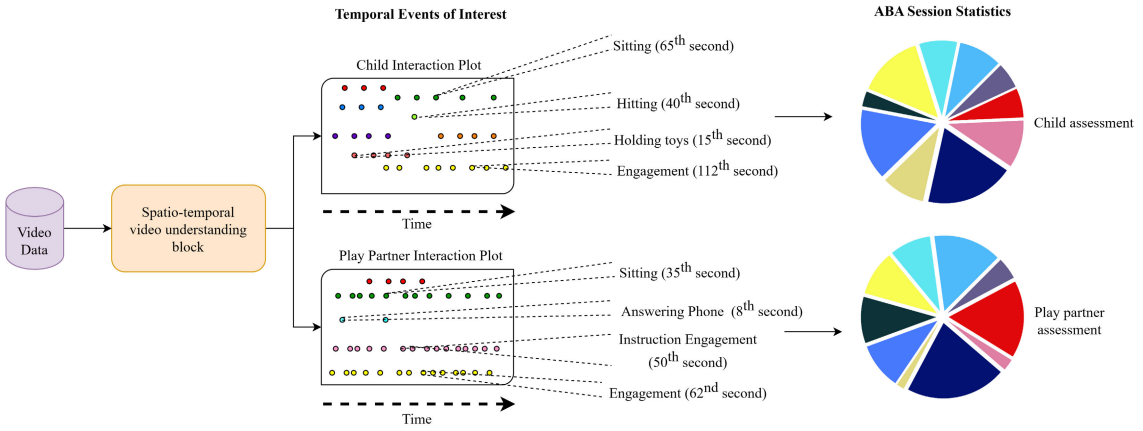

FIGURE 3. Session monitoring and assessment with activity comprehension model.

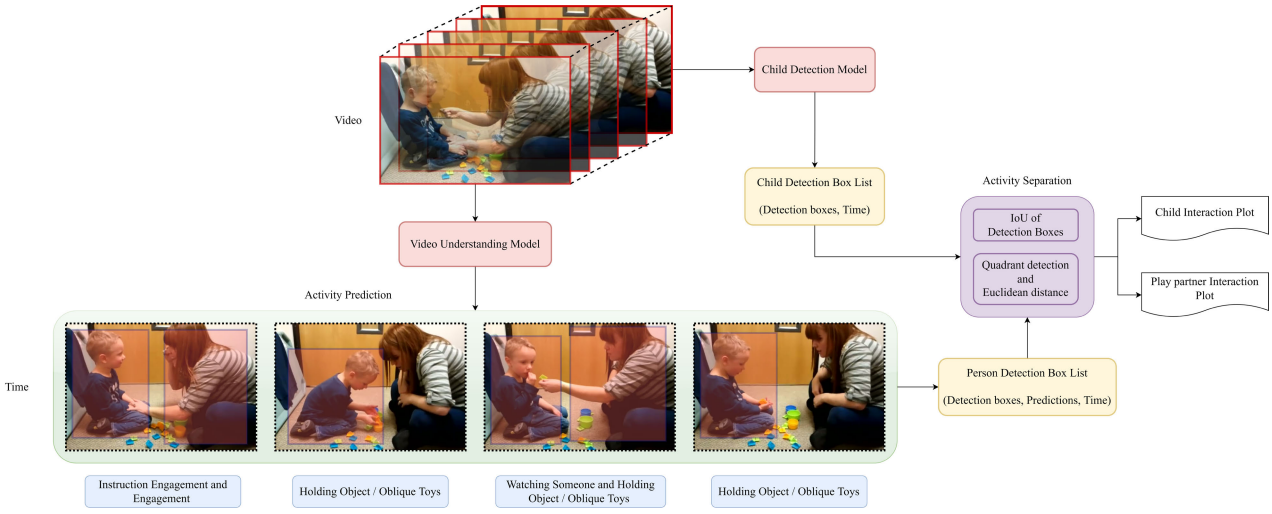

FIGURE 4. Implementation of spatio-temporal activity comprehension model for ABA assessment.

model is used to understand a child's interactions, activity level, and attention with a play partner by analyzing ten activities, including running, sitting, standing, engagement, instruction engagement, hitting/fighting someone, watching someone, holding objects/oblique toys, walking, and answering the phone. Any user who wants to assess engagement and understand the ABA session can upload an ABA video and gets predictions. These predictions are different activity classes predicted by a spatiotemporal action recognition model (subsection IV-B1 and IV-B2). We get predictions for the entire video length with specific activities shown as a scatter plot at time intervals of seconds. The model can efficiently assess videos from a camera with a tripod stand in clinical sessions and from a mobile camera without a stable tripod stands from home videos. The ABA sessions with the Activity Comprehension model will help the learners concentrate on learning goals and the clinicians or therapists to augment their decisions on ABA session outcomes.

The model also assesses the play partner and learner for various activities shown in Table 1. A video can be uploaded to analyze any of these activities, and a scatter plot of engagement and non-engagement with time intervals would be generated for the respective inputs. This scatter plot tells crucial information with a frame-by-frame analysis of the learner and the play partner's activities during the session. Each point shows the action class with a timestamp. The discontinuities in the scatter plot indicate the time intervals of no particular interest to the ABA outcomes. Therefore, these scatter plot patterns can be used to trace the success of therapy and intervention delivery by the therapist easily. The assessment recordings and scatter plot predictions of the model can predict a child's skill level, leading to the line of treatment. For instance, the therapist would be prepared to deal with a violent child if the model had predictions of hitting activity (see Fig. 5). The therapist can then work closely on various attributes of the child's behavior with prior predictions reported from the model. The scatter plot not only

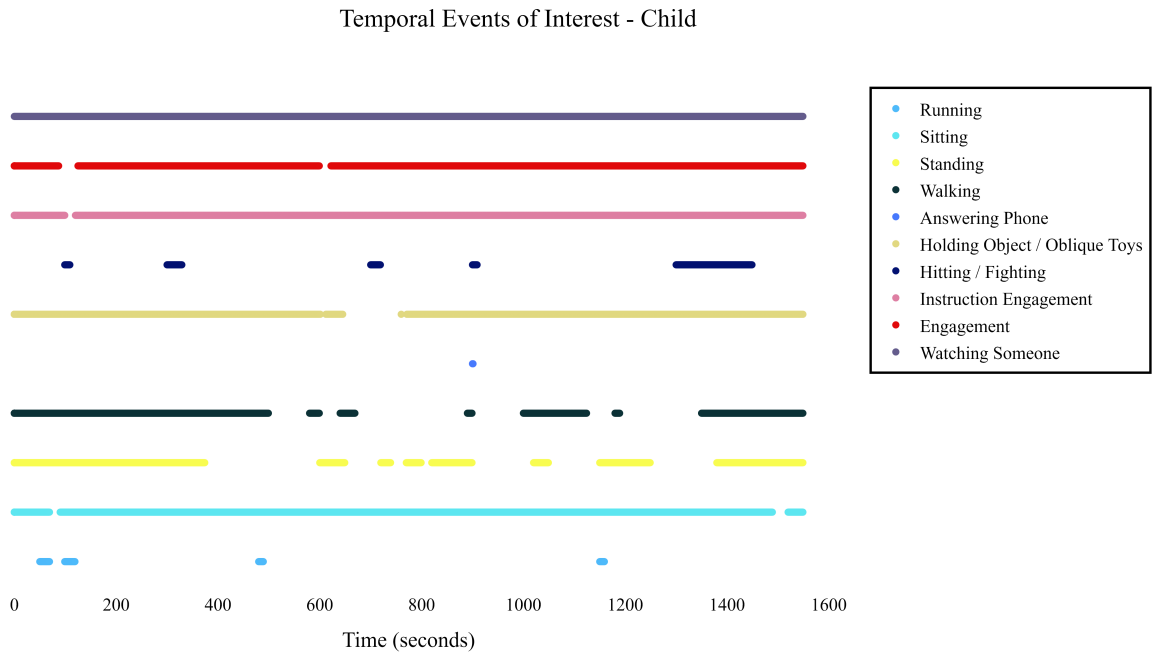

FIGURE 5. Scatter plot of a child indicating hitting and running behavior.

highlights the presence of a particular class of activity but marks the absence (if any) in a learner's or therapist's behavior. In a child showing less attentive behavior towards the play partner's commands (i.e., lower engagement), the therapist can be proactive to work on weaker behavior attributes, strengthen them in the upcoming sessions, and track progress for specific class activities. In this way, the ABA video activity recognition model is of great clinical and therapeutic significance.

The activities such as hitting, running, walking, and repetitive behavior indicate a low level of attention by the learner towards commands of the play partner. Thus, with a plot indicating such results, the customization and prognosis of the therapy can be decided (see Fig. 6). Similarly, if the plot of the play partner/therapist points towards using a phone, the play partner could be replaced or evaluated for their actions accordingly. Fig. 7 shows the pie charts of percentages of activities for a child and therapist after conducting the ABA session. To conclude, the model aids in diagnosis, prognosis, customization of the therapy, and evaluation of the learner and therapist's progress/activities and so it is vital for the ABA sessions and treatment of ASD children.

C. FACIAL EXPRESSION RECOGNITION MODEL

The experiments are conducted on the well-known FER2013 public dataset which has a collection of 35887 greyscales (48×48) images in total [92] and other publicly accumulated datasets. In the FER2013 dataset, each image gathered by the Google image search API is labeled with one of seven categories: anger, disgust, fear, happy, sad, surprise, and neutral. However, the dataset contains several faulty samples (e.g., non-face photos or images with faces wrongly cropped), and the distribution of images among emotion categories is not uniform. There are almost 6,000 photographs depicting happiness, but only about 500 depict disgust. Additionally, as none of the datasets had images of teenagers and toddlers crying or laughing, we, therefore, compiled images from the popular action recognition datasets (Kinetics [67], Moments in Time [93], HMDB [66]), and from our video dataset with 300 ASD children. We accumulate 9882 images of toddlers and teenagers crying and 10268 images of them laughing. The final enhanced dataset contains 51037 training images, 3000 images for validation, and 2000 images for testing (see Table 2). We trained a Resnet-34 backbone-based facial expression recognition model for nine output expression classes namely anger, disgust, fear, happy, sad, surprise, laugh, cry, and neutral. We train a residual masking network of four primary residual masking blocks. Each Residual Masking Block is comprised of a Residual Layer and a Masking Block which acts on different feature sizes. A 3×3 convolutional layer will first process a 224×224 input image with stride 2 followed by a 2×2 max-pooling layer, resulting in a spatial size reduction to 56×56. The corresponding forward layers of four residual masking blocks generate feature maps of four spatial sizes (56×56, 28×28, 14×14, and 7×7) from the feature maps produced by the preceding pooling layer. The network ends with an average pooling layer and a 9-way fully-connected softmax layer producing outputs for 9 facial expression classes (''Angry,'' ''Sad,'' ''Fear,'' ''Happy,'' ''Surprise,'' ''Cry,'' ''Disgust,'' ''Laugh,'' ''Neutral''). The model is trained for 250 epochs with a batch size of 48, with the SGD optimizer, with a

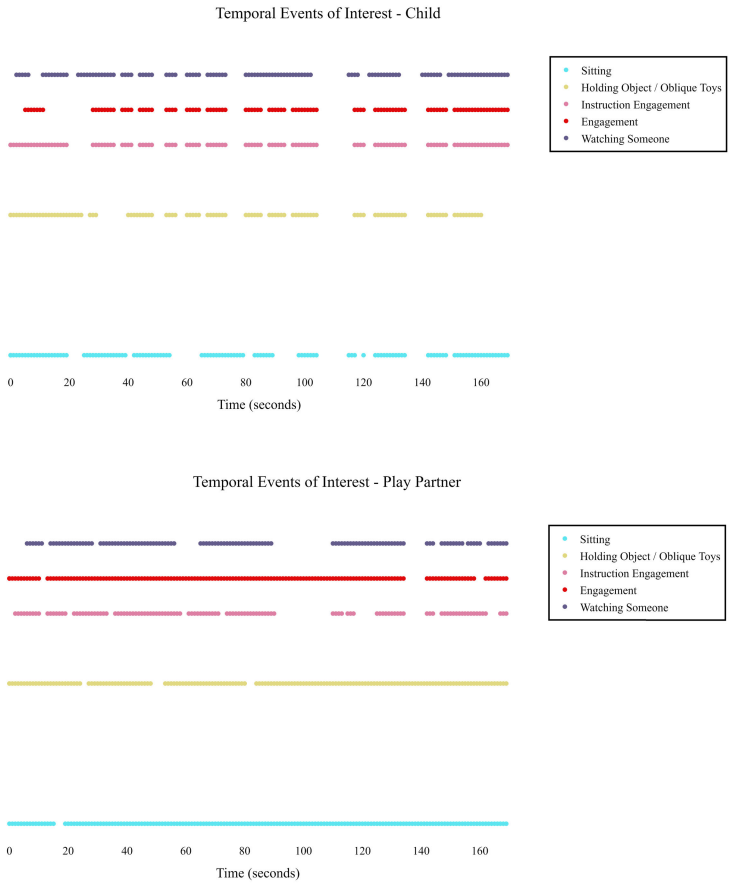

FIGURE 6. A regular output scatter plot of child and play partner to analyze per session response.

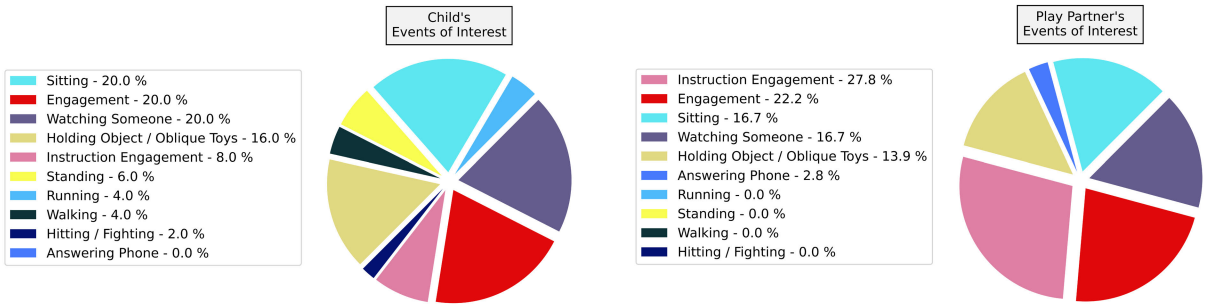

FIGURE 7. Assessment output of child and play partner interactions.

learning decay rate of 0.9 and weight decay of $5e^{-4}$. The learning rate is set to 0.001. Before the training process, the original training images are resized to 224 × 224 and transformed to RGB to support ImageNet pre-trained models. In addition, the training photos are enhanced to prevent overfitting. The dataset augmentation techniques included left-right flipping, brightness variation, and rotation inside the interval [-30, 30] degrees.

The accuracy measure is the evaluation metric for the classification tasks (see Appendix Table 13). The accuracy of the model on the validation set images is 74.4% and the accuracy of the model on the test images is 73.9%. The confusion matrix is shown in Fig. 8. The classes with the highest scores are Happy, Sad, Surprise, and Neutral, while those with the lowest scores were Laughing, Fear, and Disgust. The evaluation scores per class for the 20 ASD test videos are listed in Table 4.

The model is deployed on a Linux server with one NVIDIA V100 GPU to provide real-time facial expression assessment. Fig. 9 illustrates a child's monitoring and assessment

TABLE 4. Evaluation metrics per class on the test set of the facial expression recognition model.

Metric	Classes								
	Angry	Disgust	Fear	Happy	Sad	Cry	Surprise	Neutral	Laugh
Precision	65.22	68.81	62.03	84.62	95.56	71.54	81.41	79.63	50.18
Recall	65.22	68.81	62.03	84.62	95.56	71.54	81.41	79.63	50.18
Accuracy	97.57	99.06	97.21	95.89	95.83	96.08	95.25	98.48	92.38
F1 Score	65.22	68.81	62.03	84.62	95.56	71.54	81.41	79.63	50.18
Specificity	98.74	99.52	98.55	97.63	96.07	97.90	97.28	99.21	95.88
Negative Predictive Value	98.74	99.52	98.55	97.63	96.07	97.90	97.28	99.21	95.88

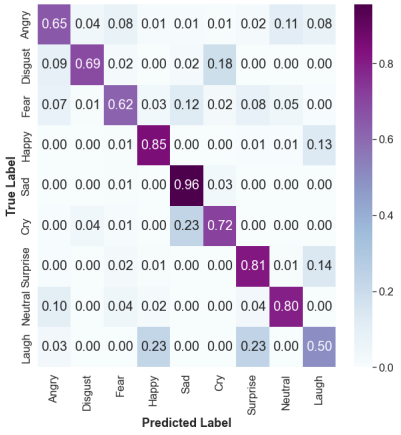

FIGURE 8. Confusion matrix scores on the test set of the facial expression recognition model.

approach. The FER model processes the video data stream and outputs predictions of emotions along with time stamps of their occurrence in a scatter plot. The therapist can view the emotional assessment of the child using a pie chart, as shown in Fig. 10. During our experiments, it is observed that the detection of laughing or crying is more accurate for numerous consecutive frames over a short duration than for a single frame. The confidence of the network predictions for sad/cry and happy/laugh is visually similar if a single frame is processed. Hence, we grouped 20 consecutive frames of the images and determined whether the confidence of the crying or laughing class increased during subsequent frame predictions, indicating greater and longer persistence of sadness or happiness. Given the volatile nature of emotions in children with ASD, understanding of the child's emotions enables better planning of activities suited to the child's needs and remote management and monitoring of the children with ASD.

D. JOINT ATTENTION DETECTION MODELS

Fig. 11 illustrates the methodology of our joint attention detection model. We implemented two deep neural network models for two crucial types of joint attention that therapists frequently use to differentiate ASD from typically developing children. First, joint attention follow-gaze is an activity in which the therapist indicates to a child to look at somewhere or something with his or her eyes, and the child must then follow the therapist's gaze. Successful joint attention indicates that the child actively responded to the therapist's gaze by turning their head and trying to look in the same direction. Another critical skill assessment in JA is hand pointing.

1) JOINT ATTENTION – FOLLOW GAZE

The task of determining the orientation of people's heads from images or videos is known as Head Pose Estimation (HPE). HPE has garnered a lot of attention [94], [95] with the development of a variety of methods such as appearance templates, detector arrays, geometric, regression models, tracking, and hybrid methods. In low-resource clinical settings, the use of costly, complex sensor systems such as magnetic sensors, inertial sensors, lasers, and optical motion capture systems makes it difficult not only to collect ground truth data for developing effective HPE methods but also makes it impossible in the deployed applications without these sensor systems. However, RGB video cameras are cheap, easy to use, and can be installed in clinics. The data collected from the cameras provide novel paths to explore HPE methods that can work on real-world videos. Existing methods are effective for frontal views but not always for head poses from all angles. Additionally, we need to identify the full range of yaw angles for follow gaze detection. Full-range yaw estimation is significantly less common than narrow-range estimation, as most known HPE datasets concentrate mainly on frontal to profile views. To determine yaw, recent approaches classify poses into coarse-grained bins/classes [96], [97], [98]. The limitations of using these existing methods are mentioned as follows: they do not predict pitch and roll; full-range yaw estimation is not robust to occlusions; unreliable pose estimation in heavy noise and jitter environments; low-resolution video; and utilization of multi-camera over monocular images. Therefore, we aim to implement a deep network to predict Euler angles over a single RGB image's entire range of head yaws.

In our clinical setting, a play partner interacts with a child. One of the goals of ABA therapy is to improve a child's joint attention (JA) skills. A child is considered to have good JA follow gaze skills if they respond and actively interact with their play partner's stimuli. For the JA follow gaze stimulus, the play partner performs a subtle gesture with their eyes or pointing hand, asking the child to look towards a direction, typically behind or beyond the sight of the child, so that the effectiveness of the JA follow gaze can be observed. As a response to the stimulus, an attentive child would turn their

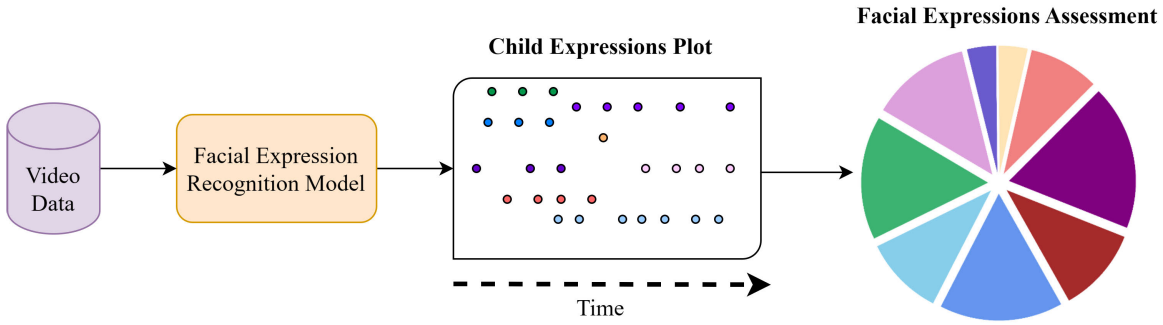

FIGURE 9. Session monitoring and assessment with facial expression recognition model.

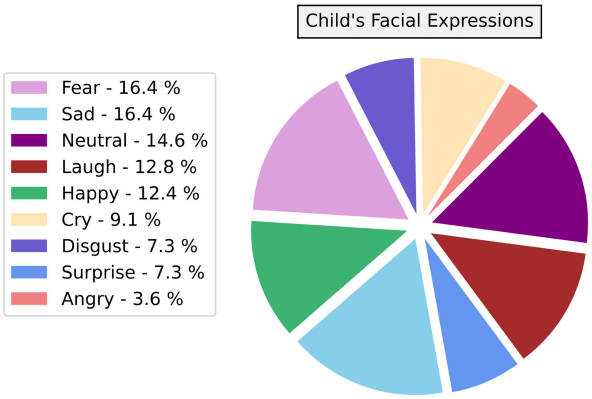

FIGURE 10. Assessment output of child's emotions.

head and look in that direction, following their play partner's gaze. We automatically identify and mark these key JA scenes in the video. To solve the problem of detecting the follow gaze, initially, we need to identify people's heads, track the Euler angles predicted by the pose estimation model, and then detect a full-range yaw rotation. Fig. 11 illustrates our Joint Attention Follow Gaze block. The input video consists of a play-based interaction of a child with a play partner. First, head detection datasets [99], [100] are used to train a YOLO-v3 object detector [89]. With the location of the spatially identified heads, we track the Euler angles with a model trained on the network that predicts pitch, yaw, and roll using a multi-loss metric [101]. In computer vision, Euler angles are commonly used to define any three-dimensional object's rotation by combining three sequential elemental rotations along three distinct axes (X, Y, and Z) which represent the rotation by three parameters: yaw, pitch, and roll. In human head posture, roll computes the amount of X-axis rotation. In head movement, it is the same as moving your head to the left or right. The amount of rotation about the Y-axis is computed via pitch which is comparable to glancing up or down at a person's head. Yaw determines the amount of rotation about the Z-axis. For the human head, it might be interpreted as either looking left or right. Fig. 11 illustrates head pose predictions of successive images. The pose predictions help track a human rotating through a complete revolution of yaw but do not distinguish between a child, a play partner, or other people in the video.

Since we are primarily interested in the full-range yaw rotation of the child, we develop a child detector model using 3274 annotated images of children (see Table 2) by fine-tuning weights from Faster R-CNN with Resnet-50 (v1) [91]. Table 3 summarizes detector model training parameters and evaluation. The child detection model predicts the child's location in each video frame and generates a bounding box for the child. We store the detected bounding boxes alongside the time interval and compare whether the x and y coordinates of the Origin (O) of the yaw, pitch, and roll axes are within the child detector's bounding box. We collect all the yaw values with the child's timestamps for the entire video duration. We develop an algorithm for detecting the rotation based on yaw values. The sign change in the yaw angles is a clear indication of the rotation of the head from left to right or vice versa when a child follows the gaze of a play partner. Upon analyzing each frame of the video, a sign change from positive to negative or vice versa indicating that the child turned from one side to the other. The magnitude of the sign change indicates the degree or magnitude of rotation (a smaller magnitude of change indicates a slight head turn and a larger magnitude indicates a complete head turn). We record the time interval during which a child responds to a follow-gaze stimulus, and detecting a change in the child's head position indicates a successful follow-gaze response.

2) JOINT ATTENTION – POINTING TO SOMEONE OR SOMETHING

One of the goals of the JA assessments is to improve a child's JA hand-pointing skills to understand how well they recognize and interact in their environment. For the JA hand-pointing stimulus, the play partner verbally asks the child a simple question to elicit an expected response of pointing fingers toward a particular direction, typically in front or behind the sight of the child, so that the effectiveness of JA finger-pointing can be observed. As a response to the stimulus, the child points a finger in that direction to answer the question. For example, the play partner might ask, "Where is

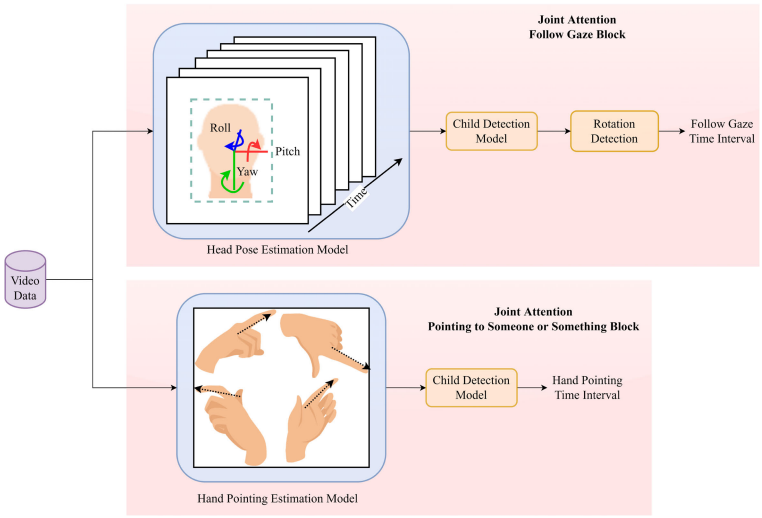

FIGURE 11. Illustration of joint attention models with follow gaze and hand pointing blocks.

your mother?'', ''Where is the toy?'' or ''Where is the animal picture in the book?'' and the child will either point with their index finger and a closed thumb finger or point with their index finger and an open thumb finger. Fig. 11 illustrates different orientations of finger pointing in the Joint Attention Pointing to someone/something block. We collected only hand/finger pointing from various hand gestures and hand-pointing annotated datasets [102], [103], [104], [105]. We trained a finger-pointing detector using 24369 annotated images with their bounding boxes (see Table 2) by fine-tuning weights from Faster R-CNN with Resnet-50 (v1) [91] to obtain a mAP@0.5IOU score of 0.917 and an average recall of 77%. Table 5 summarizes the hyperparameters used and the results of the hand-pointing detector model evaluation. Whenever the child points to someone or something, the model detects it, and we record all the time stamps of the detection in the entire video.

TABLE 5. Finger pointing detector model hyperparameters and results.

Metric	Name/Value
Similarity calculation	IOU
Regularizer	L2
Initial learning rate	005
Momentum optimizer value	0.90
mAP@.50IOU	0.917
Box Classifier Loss/classification loss	0.081
Box Classifier Loss/localization loss	0.095
RPN Loss/localization loss	0.034
RPN Loss/objectness loss	0.072
Total Loss	0.052
AR@100 (Detection Boxes Recall)	0.77

V. RESULTS

In this section, we describe our experiments conducted on previously unseen test data, along with evaluation metrics and results.

A. TEST DATA

The procedure for clinical and publicly available data processing for all the models during testing is the same as described in Section IV-A.

1) Activity Comprehension model: 48 Videos (21 videos from Clinic, 27 publicly available)
2) Facial Expression Recognition model: 10 Videos
3) Joint Attention Recognition model: 10 Videos

All the videos are collected and annotated at the frame level and these human annotations will be used as ground-truth labels for comparison of the results.

B. FACIAL EXPRESSION RECOGNITION MODEL

Ten videos of children exhibiting any of the nine facial expressions were gathered. The facial expression recognition (FER) model, when inferred, stores the predictions for each video frame. For each frame, we compare the model prediction to the ground truth. A prediction is considered as correct if it is of the same class as the ground truth and has a confidence level of at least 85%. The confusion matrix of the FER model is shown in Fig. 12. Table 6 shows the evaluation scores per class for all ten videos. The class ''None'' comprises the frames of the background and the times when no face is visible. From the experimental results, it is observed that the accuracy is above 93% for all the expression categories.

C. ACTIVITY COMPREHENSION MODEL

1) CLINICAL AND PUBLICLY AVAILABLE ABA VIDEOS

Twenty-one recordings of children participating in ABA sessions are collected and evaluated using computer vision (Activity Comprehension model) for the activities mentioned in Table 1. The videos are inferred utilizing the Activity Comprehension model, and the predictions for each video frame are compared to the ground truth. The prediction is

TABLE 6. Evaluation metrics per class on unseen test videos of children using the facial expression recognition model.

Metric	Classes									
	Angry	Disgust	Fear	Happy	Sad	Cry	Surprise	Neutral	Laugh	None
Precision	64.29	55.26	16.39	92.98	78.08	41.84	41	96.12	7.84	100
Recall	67.74	55.26	25.32	90.93	77.03	34.45	19.25	87.12	60	100
Accuracy	94.38	97.06	93.04	97.49	94.29	94.16	93.01	96.97	93.56	100
F1 Score	65.97	55.26	19.90	91.94	77.55	37.79	26.20	91.40	13.87	100
Specificity	96.71	98.48	95.43	98.72	96.83	97.40	97.19	99.20	93.85	100
Negative Predictive Value	97.16	98.48	97.31	98.31	96.63	96.48	92.22	97.14	99.63	100

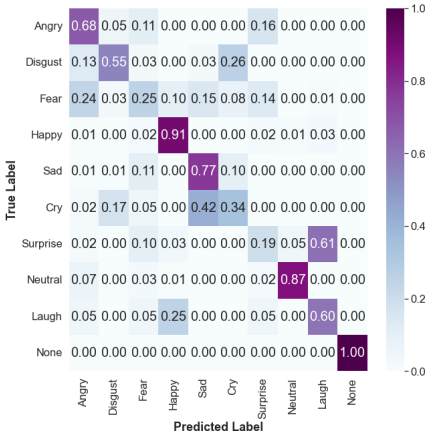

FIGURE 12. Confusion matrix scores on unseen test videos of children with the facial expression recognition model.

considered correct if the class prediction is similar to the ground truth and has a confidence level of ≥ 50 percent. The evaluation scores per class for all 21 videos of the children are listed in Table 7. The model can detect most of the children's activities with at least 70% accuracy. Precision, recall, and accuracy are highest for actions such as sitting and holding objects or oblique toys for children. The per-class metrics corresponding to the play partners in the videos are shown in Table 8. Since the child or the play partner may have more than one label class (sitting and instruction engagement, sitting and holding objects) at each instant, it is necessary to evaluate the temporal alignment of different actions against the ground truth. We define the temporal Intersection over Union (t-IoU) metric [106] for evaluating the action prediction metrics considering each class's start and end of time instants. t-IoU is the sum of the intersections between the predicted and observed time intervals for each predicted class, divided by the sum of their unions. The true-positive t-IoU threshold is set at 0.30. mt-IOU@0.3 is the mean t-IOU across all test videos with a threshold of 0.3 or above. The mt-IoU@0.3 metric severely penalizes any temporal misalignment when examining the model's true performance. Many false positives and false negatives are observed, and some actions, such as hitting and answering the phone, are not captured on video, and their evaluations are represented as null (-).

Similarly, we test the performance of twenty-seven publicly available ABA videos of children collected from YouTube and Vimeo video search engines. Unlike the previous videos, these videos do not conform to play-based activities and include multiperson interactions other than child and play partner interactions, and significantly vary from the training video samples of the AVA dataset [64]. Hence, testing our Activity Comprehension model on these videos will also be helpful to evaluate the robustness of adversarial attacks and provide insights to address out-of-distribution detection. Tables 9 and 10 show the evaluation scores for each class of the children and play partners across all the videos. The model identifies potentially dangerous behaviors, such as hitting or fighting, and performs best when recognizing children's actions, such as holding objects, sitting down, and following instructions and is also capable of determining whether a play partner uses a phone, how engaged they are with children, and how they interact with toys and objects.

D. JOINT ATTENTION

1) PUBLICLY AVAILABLE VIDEOS

We gather ten videos of children participating in ABA sessions from publicly accessible videos of children on JA gaze following and finger-pointing skills. Table 11 lists the evaluation scores of joint attention (follow gaze and joint attention - pointing to someone or something) models for all ten videos. The experimental results reveal that both models of joint attention are accurate to millisecond levels in terms of detecting the instant at which joint attention is successful. The majority of errors are false negatives caused by different configurations of the finger pointing away from the camera's line of sight.

E. EVALUATION ON BASELINE METHODS

We evaluate the performance of our FER model on the most competitive well-known baselines on our enhanced image data with 2000 images used in our test set and on the public dataset FER2013 test images. Since the FER2013 does not have the cry and laugh classes, both classes are mapped to sad and happy, respectively, for a fair comparison. Our model outperforms all the models in the baselines on our test set with an accuracy of 73.9% followed by the Ensemble 8-CNN framework with 73.50% and Ensemble ResMasknet with 73.20%. Additionally, our model is competitive with ensemble-based methods for the FER2013 dataset. Table 12 lists the baseline methods and compares the accuracy between our test set

TABLE 7. Evaluation metrics per class of children on the clinical test set using the activity comprehension model.

Metric	Classes									
	Run	Sit	Stand	Engagement	Instruction Engagement	Hit/ Fight someone	Watch someone	Hold object /Oblique toys	Walk	Answer phone
Precision	73.52	88.54	62.43	76.20	82.46	-	76.40	82.15	92.42	-
Recall	67.80	92.31	54.06	88.54	85.50	-	68.45	74.60	88.65	-
Accuracy	72.83	93.70	50.39	75.90	89.77	73.20	59.30	80.40	90.32	42.70
F1 Score	70.54	90.38	57.94	81.91	83.95	-	72.21	78.19	90.50	-
Specificity	80.40	95.47	67.32	76.24	85.64	74.26	58.56	68.70	87.41	65.28
Negative Predictive Value	82.54	94.60	63.70	77.38	78.80	70.05	65.50	72.82	91.20	64.04
mt-IoU@0.3	43.54	79.89	55.22	32.17	37.6	-	22.57	64.4	31.4	-

TABLE 8. Evaluation metrics per class of play partners on the clinical test set using the activity comprehension model.

Metric	Classes									
	Run	Sit	Stand	Engagement	Instruction Engagement	Hit/ Fight someone	Watch someone	Hold object /Oblique toys	Walk	Answer phone
Precision	-	94.50	71.40	89.35	64.35	-	65.46	86.52	90.02	-
Recall	-	88.34	66.59	92.30	57.81	-	68.3	81.10	88.46	-
Accuracy	72.19	85.18	59.14	87.50	55.40	61.21	59.16	80.40	92.50	65.29
F1 Score	-	91.32	68.91	90.80	60.90	-	66.85	83.72	89.23	-
Specificity	66.53	93.40	82.07	77.48	74.62	54.28	73.50	74.70	82.63	56.46
Negative Predictive Value	71.84	89.96	78.47	81.95	62.92	78.43	77.46	62.67	84.56	87.36
mt-IoU@0.3	-	84.32	44.2	50.85	31.56	-	35.5	57.3	37.21	-

TABLE 9. Evaluation metrics per class of children on the publicly available videos using the activity comprehension model.

Metric	Classes									
	Run	Sit	Stand	Engagement	Instruction Engagement	Hit/ Fight someone	Watch someone	Hold object /Oblique toys	Walk	Answer phone
Precision	86.20	95.30	67.25	88.36	90.94	92.35	76.74	93.20	73.60	-
Recall	77.47	91.20	70.89	88.00	88.61	90.49	79.11	90.51	77.90	-
Accuracy	76.32	88.71	88.24	76.90	89.37	95.36	80.49	94.62	65.04	67.80
F1 Score	81.60	93.20	69.02	88.18	89.76	91.41	77.91	91.84	75.69	-
Specificity	81.65	82.90	77.72	87.38	91.60	92.60	84.52	90.02	68.24	52.75
Negative Predictive Value	83.39	84.40	75.35	83.60	83.62	89.06	80.64	89.82	72.50	78.18
mt-IoU@0.3	42.64	52.82	47.1	39.5	60.41	73.2	51.21	69.5	32.6	-

TABLE 10. Evaluation metrics per class of play partners on the publicly available videos using the activity comprehension model.

Metric	Classes									
	Run	Sit	Stand	Engagement	Instruction Engagement	Hit/ Fight someone	Watch someone	Hold object /Oblique toys	Walk	Answer phone
Precision	-	88.40	75.05	82.60	78.30	-	72.40	94.50	68.00	76.40
Recall	-	90.31	71.30	80.80	62.75	-	70.05	95.22	75.40	61.15
Accuracy	67.90	90.30	77.10	86.32	69.50	64.20	74.82	91.10	69.52	82.43
F1 Score	-	89.34	73.13	81.69	69.67	-	71.21	94.86	71.51	67.93
Specificity	72.21	85.04	80.60	75.50	73.30	57.70	64.80	95.26	70.80	88.35
Negative Predictive Value	80.45	83.25	81.26	85.17	79.54	75.40	86.30	92.78	73.60	75.4
mt-IoU@0.3	-	62.6	27.32	57.6	32.42	-	42.4	55.25	37.5	31.5

and the FER2013 dataset. The baseline implementations are evaluated in TensorFlow, and the rest of them are implemented with the code provided in the studies. The method Ensemble ResMasknet outperforms with 76.82% accuracy, followed by the method CNNs and BOVW with SVM having 75.42% followed by Ensemble 8 CNN-based method with an accuracy of 75.20% and our model with an accuracy of 74.15%.

TABLE 11. Evaluation metrics of joint attention models on ten publicly available videos.

Metric	Model	
	Follow Gaze	Pointing
Precision	97	92
Recall	95.50	90
Accuracy	97	93.4
F1 Score	96.24	90.90
Specificity	98	97
Negative Predictive Value	96	95

TABLE 12. Evaluation of the performance of the FER model using well-known methods on 2000 images in our test set and test set of FER2013 dataset.

Method	Accuracy (%)	
	Our test set	FER2013
Resnet152	71.42	73.22
Cbam_Resnet50	72.40	73.39
VGG19	67.50	70.80
Densenet121	70.16	73.16
ResAttNet56	72.45	72.63
LHC-Net	72.70	74.42
Deep-Emotion	67.40	70.02
DL-SVM	70.15	71.16
Ensemble 8 CNNs	73.50	75.20
Ensemble ResMasknet	73.20	76.82
CNN-SIFT	71.50	73.40
CNNs and BOVW with SVM	71.82	75.42
Ours	73.90	74.15

For the activity comprehension model, we only utilized 10 of the 80 action categories, and we are only interested in two-person interactions. Since popular baseline models evaluate multi-person interactions and also due to resource constraints, we only performed the evaluation on 48 real-world videos, and these results are listed in Tables 7-10. More details on competitive baselines can be found in detail in the study [88].

Joint Attention Follow Gaze detection requires a pipeline of multiple models involving head detection, child detection, and finally head pose estimation via Euler angle prediction; therefore, there is no direct comparison in the clinical literature for a fair evaluation. We are only interested in real-time, long-duration videos, so a comparison of individual blocks to popular image datasets is outside the scope of this paper due to hardware resource constraints. However, details on full-range head pose estimation and comparison on specific head-image datasets are described in the studies [96], [97], [98], [101]. For Joint Attention Hand pointing estimation, we do not compare against 2D or 3D gesture recognition datasets because our model is distinct from the majority of conventional approaches, where our model is trained using images collected and curated from four hand- or finger-pointing datasets (see section IV-D2).

VI. DISCUSSION

Three independent vision-based paradigms are designed as follows: a model for Activity Comprehension for child-therapist interaction, a model for FER of children, and a real-time approach for automatic joint attention recognition. The demand for low-cost diagnostics, universal screening guidelines, and research funding availability have prompted endeavors to create technology-based ASD screening [9], [39], [48]. Technological advances and the availability of low-cost cloud infrastructure have motivated researchers to automate the creation and processing of video data by constructing data pipelines. Integrating data pipelines with ML technology has advanced the development of cost-effective ASD detection and assessment methods [31], [38], [47]. However, ASD diagnostic services are not always accessible, cost-effective, or data-driven. Our findings indicate that the technology-based ASD approaches can be generalized to the broader population with neurodevelopmental disorders along with few technological modifications and can serve wider population groups with enhanced quality, access, and affordability. In addition, technology-enabled innovations are anticipated to supplement traditional detection methods for the following reasons: Diagnostic methods based on ML and DL can be trained on a large volume of involuntary multimodal data generated from various activities to detect children at risk for ASD. Few diagnostic techniques, such as CARS-2 [30], can diagnose children older than two years. Moreover, children do not develop social communication, language, and other crucial milestones until the second or third year of life. An untrained clinician may therefore receive contradictory results when evaluating ASD risk in infants under two years of age [1], [33]. Abbas et al. [39], Gupta et al. [38], Kohli et al. [9], and Uddin et al. [48] have highlighted the novel ML methods and their feasibility of analysis on ASD and other neurodevelopmental landmarks from behavioral, eye gaze, audio, facial expression, postural, and EHR assessment data to identify children at risk of ASD at an early age, circumventing the age restrictions and limitations of the traditional diagnostic instruments.

Researchers have collected multimodal data from hospital EHRs and constructed enormous multimodal data lakes which enable DL and ML algorithms to discover clinically significant patterns for recognizing ASD, tracking patients over time, prescribing and tailoring treatments, and alleviating ASD severity [39], [47].

A. PRACTICAL IMPLICATIONS

The ABA treatment enhancement efforts described in this research can advance the field of neuroscience by increasing early identification and consequently expanding access to early intervention treatments which can be used by clinicians accessible via mobile or web applications, significantly enhancing their capacity and meeting the needs of children with ASD and other developmental delays (speech, development, and intellectual delay). By supporting the adoption of these technologies through controlled pilots with stakeholders such as parents, doctors, and schools and digitizing downstream detection processes, evaluations, and therapies, a computerized, human-supported ASD diagnosis, and

management framework can be launched and migrated to an autonomous and personalized digital model to optimize cost, maximize scale, and fast-track access to referrals and intervention. According to our knowledge, this is the first attempt to create an automated, integrated ABA assessment framework deployed on the cloud for real-time assessment using video data.

Moreover, to decrease bias and ensure the internal and external validity of the implemented models, there is an urgent need to undertake large clinical trials, including the participation of researchers and doctors from many nations with diverse backgrounds, and ethnicity. The purpose of the partnership is to validate the results, determine efficacy, address potential technology edge cases, and design approaches to incorporate children from various backgrounds into research investigations. Our experimental results from testing joint attention, Activity Comprehension, and facial expression models with YouTube videos demonstrate robustness in chaotic and natural videos.

It is appropriate for medical experts in the ABA and clinical psychology fields to evaluate the findings' validity. The study's strengths include the use of 68 real-world clinical videos and comparisons against ground-truth video annotations provided by clinicians. However, medical professionals have the expertise and experience to interpret the findings and provide additional insights that may not be immediately apparent to those outside the clinical context. In our case, we cross-reviewed the annotations provided by a clinician to ensure a coherent interpretation of video annotations and minimize annotator bias. In addition, medical professionals assessed the study's methodology, identified a few possible limitations, and made suggestions for future research, such as in the case of multiple children handled by a single therapist for Activity Comprehension models. Involving medical professionals in the evaluation of the study's findings increased the study's rigor and credibility and ensured that the results were appropriately interpreted and implemented in clinical practice.

B. LIMITATIONS AND FUTURE DIRECTIONS

While developing the Activity Comprehension model, the current study provides solutions to only a subset of the many activities performed during ABA therapy; many other action classes specific to children can be incorporated if sufficient resources are allocated to model development. As ML technology develops, it is possible to reduce false positives and negatives in the FER model, thereby increasing its sensitivity, precision, and specificity. Even though we have collected as many images of children's faces as possible to train an online FER model, there is still room for additional data and microexpressions. In addition, each model assumes a particular video data distribution: (i) the Activity Comprehension model assumes person-person or person-object interactions, (ii) the FER model assumes frontal face visibility, and joint attention models perform sub-optimally in crowded scenarios. Most of the children's videos we collected lack clinical diagnosis information for ASD or other neurodevelopmental disorders. Future work should include a large clinical trial testing the models on grouped cohorts of ASD, neurotypical, and other developmental disorders that can reveal the level of efficacy and identify areas for further development. Lastly, more real-world test cases may uncover unforeseen edge cases that hamper model performance and generalizability. Future studies can incorporate clinicians' survey responses to determine the efficacy of computer vision models that aid in accurate, timely diagnosis and treatment monitoring. Future studies can integrate various methods into a single pipeline or architecture with a unified model that is trained in a multitask fashion for analyzing human behavior, joint attention, social communication skills, facial expression and motor imitation recognition, and eye contact detection. Further, speech and auditory features can be incorporated that provide rich features in the case of social interaction and communication to develop a multimodal vision-speech model that can identify abnormalities in speech and social behaviors.

VII. CONCLUSION

The paper investigates the viability of a computer vision and deep learning-based ABA treatment and assessment that experts or non-experts can use to detect important behavioral activities, emotions, and JA using videos. Experiments with 68 clinical and public videos from the real world reveal that the activity comprehension model reports an overall accuracy of 72.32%, the joint attention models show an accuracy of 97% for gaze following and 93.4% for hand pointing, and the facial expression recognition model has an overall accuracy of 95.1%. During the development of the activity comprehension and facial expression recognition model, the proposed methodology incorporates diversity and fairness to low-income and middle-income populations by collecting videos of children of different ages, socioeconomic statuses, and ethnicity. The models' predictions help to make real-time monitoring and assessment reports that help clinicians to make decisions about ABA services.

APPENDIX A

A. MODEL EVALUATION METRICS

The robustness of the machine learning model can be evaluated on various metrics items listed below.

a) **Accuracy** is the number of correct predictions divided by the total number of predictions.
b) **True Positive (TP)** signifies how many positive class samples the model predicted correctly.
c) **True Negative (TN)** signifies how many negative class samples the model predicted correctly.
d) **False Positive (FP)** signifies how many negative class samples the model predicted incorrectly.
e) **False Negative (FN)** signifies how many positive class samples the model predicted incorrectly.
f) **Precision** is the ratio of true positives and total positives predicted.

g) **Recall or Sensitivity** is the ratio of true positives to all the positives in the ground truth.
h) **Specificity** is defined as the proportion of actual negative class samples, which got predicted as the true negatives.
i) **F1 Score** is the harmonic mean of precision and recall.
j) **Negative Predictive Value** is the ratio of the number of true negatives to the total number of class samples that test negative.

Table 13 lists evaluation metrics with mathematical notations.

TABLE 13. Machine learning model performance metrics.

Metric Name	Notation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
F1 score	$2 \times \frac{(Precision)(Recall)}{(Precision+Recall)}$
Negative Predictive Value	$\frac{TN}{TN+FN}$

TP: True Positives, TN: True Negatives, FP: False Positives, and FN: False Negatives.

ACKNOWLEDGMENT

The authors would like to acknowledge the clinical experts and engineers at SM Learning Skills Academy for Special Needs Pvt. Ltd. for their support with video data collection and annotation. They also thank Microsoft Azure and IBM Cloud for giving them cloud credits that allowed them to store huge amounts of video data and run GPU virtual machines.

DISCLOSURE STATEMENT

The authors Manu Kohli and Dr. A. P. Prathosh report a relationship with the company SM Learning Skills Academy for Special Needs Pvt. Ltd. that includes: equity or stocks. The remaining authors declare they have no financial or non-financial conflicts of interest.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Conceptualization: Manu Kohli, Swati Kohli, A. P. Prathosh; Data curation: Varun Ganjigunte Prakash, Manu Kohli, Swati Kohli, Tanu Wadhera, Diptanshu Das, Debasis Panigrahi; Formal analysis: A. P. Prathosh, Tanu Wadhera, Diptanshu Das, Debasis Panigrahi; Funding acquisition: Manu Kohli, A. P. Prathosh; Methodology: Varun Ganjigunte Prakash, Manu Kohli, Swati Kohli, A. P. Prathosh; Project administration: Manu Kohli, Swati Kohli; Resources: Manu Kohli, Swati Kohli, A. P. Prathosh, Tanu Wadhera, Diptanshu Das, Debasis Panigrahi; Software: Varun Ganjigunte Prakash, Manu Kohli; Supervision: Manu Kohli, Swati Kohli, A. P. Prathosh, Tanu Wadhera, Diptanshu Das, Debasis Panigrahi; Writing—review and editing: Varun Ganjigunte Prakash, Manu Kohli, Tanu Wadhera; Visualization: Varun Ganjigunte Prakash, Manu Kohli.

ETHICAL APPROVAL

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

INFORMED CONSENT

Informed consent was obtained from all individual participants included in the study.

DATA AVAILABILITY STATEMENT

Due to the nature of the research and participants' consent agreements, data is not available. However, case-by-case analysis can be made as per the request and data access requirements.

REFERENCES

[1] CDC. (2021). *Data and Statistics on Autism Spectrum Disorder*. [Online]. Available: https://www.cdc.gov/ncbddd/autism/data.html
[2] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Developmental Disorders*, vol. 30, no. 3, pp. 205–223, Jun. 2000.
[3] L. Jurek, M. Baltazar, S. Gulati, N. Novakovic, M. Núñez, J. Oakley, and A. O'Hagan, "Response (minimum clinically relevant change) in ASD symptoms after an intervention according to CARS-2: Consensus from an expert elicitation procedure," *Eur. Child Adolescent Psychiatry*, vol. 31, no. 8, pp. 1–10, Aug. 2022.
[4] M. L. Sundberg, *VB-MAPP Verbal Behavior Milestones Assessment and Placement Program: A Language and Social Skills Assessment Program for Children with Autism or Other Developmental Disabilities*. Concord, CA, USA: AVB Press, 2008.
[5] H. S. Roane, W. W. Fisher, and J. E. Carr, "Applied behavior analysis as treatment for autism spectrum disorder," *J. Pediatrics*, vol. 175, pp. 27–32, Sep. 2016.
[6] R. K. Dogan, M. L. King, A. T. Fischetti, C. M. Lake, T. L. Mathews, and W. J. Warzak, "Parent-implemented behavioral skills training of social skills," *J. Appl. Behav. Anal.*, vol. 50, no. 4, pp. 805–818, Oct. 2017.
[7] D. M. Bhatia. (Aug. 2020). *How to Help Low-Income Children With Autism*. [Online]. Available: https://www.doctorbhatia.com/autism-research-treatment/autism-diagnosis-how-is-autism-diagnosed-clinical-screening-cars-blood-gentic-tests/
[8] A. Opar. (Jan. 2019). *How to Help Low-Income Children With Autism*. [Online]. Available: https://www.spectrumnews.org/features/deep-dive/help-low-income-children-autism/
[9] M. Kohli, A. K. Kar, and S. Sinha, "The role of intelligent technologies in early detection of autism spectrum disorder (ASD): A scoping review," *IEEE Access*, vol. 10, pp. 104887–104913, 2022.
[10] M. Kohli and S. Kohli, "Electronic assessment and training curriculum based on applied behavior analysis procedures to train family members of children diagnosed with autism," in *Proc. IEEE Region 10 Humanitarian Technol. Conf. (R10-HTC)*, Dec. 2016, pp. 1–6.

[11] A. Ali, F. F. Negin, F. F. Bremond, and S. Thümmler, "Video-based behavior understanding of children for objective diagnosis of autism," in *Proc. 17th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Feb. 2022, pp. 1–11. [Online]. Available: https://hal.inria.fr/hal-03447060

[12] Q. Tariq, S. L. Fleming, J. N. Schwartz, K. Dunlap, C. Corbin, P. Washington, H. Kalantarian, N. Z. Khan, G. L. Darmstadt, and D. P. Wall, "Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: Development and validation study," *J. Med. Internet Res.*, vol. 21, no. 4, Apr. 2019, Art. no. e13822.

[13] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *J. Autism Develop. Disorders*, vol. 24, no. 5, pp. 659–685, Oct. 1994.

[14] P. U. Putra, K. Shima, S. A. Alvarez, and K. Shimatani, "Identifying autism spectrum disorder symptoms using response and gaze behavior during the Go/NoGo game CatChicken," *Sci. Rep.*, vol. 11, no. 1, p. 22012, Nov. 2021, doi: 10.1038/s41598-021-01050-7.

[15] L. Billeci et al., "Disentangling the initiation from the response in joint attention: An eye-tracking study in toddlers with autism spectrum disorders," *Transl. Psychiatry*, vol. 6, no. 5, p. e808, May 2016, doi: 10.1038/tp.2016.75.

[16] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: A scoping review," *Transl. Psychiatry*, vol. 10, no. 1, p. 116, Apr. 2020, doi: 10.1038/s41398-020-0780-3.

[17] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *npj Digit. Med.*, vol. 4, no. 1, p. 5, Jan. 2021, doi: 10.1038/s41746-020-00376-2.

[18] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda," *J. Ambient Intell. Humanized Comput.*, vol. 2022, pp. 1–28, Jan. 2022, doi: 10.1007/s12652-021-03612-z.

[19] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, and S. Jayarathna, "EEG-based processing and classification methodologies for autism spectrum disorder: A review," *J. Comput. Sci.*, vol. 15, no. 8, pp. 1161–1183, Aug. 2019.

[20] L. Klintwall and S. Eikeseth, *Early and Intensive Behavioral Intervention (EIBI) in Autism*. New York, NY, USA: Springer, 2014, pp. 117–137, doi: 10.1007/978-1-4614-4788-7_129.

[21] J. J. Wood, A. Drahota, K. Sze, K. Har, A. Chiu, and D. A. Langer, "Cognitive behavioral therapy for anxiety in children with autism spectrum disorders: A randomized, controlled trial," *J. Child Psychol. Psychiatry*, vol. 50, no. 3, pp. 224–234, Mar. 2009, doi: 10.1111/j.1469-7610.2008.01948.x.

[22] M. E. Król and M. Król, "A novel machine learning analysis of eye-tracking data reveals suboptimal visual information extraction from facial stimuli in individuals with autism," *Neuropsychologia*, vol. 129, pp. 397–406, Jun. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002839321930106X

[23] R. Aishworiya, T. Valica, R. Hagerman, and B. Restrepo, "An update on psychopharmacological treatment of autism spectrum disorder," *Neurotherapeutics*, vol. 19, no. 1, pp. 248–262, Jan. 2022, doi: 10.1007/s13311-022-01183-1.

[24] X. Wang, J. Zhao, S. Huang, S. Chen, T. Zhou, Q. Li, X. Luo, and Y. Hao, "Cognitive behavioral therapy for autism spectrum disorders: A systematic review," *Pediatrics*, vol. 147, no. 5, May 2021, doi: 10.1542/peds.2020-049880.

[25] D. Ung, R. Selles, B. J. Small, and E. A. Storch, "A systematic review and meta-analysis of cognitive-behavioral therapy for anxiety in youth with high-functioning autism spectrum disorders," *Child Psychiatry Hum. Develop.*, vol. 46, no. 4, pp. 533–547, Aug. 2015, doi: 10.1007/s10578-014-0494-y.

[26] F. Mohammadzaheri, L. K. Koegel, M. Rezaee, and S. M. Rafiee, "A randomized clinical trial comparison between pivotal response treatment (PRT) and structured applied behavior analysis (ABA) intervention for children with autism," *J. Autism Develop. Disorders*, vol. 44, no. 11, pp. 2769–2777, Nov. 2014, doi: 10.1007/s10803-014-2137-3.

[27] Q. Yu, E. Li, L. Li, and W. Liang, "Efficacy of interventions based on applied behavior analysis for autism spectrum disorder: A meta-analysis," *Psychiatry Invest.*, vol. 17, no. 5, pp. 432–443, May 2020.

[28] H. Manohar, P. Kandasamy, V. Chandrasekaran, and R. P. Rajkumar, "Early diagnosis and intervention for autism spectrum disorder: Need for pediatrician-child psychiatrist liaison," *Indian J. Psychol. Med.*, vol. 41, no. 1, pp. 87–90, 2019, doi: 10.4103/IJPSYM.IJPSYM_154_18.

[29] D. L. Robins, K. Casagrande, M. Barton, C.-M.-A. Chen, T. Dumont-Mathieu, and D. Fein, "Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F)," *Pediatrics*, vol. 133, no. 1, pp. 37–45, Jan. 2014.

[30] C. A. Vaughan, "Test review: E. Schopler, ME Van Bourgondien, GJ Wellman, & SR Love childhood autism rating scale. Los Angeles, CA: Western psychological services, 2010," *J. Psychoeducational Assessment*, vol. 29, no. 5, pp. 489–493, Oct. 2011.

[31] M. Marlow, C. Servili, and M. Tomlinson, "A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: Recommendations for use in low- and middle-income countries," *Autism Res.*, vol. 12, no. 2, pp. 176–199, Feb. 2019, doi: 10.1002/aur.2033.

[32] K. Bauer, K. L. Morin, T. E. Renz, and S. Zungu, "Autism assessment in low- and middle-income countries: Feasibility and usability of western tools," *Focus Autism Other Develop. Disabilities*, vol. 37, no. 3, pp. 179–188, Sep. 2022, doi: 10.1177/10883576211073691.

[33] R. Choueiri, W. T. Garrison, and V. Tokatli, "Early identification of autism spectrum disorder (ASD): Strategies for use in local communities," *Indian J. Pediatrics*, vol. 90, no. 4, pp. 377–386, May 2022, doi: 10.1007/s12098-022-04172-6.

[34] N. J. Hidalgo, L. L. McIntyre, and E. H. McWhirter, "Sociodemographic differences in parental satisfaction with an autism spectrum disorder diagnosis," *J. Intellectual Develop. Disability*, vol. 40, no. 2, pp. 147–155, Apr. 2015.

[35] A. J. Kumm, M. Viljoen, and P. J. de Vries, "The digital divide in technologies for autism: Feasibility considerations for low- and middle-income countries," *J. Autism Develop. Disorders*, vol. 52, no. 5, pp. 2300–2313, May 2022.

[36] N. Malik-Soni, A. Shaker, H. Luck, A. E. Mullin, R. E. Wiley, M. E. S. Lewis, J. Fuentes, and T. W. Frazier, "Tackling healthcare access barriers for individuals with autism from diagnosis to adulthood," *Pediatric Res.*, vol. 91, no. 5, pp. 1028–1035, Apr. 2022.

[37] M. P. Kelly and P. Reed, "Examination of stimulus over-selectivity in children with autism spectrum disorder and its relationship to stereotyped behaviors and cognitive flexibility," *Focus Autism Other Develop. Disabilities*, vol. 36, no. 1, pp. 47–56, Mar. 2021.

[38] C. Gupta, P. Chandrashekar, T. Jin, C. He, S. Khullar, Q. Chang, and D. Wang, "Bringing machine learning to research on intellectual and developmental disabilities: Taking inspiration from neurological diseases," *J. Neurodevelopmental Disorders*, vol. 14, no. 1, p. 28, May 2022, doi: 10.1186/s11689-022-09438-w.

[39] H. Abbas, F. Garberson, S. Liu-Mayo, E. Glover, and D. P. Wall, "Multi-modular AI approach to streamline autism diagnosis in young children," *Sci. Rep.*, vol. 10, no. 1, p. 5014, Mar. 2020, doi: 10.1038/s41598-020-61213-w.

[40] D.-Y. Song, S. Y. Kim, G. Bong, J. M. Kim, and H. J. Yoo, "The use of artificial intelligence in screening and diagnosis of autism spectrum disorder: A literature review," *J. Korean Acad. Child Adolescent Psychiatry*, vol. 30, no. 4, pp. 145–152, Oct. 2019, doi: 10.5765/jkacap.190027.

[41] G. S. Young, J. N. Constantino, S. Dvorak, A. Belding, D. Gangi, A. Hill, M. Hill, M. Miller, C. Parikh, A. J. Schwichtenberg, E. Solis, and S. Ozonoff, "A video-based measure to identify autism risk in infancy," *J. Child Psychol. Psychiatry*, vol. 61, no. 1, pp. 88–94, Jan. 2020.

[42] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *J. Autism Develop. Disorders*, vol. 44, no. 10, pp. 2413–2428, Oct. 2014.

[43] K. L. H. Carpenter, J. Hahemi, K. Campbell, S. J. Lippmann, J. P. Baker, H. L. Egger, S. Espinosa, S. Vermeer, G. Sapiro, and G. Dawson, "Digital behavioral phenotyping detects atypical pattern of facial expression in toddlers with autism," *Autism Res.*, vol. 14, no. 3, pp. 488–499, Mar. 2021.

[44] R. Rahman, A. Kodesh, S. Z. Levine, S. Sandin, A. Reichenberg, and A. Schlessinger, "Identification of newborns at risk for autism using electronic medical records and machine learning," *Eur. Psychiatry*, vol. 63, no. 1, p. e22, 2020.

[45] L. Ouss, G. Palestra, C. Saint-Georges, M. L. Gille, M. Afshar, H. Pellerin, K. Bailly, M. Chetouani, L. Robel, B. Golse, R. Nabbout, I. Desguerre, M. Guergova-Kuras, and D. Cohen, "Behavior and interaction imaging at 9 months of age predict autism/intellectual disability in high-risk infants with west syndrome," *Transl. Psychiatry*, vol. 10, no. 1, pp. 1–7, Feb. 2020.

[46] J. Hashemi, G. Dawson, K. L. H. Carpenter, K. Campbell, Q. Qiu, S. Espinosa, S. Marsan, J. P. Baker, H. L. Egger, and G. Sapiro, "Computer vision analysis for quantification of autism risk behaviors," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 215–226, Jan. 2021.

[47] M. Kohli, A. K. Kar, A. Bangalore, and P. Ap, "Machine learning-based ABA treatment recommendation and personalization for autism spectrum disorder: An exploratory study," *Brain Informat.*, vol. 9, no. 1, p. 16, Jul. 2022, doi: 10.1186/s40708-022-00164-6.

[48] M. Uddin, Y. Wang, and M. Woodbury-Smith, "Artificial intelligence for precision medicine in neurodevelopmental disorders," *npj Digit. Med.*, vol. 2, no. 1, p. 112, Nov. 2019, doi: 10.1038/s41746-019-0191-0.

[49] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders," *Sci. Robot.*, vol. 5, no. 39, Feb. 2020, doi: 10.1126/scirobotics.aaz3791.

[50] M. Toshpulatov, W. Lee, S. Lee, and A. H. Roudsari, "Human pose, hand and mesh estimation using deep learning: A survey," *J. Supercomput.*, vol. 78, no. 6, pp. 7616–7654, Apr. 2022, doi: 10.1007/s11227-021-04184-7.

[51] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation," *IEEE Access*, vol. 8, pp. 133330–133348, 2020.

[52] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," 2020, *arXiv:2012.13392*.

[53] W. Zhang, J. Fang, X. Wang, and W. Liu, "EfficientPose: Efficient human pose estimation with neural architecture search," *Comput. Vis. Media*, vol. 7, no. 3, pp. 335–347, Sep. 2021, doi: 10.1007/s41095-021-0214-z.

[54] S. Dubey and M. Dixit, "A comprehensive survey on human pose estimation approaches," *Multimedia Syst.*, vol. 29, no. 1, pp. 167–195, Aug. 2022, doi: 10.1007/s00530-022-00980-0.

[55] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2D human pose estimation: A survey," 2022, *arXiv:2204.07370*.

[56] G. Sciortino, G. M. Farinella, S. Battiato, M. Leo, and C. Distante, "On the estimation of children's poses," in *Image Analysis and Processing—ICIAP*, S. Battiato, G. Gallo, R. Schettini, and F. Stanco, Eds. Cham, Switzerland: Springer, 2017, pp. 410–421.

[57] J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. D. Reetzke, M. F. Vignos, and R. T. Roemmich, "Applications of pose estimation in human health and performance across the lifespan," *Sensors*, vol. 21, no. 21, p. 7315, Nov. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/21/7315

[58] D. Cazzato, P. L. Mazzeo, P. Spagnolo, and C. Distante, "Automatic joint attention detection during interaction with a humanoid robot," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Cham, Switzerland: Springer, 2015, pp. 124–134.

[59] K. Kim and P. Mundy, "Joint attention, social-cognition, and recognition memory in adults," *Frontiers Hum. Neurosci.*, vol. 6, p. 172, Jun. 2012.

[60] P. Nyström, E. Thorup, S. Bölte, and T. Falck-Ytter, "Joint attention in infancy and the emergence of autism," *Biol. Psychiatry*, vol. 86, no. 8, pp. 631–638, Oct. 2019, doi: 10.1016/j.biopsych.2019.05.006.

[61] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng, Y. Lin, and J. Kong, "Applying eye tracking to identify autism spectrum disorder in children," *J. Autism Develop. Disorders*, vol. 49, no. 1, pp. 209–215, Jan. 2019.

[62] W. Zhao and L. Lu, "Research and development of autism diagnosis information system based on deep convolution neural network and facial expression data," *Library Hi Tech*, vol. 38, no. 4, pp. 799–817, Mar. 2020.

[63] G. Alvari, C. Furlanello, and P. Venuti, "Is smiling the key? Machine learning analytics detect subtle patterns in micro-expressions of infants with ASD," *J. Clin. Med.*, vol. 10, no. 8, p. 1776, Apr. 2021. [Online]. Available: https://www.mdpi.com/2077-0383/10/8/1776

[64] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, May 2018, pp. 6047–6056.

[65] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. (2014). *THUMOS Challenge: Action Recognition With a Large Number of Classes*. [Online]. Available: http://crcv.ucf.edu/THUMOS14/

[66] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[67] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[68] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose motion representation for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7024–7033.

[69] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[70] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[71] 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.

[72] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers Neurosci.*, vol. 14, p. 779, Oct. 2020.

[73] P. Pandey, P. Ap, M. Kohli, and J. Pritchard, "Guided weak supervision for action recognition with scarce data to assess skills of children with autism," *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 1, pp. 463–470. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5383

[74] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Aug. 2017, pp. 4724–4733.

[75] S. F. Dos Santos, N. Sebe, and J. Almeida, *CV-C3D: Action Recognition on Compressed Videos with Convolutional 3D Networks*. Porto Alegre, RS, Brasil: SBC, 2019. [Online]. Available: https://sol.sbc.org.br/index.php/sibgrapi/article/view/9782

[76] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5794–5803.

[77] A. Richard and J. Gall, "A bag-of-words equivalent recurrent neural network for action recognition," *Comput. Vis. Image Understand.*, vol. 156, pp. 79–91, Mar. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314216301680

[78] J. Tang, J. Xia, X. Mu, B. Pang, and C. Lu, *Asynchronous Interaction Aggregation for Action Detection*. New York, NY, USA: Springer-Verlag, 2020, pp. 71–87, doi: 10.1007/978-3-030-58555-6_5.

[79] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3599–3608.

[80] V. Escorcia, F. Caba Heilbron, J. C. Niebles, and B. Ghanem, "DAPs: Deep action proposals for action understanding," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 768–784.

[81] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4415–4423.

[82] M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, and R. Cucchiara, "Video action detection by learning graph-based spatio-temporal interactions," *Comput. Vis. Image Understand.*, vol. 206, May 2021, Art. no. 103187. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731422100031X

[83] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314221000631

[84] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 413–431.

[85] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jan. 2019, pp. 6202–6211.

[86] H. Xu, L. Yang, S. Sclaroff, K. Saenko, and T. Darrell, "Spatio-temporal action detection with multi-object interaction," 2020, *arXiv:2004.00180*.
[87] M. Tapaswi, V. Kumar, and I. Laptev, "Long term spatio-temporal modeling for action detection," *Comput. Vis. Image Understand.*, vol. 210, Sep. 2021, Art. no. 103242. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314221000862
[88] J. Tang, J. Xia, X. Mu, B. Pang, and C. Lu, "Asynchronous interaction aggregation for action detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 71–87.
[89] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[90] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, F. Ingham, J. Poznanski, J. Fang, L. Yu, M. Wang, N. Gupta, O. Akhtar, and P. Rai, "ultralytics/yolov5: V3.1—Bug fixes and performance improvements," Ultralytics, Los Angeles, CA, USA, Tech. Rep. 7.0, Oct. 2020, doi: 10.5281/zenodo.4154370.
[91] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2015.
[92] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015.
[93] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, and A. Oliva, "Moments in time dataset: One million videos for event understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 1–8, Feb. 2019.
[94] J. Bidwell, A. Rozga, I. Essa, and G. Abowd, "Measuring child visual attention using markerless head tracking from color and depth sensing cameras," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 447–454.
[95] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.
[96] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians' head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, Jan. 2018, doi: 10.1016/j.neucom.2017.07.029.
[97] E. Rehder, H. Kloeden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 2292–2297.
[98] D. Heo, J. Nam, and B. Ko, "Estimation of pedestrian pose orientation using soft target training based on teacher–student framework," *Sensors*, vol. 19, no. 5, p. 1147, Mar. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/5/1147
[99] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.
[100] T.-H. Vu, A. Osokin, and I. Laptev, "Context-aware CNNs for person head detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2893–2901.
[101] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," 2020, *arXiv:2005.10353*.
[102] Y. Huang, X. Liu, L. Jin, and X. Zhang, "DeepFinger: A cascade convolutional neuron network approach to finger key point detection in egocentric vision with mobile camera," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2015, pp. 2944–2949.
[103] Y. Huang, X. Liu, X. Zhang, and L. Jin, "A pointing gesture based egocentric interaction system: Dataset, approach and application," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 16–23.
[104] D. Shukla, O. Erkent, and J. Piater, "Probabilistic detection of pointing directions for human–robot interaction," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2015, pp. 1–8.
[105] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai, "IPN hand: A video dataset and benchmark for real-time continuous hand gesture recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 4340–4347.
[106] H. Xia and Y. Zhan, "A survey on temporal action localization," *IEEE Access*, vol. 8, pp. 70477–70487, 2020.

VARUN GANJIGUNTE PRAKASH received the B.E. degree in electronics and communication from Sri Jayachamarajendra College of Engineering, India, in 2018. He has five years of experience in developing machine learning and computer vision-based technological solutions for multiple industries and startups. He has spent time working on many computer vision and robotics challenges. His research interests include solving problems at the intersection of computer vision and robotics, computer vision, robotic manipulation, control systems, autonomous mobile robots, deep learning, deep reinforcement learning, robotic system design, and machine learning.

MANU KOHLI is currently pursuing the Ph.D. degree with the Indian Institute of Technology, Delhi (IIT Delhi). He has 18 years of experience in executing large-scale business and digital transformation projects in multiple countries. He has undertaken leadership positions in Fortune 500 organizations globally and has worked on multiple technologies, such as SAP, Saas, and machine learning. He is the Chief Technology Officer of CogniAble, where he has developed innovative artificial intelligence solutions for detecting and managing developmental challenges, including autism, with outstanding psychometric properties. He has led the formation of new technology-enabled businesses and ensured their commercialization. He has authored multiple publications in peer-reviewed journals and books by SAP PRESS. His research interests include developing cutting-edge machine learning, deep learning, and computer vision methods to solve complex business and healthcare problems. He has received numerous honors, including the UNICEF Blue Ribbon, AI Gamechangers, and cash prizes from Lockheed Martin, Tata-Trusts, Western Digital, NASSCOM, NTT-Data, and the Ministry of Electronics for his innovations.

SWATI KOHLI received the Diploma degree, in 1998, the B.Ed. degree in special education, in 2002, the Postgraduate Diploma degree in early intervention, and the M.A. degree in psychology. She completed the ABA coursework from the Florida Institute of Technology. She is currently the Clinical Director of CogniAble. She has more than 18 years of experience in working for special needs children with neuro-developmental delays in school, center, and clinic-based settings.

A. P. PRATHOSH received the B.Tech. degree, in 2011, and the Ph.D. degree in temporal data analysis from the Indian Institute of Science (IISc), Bengaluru, in 2015. He submitted the Ph.D. thesis three years after the B.Tech. degree, with many top-tier journal publications. He also happens to be a Student of the Sanskrit language and Indian Philosophical Sciences. He was with corporate research labs, including Xerox Research India, Philips Research, and a start-up in CA, USA. He has co-founded CogniAble which builds learning algorithms for behavioral healthcare using video analytics (first-place winner of the recent AI startup challenge by the Government of India) and also actively engaged with several corporate industries, start-ups, and medical centers (e.g., AIIMS) in solving interesting technical problems. He joined the Computer Technology Group of Electrical Engineering, IIT Delhi, in 2017, as an Assistant Professor, where he was engaged in research and teaching of machine and deep learning courses. He is currently a Faculty Member with the Department of ECE, IISc. His work in the industry, focusing on healthcare analytics, led to the generation of several IPs, comprising 15 (U.S.) patents of which ten are granted and six are commercialized. His current research interests include deep-representational learning, cross-domain generalization, signal processing, and their applications in computer vision and speech analytics.

TANU WADHERA received the B.Tech. degree in electronics and communication from the Guru Nanak Engineering College, Ludhiana, India, the M.Tech. degree in electronics and communication from Punjabi University, Patiala, India, and the Ph.D. degree from the National Institute of Technology Jalandhar (NIT Jalandhar), Jalandhar, Punjab, India. She completed her postdoctoral research with the Indian Institute of Technology, Delhi, India. She has a total of six years of research experience, including four years with NIT Jalandhar. She has one year of teaching experience as an Assistant Professor at NIT Jalandhar. Based on her contribution to the field of computational healthcare, especially autism spectrum disorder and other disabilities lying on the same spectrum, she is currently a Project Engineer with the Indian Institute of Technology in collaboration with AIIMS, Delhi. She is also an Assistant Professor with the School of Electronics, Indian Institute of Information Technology Una, Una, India. She has experience publishing work in reputed journals and editing and/or authoring books for several journals. Her research interests include artificial intelligence, assistive technology, behavioral modeling, biomedical signal processing, cognitive neuroscience, and machine learning.

DIPTANSHU DAS received the M.B.B.S. degree from the Medical College, Kolkata, the M.H.Sc. degree in clinical child development from the University of Kerala, and the M.D. degree in pediatric neurology (level 2 master's) from the University of Rome, Tor Vergata. He is a Pediatric Neurologist by education and profession. He was formerly a Consultant with Medica Superspecialty Hospital (MSH), Kolkata, and the Institute of Neurosciences Kolkata (I-NK). He is currently a Consultant with the Institute of Child Health, Kolkata. He is the Founder/Co-Founder of the Institute of NeuroDevelopment, Kolkata, India. He is also a long-term Wikipedian and an Advocate for open access and open education. He is a General Practitioner and an eminent Pediatric Neurologist. He is proficient in treating patients with epilepsy or neurodevelopmental disorders, with 16 years of experience. He has several academic publications. He is a member of the Education Council and the Publication Council of the International League Against Epilepsy (ILAE). He is an Executive Board Member of the Association of Child Neurology (AOCN), India, and the Editor-in-Chief of the ILAE Project.

DEBASIS PANIGRAHI received the M.B.B.S. degree from the Veer Surendra Sai Institute of Medical Sciences and Research, Sambalpur, Odisha, in 2006, and the M.D. degree in pediatrics from the Sriram Chandra Bhanj Medical College, Cuttack, in 2011. He received a fellowship in pediatric neurology from the Kanchi Kamakoti Child Trust Hospital, Chennai, in 2013. He has 16 years of experience as a Pediatrician and a Pediatric Neurologist in Bhubaneswar. He currently practices with the Child Neuro Clinic, Jagannath Hospital, Bhubaneswar. He is a member of the Indian Academy of Paediatrics (IAP), the Association of Child Neurologists, India, and the IAP's Developmental Pediatric Chapter.

JOHN VIJAY SAGAR KOMMU received the bachelor's degree in medical training from the S. V. Medical College, Tirupati, and the Doctor of Medicine (M.D.) degree in psychiatry from the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru. He has extensive academic, clinical, and research experience of over 20 years, having worked in reputed institutions, such as Christian Medical College, Vellore, and JIPMER, Pondicherry. He was trained with the McKnight Brain Institute, University of Florida. He was recently appointed to Affiliate Faculty with the School of Biological and Population Health Sciences, Oregon State University. He has been a Consultant Child Psychiatrist with the Department of Child and Adolescent Psychiatry, NIMHANS, for over ten years. He is currently a Professor and the Head of the Department of Child and Adolescent Psychiatry, NIMHANS. He has authored 60 research articles and seven book chapters. His current research interests include neurodevelopmental disorders especially ASD, pediatric psychopharmacology, adolescent mental health, child abuse, and neurobiology of child psychiatric disorders. He was awarded the prestigious NIH (USA) funded Indo–U.S. Fogarty Fellowship, in 2014.