Deep learning approaches to predicting Autism Spectrum Disorder diagnosis from video data

Shotaro Hato

Supervised by Ulrike Gehring and Sonja de Zwarte

MSc Bioinformatics and Biocomplexity

Utrecht University

Title: Deep learning approaches to predicting Autism Spectrum Disorder diagnosis from video data

Shotaro Hato
s.hato@students.uu.nl
Student number: 1066757
MSc Bioinformatics and Biocomplexity
Utrecht University

Writing Assignment (December 17th, 2023 – January 15th, 2024)
Under the supervision of Ulrike Gehring and Sonja de Zwarte
Utrecht University
Yalelaan 2
3584 CM Utrecht
The Netherlands

Contents

Abstract

Autism Spectrum Disorder (ASD), a neurodevelopmental condition, affects approximately 1 in 100 individuals worldwide. Confirming a clinical diagnosis of ASD relies predominantly on interviews and questionnaires, yet these approaches have inherent limitations. Presently, the rapid development of Artificial Intelligence (AI) technologies across various domains, including medical research, has spurred considerable interest among researchers exploring AI applications in ASD studies. A noteworthy approach is the utilization of video-based ASD diagnosis with AI, offering advantages in terms of accessibility and information volume compared to other data modalities, such as facial and brain images.

In this study, we conducted a search for video and AI-based ASD diagnosis studies published between 2018 and 2024, identifying nine pertinent papers. Our analysis and discussion of these papers, segregated by input features, 1. Facial features, 2. Pose and gait features and 3. Multimodal features. These input features resulted in a promising ASD prediction accuracy on the test data range of 79.7-96.39%. However, we also highlighted certain issues and areas for improvement like out-of-cohort validation, sample size, the black box problem with AI, low specificity, and the establishment of robust and easy video-capturing protocol. These insights contribute valuable information for future clinical applications in the domain of ASD diagnosis.

Layman’s summary

According to a report by the World Health Organization (WHO), Autism Spectrum Disorder (ASD) affects 1 in 100 children. Neurodevelopmental disorders, including ASD and Attention Deficit Hyperactivity Disorder (ADHD), significantly impact individuals' social lives. ASD patients often encounter challenges in focusing on tasks, leading to difficulties in school activities. ASD, being a spectrum disorder, exhibits varying degrees of severity among individuals. The co-occurrence of ASD and ADHD in an individual can result in an intense concentration on specific interests but vulnerability to distraction by other stimuli. Pinpointing causal factors or genes is challenging due to the diversity in the ASD population and its developmental nature within brain neurons.

Despite these complexities, the substantial impact of ASD on individuals' lives has spurred research into diagnosis and treatment. Traditional diagnostic methods rely on interviews with clinicians and questionnaires completed by parents and children. Generally, ASD can be diagnosed at the age of months 18 and 24, but symptoms are more distinct at older ages. Therefore, a definitive diagnosis can be made later. Still, Early diagnosis makes early interventions possible to prevent developmental delays. However, the interview-based approach has faced limitations, such as the difficulty of early-age diagnosis, the time-consuming nature of interviews, and the stress imposed on medical experts, children, and parents. The accuracy of traditional diagnoses with interviews remains a concern, as even with sophisticated and standardized instructions for ASD diagnosis, interviewers may overlook minor symptoms in affected individuals. Furthermore, comprehending the diverse spectrum of ASD poses a formidable challenge for human evaluators.

To address these issues, there is growing interest in leveraging Artificial Intelligence (AI) technologies for ASD diagnosis. AI, with its capacity to handle extensive datasets and solve complex tasks swiftly, is gaining prominence in various scientific studies. Many ASD researchers advocate for AI applications in diagnosis, utilizing multiple inputs such as facial images, brain images, and audio data. Videos, as a promising input for ASD diagnosis, provide rich information, capturing specific behavioral and facial movements exhibited by ASD patients. The accessibility of video data further enhances its appeal for ASD diagnosis in young children.

Thus, in this study, we review papers about AI approaches in ASD diagnosis specifically focusing on video data.

1. Introduction

1.1. ASD characteristics

Autism Spectrum Disorder (ASD) constitutes a neurodevelopmental condition that impacts social activities such as communication, education, and vocational pursuits. The prevalence of ASD is reported to be 1 in 100 children (Zeidan et al., 2022). Recently, the ASD population has continued to rise. A comprehensive review paper addressing the escalation of ASD (Matson and Kozlowski, 2011) posited that changes and advancements in ASD diagnosis might contribute to the observed increase. Environmental and genetic factors may underlie this phenomenon; however, the intricate nature of ASD etiology renders it still largely unknown.

Individuals with ASD exhibit distinctive behaviors in their daily lives. Echolalia, characterized by the repetitive use of phrases, is a notable trait. Additional ASD-specific patterns encompass challenges in communication, difficulty in transitioning between activities, and a tendency to adhere rigidly to specific behaviors or objects (WHO, 2023). Moreover, individuals with ASD display reduced facial expressions and emotional expressions. In stressful situations, individuals with ASD may exhibit an intense response known as a meltdown. These deviations from typical development (TD) can impede various tasks, including academic studies and occupational responsibilities.

Contemporary parlance designates autism as a spectrum, acknowledging the existence of a continuum of traits. The severity of symptoms correlates with the requisite level of support. Due to the comorbidity of ASD with conditions such as depression, Attention Deficit Hyperactivity Disorder (ADHD), and other mental disorders, investigating not only ASD cohorts but also cohorts focussing on these related conditions is significant to make accurate diagnoses for subsequent medical examinations and therapeutic interventions. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; American Psychiatric Association, 2013) stands as a comprehensive diagnostic manual encompassing over 70 disorders, including ASD. Involving more than 200 experts, the criteria in DSM-5 are highly credible. According to DSM-5, ASD can be categorized into three levels: Requiring Support (Level 1), Requiring Substantial Support (Level 2), and Requiring Very Substantial Support (Level 3). Subjects are classified based on the severity of repetitive movements, insistence on sameness, fixated interests, etc.

Researchers and healthcare professionals worldwide express keen interest in the exploration of ASD studies and therapeutic interventions. Conventional therapeutic approaches include parent training and Applied Behavior Analysis (ABA). The former entails the active involvement of parents in the training of children with ASD, while the latter necessitates intensive intervention programs (Brentani et al., 2013). Although there

exist over a thousand different strategies for treating ASD symptoms, the efficacy of these approaches varies among individuals.

1.2. Traditional ASD diagnosis and its limitations

Traditionally, ASD diagnoses rely on interviews and questionnaires based on criteria like DSM-5. For example, the Modified Checklist for Autism in Toddlers (M-CHAT; Robins et al., 2001) and M-CHAT with Follow-Up (M-CHAT/F) serve as reliable screening criteria for early ASD detection globally (Kleinman et al., 2008). The M-CHAT comprises 23 yes/no questions for parents addressing children's behavior and development. Despite the demonstrated high reliability in an experiment involving 1293 children, it does not necessitate specialized devices (Robins et al., 2001).

The Screening Tool for Autism in Toddlers & Young Children (STAT) offers another viable method for ASD identification, targeting children aged 2-3 years. STAT requires only 20 minutes for assessment by a clinician and consists of 12 items pertaining to children's social activities, interactions, and behaviors. In many instances, psychiatrists and pediatricians cannot provide a definitive ASD diagnosis for at-risk individuals until they reach the age of 3.5 years. Nevertheless, early diagnosis using these characteristics is crucial as it facilitates early tailored therapies, stress reduction, and cost savings (Okoye et al., 2023). ASD traits may manifest in infants under 9 months, starting with a lack of response to their name and facial expressions. As the child ages, more ASD-specific traits and behaviors may become observable during interactions with parents and peers.

ASD diagnosis has been investigated by many researchers, but there are certain limitations. Firstly, as mentioned above, ASD has a diverse spectrum. Three-level classifications like DSM-5 or binary classifications are not enough. Acknowledging the spectrum and divergence of autism holds significance for therapeutic approaches and follow-up assessments. Despite that, non-binary and more accurate assessment methods by clinicians have not been developed. As highlighted earlier, early diagnoses are imperative, and numerous screening tools aim to fulfill this objective. However, since infants show fewer ASD features compared with older children, symptoms are not obvious to clinicians. Taking into account these facts, a non-human i.e. computer-based approach has been developed in ASD studies.

Additionally, there are potential drawbacks to early diagnosis. For instance, an ASD diagnosis can carry a stigma for children at a young age. Additionally, the false positive rate of early diagnosis exceeds that of late diagnosis due to certain symptoms manifesting at older ages in the context of social communication (Okoye et al., 2023). Thus, precise diagnosis, extensive research, and societal understanding are essential for future ASD studies.

2. AI approaches in ASD diagnosis

2.1. AI overview and application examples

In various academic disciplines, there is a growing emphasis on the application of artificial intelligence (AI). AI demonstrates proficiency in handling extensive datasets such as genetic data, images, videos, and natural languages. Within the domain of AI, notable subsets include Machine Learning (ML), Deep Learning (DL), and Deep Neural Network (DNN). These methodologies are adept at processing input data, engaging in learning processes, adjusting parameters, and ultimately generating values or classifications.

DL comprises a network of neurons that produce output values based on the non-linear combination of inputs (LeCun et al., 2015). Given the intricate nature of phenomena encountered in the fields of biology and medicine, coupled with the vastness of available data, there has been a notable surge in interest among biologists and medical researchers in leveraging AI technologies.

AI has made significant strides in the realm of ASD studies, with a notable application being facial affect detection in individuals with ASD. The challenges faced by individuals with ASD in expressing emotions through facial expressions can pose substantial obstacles to forming meaningful relationships. Due to the distinctive nature of facial expressions in individuals with ASD, the development of facial processing technologies tailored to this population is an ongoing endeavour (Gepner et al., 2001).

Various AI technologies, including Convolutional Neural Networks (CNN) and vision transformers (ViT), designed for image processing have emerged as promising tools for facial affect detection. A study (Awatramani and Hasteer, 2020) conducted by investigated the performance of a CNN model in facial affect recognition. The model was trained using 28,709 facial expression images and tested with 3,589 images from the FER-2013 dataset (Goodfellow et al., 2013). After categorizing emotions into seven classes (Angry, Disgust, Happy, Sad, Scared, Surprise, and Neutral), the model achieved an accuracy of 67.50%. The researchers suggested that CNNs could be applied to real-time videos, including those involving individuals with ASD.

Videos and AI demonstrate effective synergy due to the substantial volume of data inherent in video datasets, which aligns with AI's proficiency in managing extensive datasets. For example, there are successful applications of DL methods with video data in investigating head movements typical of ASD (Dawson et al., 2018; Martin et al., 2018). The field of video-based diagnosis utilizing AI technologies stands out as a promising avenue in ASD studies. The notable facial expression and behavior differences between ASD and typically developing (TD) groups make a video-based approach a potential diagnostic tool for ASD. The salient advantage of the video-based approach lies in its accessibility

and availability, particularly when compared to other data inputs. Home videos, conveniently captured by parents of children at risk of ASD, emerge as a practical means for diagnosis. If ASD diagnosis can be effectively conducted through these videos, parents can readily identify potential ASD risks without the need for extended visits to medical institutes. Even in cases where video-based diagnoses result in false positives, seeking medical checks based on such diagnoses can help alleviate parental concerns. Furthermore, the rapid development of AI technology in video processing underscores the potential efficacy of this approach.

2.2. AI approaches in ASD diagnosis with video data as input

Building upon the significance highlighted in Section 2.1, the field of video-based diagnosis utilizing AI technologies stands out as a promising avenue in ASD studies. Particularly, a home video-based approach proves to be convenient for clinicians, patients, and their families. Consequently, the primary objective of this paper is to review studies focusing on AI approaches in ASD diagnosis utilizing video data as input.

This review paper specifically selected English-language papers pertinent to AI approaches in ASD diagnosis with video data as the primary input. The literature search was conducted using Scopus and PubMed, with a restriction on the publication year from 2018 to 2024. The initial search was performed on December 30th, yielding a total of 155 papers (133 from Scopus and 22 from PubMed).
The initial search query was ("Autism" OR "Autism Spectrum Disorder") AND ("Convolutional Neural Network" OR "Deep Learning" OR "Deep Neural Network" OR "Recurrent Neural Networks" OR "Deep belief networks" OR "Neural Network" OR "Multilayer neural networks") AND ("video"). Then, we narrowed down these papers based on the following process.

1. Remove duplicates
2. Papers that did not pertain to the domains of diagnosis, video, and AI (excluding machine learning) were excluded from the selection.

In the refinement process, 17 duplicated papers were initially excluded. Subsequently, 146 additional papers were excluded based on criteria 2. These excluded papers encompassed affect recognition by individuals with ASD, ASD therapy, images as input, and various traditional machine learning methods. Consequently, only 9 papers met the specified criteria and were included for review in this paper.

The summarized findings of these 9 papers are presented in Table 1. Three distinct input features to diagnose ASD were identified:

1. Features from facial videos (e.g., facial landmarks)
2. Features from behavioral videos (e.g., gait)
3. Multimodal features from videos (the combination of 1., 2. and other features like voice, and demographic and clinical data)

We divided the following sections based on the differences in these input features.

Table 1. AI approaches studies in ASD diagnosis with video data as input as published between 2018 and 2024

Study	Sample	Method	Input	Result
Tang et al., 2020	45 high-risk ASD and 43 TD (Chinese; Age 0-2)	NN (OpenPose) + CNN (MTCNN) + SVM	Multimodal data (Head-Movement, facial and vocal)	In the split test sample Accuracy = 96.4% Sensitivity = 95.0% Specificity = 97.7%
Kojovic et al., 2021	Dataset 1: 68 ASD and 68 TD (European; Age 1-3) Dataset 2: 101 ASD (European; Age 3-4, Out-of-sample)	NN (OpenPose) + CNN-LSTM	Behavioral video (Pose + Gait)	Dataset 1: In the split test sample Accuracy = 80.9% Sensitivity = 85.4% Specificity = 76.5% Dataset 2: Accuracy = 80.2%
Wu et al., 2021	133 infants (American; Age 0-3)	NN	Multimodal data (Look face rate + Social smile rate + social vocal rate age and gender)	In the split test sample Accuracy = 82.0% Sensitivity = 92.0% Specificity = 71.0%
Cai et al., 2022	57 ASD and 25 TD (Region and age NA)	CNN	Facial video	3-fold cross-validation

				Accuracy = 95.1% Sensitivity = 92.6% Specificity = 96.5%
Chanyoung et al., 2022	50 TD children and 44 ASD (Korean; Age 2-6)	CNN + LSTM	Facial video	In the split test sample Accuracy = 91.8% Sensitivity = 95.3%
Patankar et al., 2022	27 ASD children and 43 TD (Indian; Age 0-10)	CNN + RNN	Facial video	In the split test sample Accuracy = 90.5%
Saranya and Anandan., 2022	50 ASD and non-ASD subjects (Region NA; Age 5-8, 9-12, 13-16 and 45-50)	DEAF	Multimodal data (Facial emotion and gait)	In the split test sample Accuracy = 95.4% Sensitivity = 93.5% Specificity = 94.0% (Age 5-8) Accuracy = 96.0% Sensitivity = 94.5% Specificity = 94.5% (Age 9-12) Accuracy = 95.5%

				Sensitivity = 94.5% Specificity = 94.0% (Age 13-16) Accuracy = 96.5% Sensitivity = 94.5% Specificity = 95.0% (Age 45-50)
Henderson et al., 2023	Dataset 1: 50 ASD and 50 TD (Iraqi; childcare and kindergarten) Dataset 2: 30 ASD and 30 TD (Malaysian; Age 4-14)	CNN	Behavioral video (Gait)	Dataset 1: In the split test sample TAT accuracy = 95.6% Dataset 2: In the split test sample TAT accuracy = 80.0%
Prakash et al., 2023	400 ASD, 600 neurotypical and 250 Other Developmental Delay (Indian; Age 1.5-5)	R-CNN + DNN (DeepPose) + CNN	Behavioral video (Pose)	In the split test sample Accuracy = 79.7%

Abbreviations in the table: CNN = Convolutional Neural Network, DEAF = Deep Extreme Adaptive Fuzzy, DNN = Deep Neural Network, LSTM = Long Short-term Memory, NN = Neural Network, R-CNN = Region-based Convolutional Neural Network, RNN = Recurrent Neural Network, SVM = Support Vector Machine, TAT = Test Time Augmentation

2.2.1. ASD diagnosis with facial videos

Three studies were conducted to explore AI approaches in the diagnosis of ASD using raw videos as input.

Chanyoung et al. employed a combination of Long Short-term Memory (LSTM) and Convolutional Neural Networks (CNN) for feature extraction. CNNs, commonly utilized in image and video processing, apply a weight matrix, known as a kernel, to two-dimensional image data. This process compresses the original data into an attention map, preserving significant features while reducing dimensionality. The convolutional layers in this study were pretrained using the ImageNet dataset (Deng et al., 2016). In this dataset, images of humans, animals, plants, and inorganic substance objects are included. The CNN outputs were then fed into the LSTM architecture, a subset of Recurrent Neural Networks (RNNs) capable of utilizing memory in a cyclic manner. LSTMs incorporate a forget gate, enabling the retention of relevant information while discarding unnecessary data. Without LSTM architecture, RNNs encounter the vanishing gradient problem, resulting in excessively small gradients during network updates. The output from the LSTM architecture was subsequently processed through fully connected and dropout layers to generate predictions (ASD or not). Video frames from 50 typically developing (TD) children and 44 ASD children, aged 25 to 72 months, were obtained from the Child and Adolescent Psychiatry Division of Seoul National University Hospital and preschools in Korea. Parents of these subjects were instructed to sit in front of a camera with toys to capture joint attention skills. The proposed model was evaluated using a total of 918 video clips (TD = 484, ASD = 434), comparing ASD prediction accuracy between raw videos and background-removed videos. The results for the raw videos test dataset were as follows: Accuracy = 91.6%, Precision = 90.0%, Sensitivity = 94.5%, and F1 score = 92.2%. The background-removed videos achieved Accuracy = 91.8%, Precision = 88.0%, Sensitivity = 95.3%, and F1 score = 91.5%.

Patankar et al. utilized a combined CNN-RNN model for ASD diagnosis. The CNN component employed the Inception-V3 pretrained model from TensorFlow (Szegedy et al., 2015), consisting of convolution, pooling, dropout, and fully connected layers. Inception-V3 achieved high accuracy in object classification. In this study, they applied this object classifier to detect ASD and non-ASD. CNN architecture learned facial landmarks in the training process. These facial landmarks were shown as heatmap-like activation maps. The RNN architecture employed a Gated Recurrent Unit (GRU) layer, suitable for training models with time-series or sequence data like videos. Although the CNN-RNN model was not highly complex, it focused on high-level facial landmarks as essential features, making it convenient due to the lack of necessity for detailed facial features in the analysis. The study involved 27 ASD children and 43 TD children from the

Indian region who provided questionnaires and videos via a responsive web app, AutiScan. Three approaches—(1) CNN-RNN, (2) Long-term Recurrent Convolutional Network (LRCN), and (3) Convolutional LSTM (ConvLSTM)—resulted in test accuracies of 90.48%, 69.23%, and 53.85%, respectively.

Cai et al. employed a CNN architecture for the diagnosis of ASD. In this study, frames were sampled from videos, and OpenFace 2.0 (Baltrusaitis et al., 2018) was utilized to automatically detect facial landmarks, eye movement, and other facial behaviors. The data, comprising 709 dimensions obtained from OpenFace, underwent a feature selection process, wherein significant features were chosen for input into the CNN architecture. The CNN model ResNet-50 (He et al., 2016), a network consisting of 50 layers, was used for calculating ASD scores. The average score of the video frames was used for binary ASD diagnosis. Videos, recorded on the parents' phones, featured parents capturing the attention of their children through name-calling and the use of toys. Thus, recorded videos are similar to the paper from Chanyoung et al. This approach, involving short home videos with an average length of 18.74 seconds, was convenient for both subjects and researchers.

The dataset encompassed videos from 57 children with ASD and 25 TD children. The age of the subjects was not explicitly specified in this experimental context. Employing the top 100 significant features, their methodology achieved an accuracy of 95.06%, surpassing that of machine learning (ML) classifiers, such as Support Vector Machines (SVM), which achieved an accuracy of 75.62%. The study further conducted a comparative analysis of accuracy based on different input feature numbers. Notably, an accuracy of 91.40% and 90.17% were attained with 50 and 200 features, while 100 features resulted in an accuracy of 95.06%. This comparison underscored the significance of judicious feature selection in the diagnostic process.

2.2.2. ASD diagnosis with pose and gait videos

Kojovic et al. employed deep neural network-based pose estimation software and utilized deep learning classification for the diagnosis of ASD. To isolate only the skeletal information from the gathered videos, the researchers employed OpenPose (Cao et al., 2019), a software capable of detecting multiple human poses simultaneously. The deep learning architecture of OpenPose consists of multiple convolutional layers. Given the presence of not only the target subjects but also additional individuals in the videos, this multi-person pose estimation tool proved beneficial for the study.

Eighteen key points representing the detected skeleton in the videos were input into a CNN-LSTM deep neural network architecture. As the CNN architecture, VGG16 (Simonyan and Zisserman, 2015) was utilized, comprising 16 convolutional layers and pretrained with the ImageNet

dataset to extract features of the skeleton. Subsequently, after passing through the LSTM, ASD binary classification was performed using the softmax function.

For the experimental phase, two datasets were prepared. The first dataset comprised 68 children with ASD and 68 TD children, all aged between 1 and 3 years. The second dataset exclusively included 101 children with ASD aged between 3 and 4 years. The ASD prediction model trained on the first dataset exhibited an accuracy of 80.9%, a specificity of 76.5%, and a sensitivity of 85.4%. Notably, the second dataset yielded a comparable accuracy of 80.2%, affirming the robustness of their approach.

Henderson et al. proposed a novel diagnostic approach for ASD based on gait data utilizing CNNs. The architecture of their CNN is notably uncomplicated, comprising three layers. Nevertheless, their study diverges from others in terms of the specific focal point of the selected feature and the method employed for accuracy calculation. They introduced the Joint Energy Image (JEI) as the input for the CNN, derived from the initial extraction of 3D joint positions and trajectories from video data to measure the mobility of skeletons in the pixel. These 3D joint features were subsequently translated into 2-dimensional maps denoted as JEI.

In the context of ASD classification accuracy assessment, Test Time Augmentation (TAT) accuracy was employed. During TAT calculation, classification results were initially obtained using randomly modified (augmented) frame data. TAT accuracy represents the average accuracy across augmentations. The evaluation utilized two distinct datasets. The first dataset consisted of videos capturing straight-walking behavior from 68 individuals with ASD and 50 children from childcare and kindergarten settings. The second dataset encompassed 30 individuals with ASD and 30 TD children aged 4-14 from the National Autism Society of Malaysia center.

Training the respective datasets yielded TAT accuracy values of 95.56% and 80.00%. Additionally, conventional (non-TAT, without augmentations) accuracy values were reported, namely, 88.89% and 93.33%.

Prakash et al. employed Human Action Recognition (HAR) technology in the context of ASD diagnosis. Their HAR methodology is structured into three sequential phases, encompassing Human Detection, Temporal Action Localization, and Action Recognition. In the Human Detection phase, in order to identify the location of children, the Faster R-CNN (Girshick, 2015) was deployed, which is a CNN-based Region Proposal Network (RPN). Extracting features from the CNN, the RPN proposes object detection in the form of a 2-dimensional bounding box. The Faster R-CNN algorithm successfully identified children, play partners, and objects within the video content.

For the Temporal Action Localization phase, the Asynchronous Interaction Aggregation network (AIA; Tang et al., 2020) was employed. AIA, functioning as a network system, analyzes interactions between objects within the videos through accumulated transformer blocks. Notably, AIA exhibits limitations in detecting objects when interactions are absent; consequently, simultaneous pose-tracking using DeepPose (Toshev et al., 2014) was implemented. DeepPose detected 10 key points, such as the elbow and head, concurrently.

The final phase involved constructing a behavior action recognition model with 3-dimensional convolutional layers, drawing on previous work by Carreira and Zisserman (2018). The researchers amassed a dataset comprising videos from 400 individuals with ASD, 600 neurotypical patients (defined as low ASD possibility and high developmental quotient), and 250 individuals with other developmental delays (high-risk: low-risk = 125: 125). The interaction videos between children (aged 1.5-5) and therapists had durations ranging from 7 to 12 minutes. Following the training of the model with ASD and ODD cohorts, it demonstrated accuracy rates of 79.7%, 77.2%, and 80.8% in the detection of ASD, ODD, and neurotypical cases, respectively.

2.2.3. ASD diagnosis with multimodal features from videos

Tang et al. developed a High-risk (HR) ASD classifier employing video and audio data. Unlike other studies, this study investigated HR-ASD which is not ASD patients. For HR-ASD diagnosis, the authors extracted head movement, facial appearance, and vocal data from the provided videos. The OpenPose framework was utilized to extract features related to head movement. This involved investigating the movement of the head based on detected locations such as eyes, ears, and nose. Concurrently, the CNN-based architecture MTCNN (Zhang, 2016) and OpenFace were applied to extract facial appearance features. In this phase, 68 facial landmarks were identified from faces detected by MTCNN.

In addition to the head and facial features, a 384-dimensional dataset pertaining to vocal characteristics, including frequency, energy, and spectrum, was extracted from the videos. Subsequently, these features were input into a SVM, a machine learning method. Videos spanning 2 minutes each were recorded for 45 HR-ASD and 43 TD infants aged 8-24 months. The recordings involved one frontal and two non-frontal cameras capturing the subjects, with parents seated in front of the infants.

The SVM classification achieved notable performance metrics, including an accuracy of 96.39%, sensitivity of 95.00%, specificity of 97.67%, and an Area Under the Curve (AUC) of 94.59% for HR-ASD classifications using head movement, facial appearance, and vocal features. Remarkably, identical results were obtained even when excluding head movement features.

Saranya and Anandan proposed an ASD diagnosis methodology that integrates facial emotion analysis and human gait assessment using the Deep Extreme Adaptive Fuzzy (DEAF) learning algorithm. The DEAF model consists of two key algorithms: 1. Feature extraction through Convolutional Neural Network (CNN) and 2. Fuzzy-fused Extreme Learning Machine (ELM).

Given the well-established distinctions in facial features between individuals with ASD and those without, coupled with the specific gait patterns observed in individuals with ASD (Calhoun et al., 2011), the CNN model was developed to extract features from both facial expressions and human gaits. The facial feature extraction phase involved the classification of seven emotions (Anger, Disgust, Happiness, Sadness, Surprise, and Neutral). Simultaneously, gait features such as swings of hands (swing ratio), number of steps per minute (cadence), velocity, and others were extracted. Subsequently, these features were input into the Fuzzy-based Extreme Learning Machine (FELM). The Extreme Learning Machine (ELM) is a single hidden layer network (Huang and Chen, 2007) renowned for its ability to minimize training errors and enhance approximation. Given the diverse nature of ASD cases, the authors introduced "fuzzy" to the ELM, essentially transforming it into a fuzzy inference system.

The study involved the analysis of videos from 50 subjects across various age groups (5-8, 9-12, 13-16, 45-50). Facial emotions were classified with an accuracy of 89.0% on the Karolinska Directed Emotional Faces (KDEF) datasets, while human gait movements were detected with 90.0% accuracy on the Chinese Academy of Sciences, Institute of Automation (CASIA) dataset. Specific to the age group 5-8, ASD prediction accuracy for facial emotions, human gaits, and fused features reached 87.5%, 88.5%, and 95.4%, respectively. For the age group 45-50, accuracies of 88.0%, 88.5%, and 96.5% were achieved for facial emotions, human gaits, and fused features, respectively. Thus, results do not show age dependency.

Wu et al. proposed an approach for ASD diagnosis based on Machine Learning (ML) and Deep Neural Network (DNN) analysis of facial videos. Notably, raw videos were not utilized in training the Neural Network, a decision attributed to the authors' possession of a limited video dataset and a comprehensive understanding of distinctive ASD signs. As a machine learning method, recursive Feature Elimination (RFE), Ridge Regression (RR), Mutual Information Estimation (MI), and Kolmogorov Smirnov Test (KS) were chosen as ML methodologies for selecting significant features. All these techniques are adept at discerning influential values from multiple variables, with RFE and RR assuming normal distribution in input features, while MI and KS do not. ML techniques selected features, including look face rate, vocal rate, smile rate, age, and gender. For ASD diagnosis, these features were fed into a Neural Network (NN) consisting of three fully connected layers. The

performance of ML-based feature selection was compared with that of using all features and two sampling methods: Synthetic minority Over-sampling and Tomek Links under-sampling. Over and under-sampling were implemented due to the significant disparity in the number of ASD and non-ASD samples.

The study involved the collection of 3-minute videos featuring infants aged 3 to 36 months, along with their parents, from the UC Davis MIND Institute. The videos, totaling 1707, were manually labeled for signs such as smiles. Ultimately, 547 videos from 133 infants were employed. The results indicated that over-and-under-sampling with all features and NN yielded the highest accuracy at 82%, outperforming other methods (51% for selected features).

The authors also proposed two Deep Learning (DL) methods for detecting behavioral events. The first method employed image and DNN-based detection, utilizing ResNet-18, a CNN with 18 layers to classify behavioral events like smiling, look face, and look objects. This ResNet-18 is not deeper than ResNet-50. However, deeper architecture is not necessarily more accurate. Therefore, choosing the best depth architecture is the most significant. Frame images from videos served as input to this DNN model. The second method employed OpenFace 2.0 (Baltrusaitis et al., 2018) to detect behavioral events using facial landmarks. The first method achieved accuracies of 70%, 68%, 67%, and 53% for manually annotated smile, look face, look object, and vocal detection, respectively. The second tool achieved accuracies of 68.5%, 66.0%, and 50.0% for smile, look face, and vocal detection from videos. Combining these detection tools with the proposed NN demonstrated effectiveness in diagnosing ASD. These automatically detected features were not utilized in the ASD diagnosis in this study.

3. Discussion

3.1. Discussion and comparison of nine video-based AI approaches for ASD diagnosis

The nine aforementioned papers possess both advantages and disadvantages. Regarding video data collection methods, a majority of the studies recorded videos in controlled settings such as laboratories or hospitals. Conducting experiments and diagnoses in these facilities incurs costs and time. In this context, the innovation introduced by Patankar et al., known as Autiscan, is noteworthy. Autiscan, being a web application, enables the easy involvement of subjects and their parents in the experimental process. This application accommodates questionnaires with videos acquired at home. In addition to videos, experts utilize questionnaires for ASD diagnosis, thereby potentially enhancing diagnostic accuracy and mitigating false positive and negative outcomes. Cai et al. similarly obtained semistructured videos from parents of subjects, providing explicit recording instructions for parents to attract

their children's attention through name-calling or the use of toys. Such succinct instructions are imperative for clinical applications. Conversely, Tang et al.'s study necessitated the deployment of three cameras to capture detailed facial movements around subjects, demanding the involvement of clinical experts. Nevertheless, it holds promise for expediting and simplifying the diagnostic process.

Subjects across the experiments exhibit differences among the papers. Five studies (Tang et al., Kojovic et al., Wu et al., Chanyoung et al., and Prakash et al.) focused on subjects under six years old, while other papers encompass subjects older than six. As underscored, early diagnosis is imperative for timely therapeutic interventions and parental support, yet diagnosing ASD at an early age is inherently more challenging due to delayed manifestation of certain symptoms in social interactions. Tang et al. specifically investigated HR-ASD, achieving a notable accuracy of 96.39%. However, this approach does not provide a pure ASD diagnosis applicable across the entire spectrum. Conversely, Chanyoung et al. achieved a high ASD diagnosis accuracy of 91.8% for subjects aged 2-6. Nevertheless, clinical experts at these ages can also achieve decent diagnoses using interviews and questionnaires. A study focusing on interviews and questionnaires for ASD diagnosis in children aged 1-4 (Christiansz et al., 2016) disclosed that the modified DSM-5 criteria yielded a sensitivity (true positive rate) of 97% and specificity (true negative rate) of 41%. It is noteworthy to underscore the importance of statistical measures in this context. Simultaneously achieving high sensitivity and specificity poses challenges for both human evaluators and AI. The occurrence of false negatives is a concern, as they can carry stigmatizing implications for children. Consequently, in the context of clinical applications, the judicious selection of statistical measures holds significant relevance.

As previously discussed, further investigation is warranted for the diagnosis of infants, and numerous studies have been conducted on early ASD diagnosis. An inquiry focusing on subjects aged 2 years old (Charman et al., 2005) reported that, out of 26 subjects diagnosed with ASD using the Autism Diagnostic Interview-Revised (ADI-R), 22 received a consistent ASD diagnosis at the age of 9 years. Thus, without adjustment to age, traditional ASD diagnosis can achieve accurate diagnosis to some extent. Another study (Luyster et al., 2009) adapted the Autism Diagnostic Observation Schedule for children under 30 months, achieving a specificity of 93% and sensitivity of 95%. Luyster's study involved modifications to the experimental design and criteria tailored for infants displaying fewer reactions than older children. Similarly, for the further development of video-based ASD diagnosis approaches, experimental methods, such as recording techniques with toys, should be tailored to specific target age groups.

The majority of the studies were geographically specific, leading to possible biases in diagnosis or video samples based on cultural differences. A study about cultural differences in ASD diagnosis reported

that ASD and non-ASD thresholds can be different because of the degree of development expected by others including parents (Matson et al., 2011). Chanyoung et al. conducted experiments exclusively with Korean cohorts. Furthermore, Patankar et al.'s motivation stemmed from the observation that ASD studies in the Indian region lag behind those conducted with European and US samples. In practical AI use outside of ASD diagnosis, in 2020, an AI algorithm wrongly arrested a black male because of biased training (Perkowitz et al., 2021). Despite ASD not being inherently region-specific, the potential for AI to inadvertently learn region-specific parameters during training underscores the importance of considering overfitting to training samples in AI studies.

The nine papers focused on distinct features for diagnosis, with facial studies demonstrating an accuracy range of approximately 90%-95%, pose or gait studies exhibiting about 80%-95% accuracy, and multimodal studies yielding results in the range of 80%-95%. Nevertheless, the challenge associated with obtaining high-quality video data varies across different features, with behavioral data posing greater difficulties in processing and standardization in comparison to facial data, owing to the potential inclusion of extraneous information. Consequently, it appears that facial studies possess certain advantages. However, for the sake of diagnostic robustness, multimodal studies potentially mitigate the risk of false negatives, given the diversity of symptoms and severity in the ASD population that an ASD diagnosis often necessitates consideration of multiple features rather than relying on a single aspect.

Accuracy may also be contingent on sample size, particularly in the realm of AI research. Most papers involved approximately 100 participants (ASD: TD = 50: 50). While multiple videos and frames were collected from each participant, training neural networks with such limited participants raises concerns about potential overfitting to the involved cohorts. Moreover, the global prevalence of ASD is approximately 1%, introducing a skewed learning environment due to the discrepancy in ASD ratios across the world population. Furthermore, it is essential to include related disorders such as ADHD and Schizophrenia in the sample and ensure high sensitivity in their diagnosis simultaneously. These considerations underscore the significance of out-of-sample replication. With the exception of the study by Kojovic et al. (2021), the video samples were typically divided into training and test datasets. Following the training of AI with the training dataset, performance validation occurred using the test dataset. Consequently, the performance of neural networks was not corroborated with an out-of-sample dataset, potentially leading to unintentionally high accuracy due to the risk of overfitting to the sample dataset.

Additionally, as highlighted in the introduction, autism is a spectrum disorder. Consequently, the binary classification of input data and diagnosis poses a challenge in clinical applications. One potential solution involves the utilization of a numeric score in ASD diagnosis. Many DL studies typically employ the softmax function at the conclusion of the DL

architecture. However, by omitting the softmax functions and implementing standardization, a numeric score can be derived, effectively representing the spectrum of ASD severity. While not an AI diagnostic method, the Childhood Autism Rating Scale also offers numeric scores in clinical applications (Chlebowski et al., 2010).

In summary, the video-based AI approach in ASD diagnosis exhibits numerous advantages and challenges, as discussed herein.

3.2. Machine learning approaches in ASD diagnosis

Despite the development of advanced AI methods, such as Deep Learning and Deep Neural Networks, the application of machine learning is anticipated to remain valuable across various research fields. Machine learning methods have also been employed for video-based ASD diagnosis and may provide future direction for research using advanced AI methods, such as additional input features. It is noteworthy that ML methods are generally less complex than DL methods. This characteristic renders ML more interpretable for humans.

Li et al. (2018) employed Support Vector Machines (SVM) by inputting eye-tracking data from videos. Given the distinctive eye movement patterns associated with individuals with ASD, SVM was trained with videos featuring 53 ASD and 136 TD children. This approach achieved a noteworthy 93.7% accuracy in predicting ASD solely based on eye-tracking videos, showcasing the efficacy of the combination of eye-tracking data and machine learning. However, it is important to note that recording eye-tracking videos may present challenges in household settings when compared to obtaining behavioral videos.

An alternative approach involves the creation of a mobile web portal for ASD diagnosis utilizing video data (Tariq et al., 2018). This portal leverages multiple features, such as expressive language, eye contact, and echolalia, extracted from 3-minute home videos. Three types of machine learning methods—Decision Tree, Logistic Regression, and SVM—were applied to the ASD diagnosis process. The evaluation encompassed 116 ASD and 46 non-ASD participants. Notably, one Decision Tree method demonstrated sensitivity, specificity, and accuracy rates of 100%, 22.4%, and 76.1%, respectively, while one Logistic Regression method exhibited rates of 94.5%, 77.4%, and 88.9%. This outcome underscores two significant findings: the potential for logistic regression to achieve high prediction accuracy and the observation that, despite high sensitivity (true positive rate), specificity (true negative rate) can be comparatively low even with AI approaches. The ease of capturing videos in home environments renders this approach promising for accessible ASD diagnosis.

3.3. Other modalities for AI approaches in ASD diagnosis

In conjunction with the video approach, diverse AI methodologies have been applied to the diagnosis of ASD using non-video-based modalities, focusing on neurodevelopmental aspects such as the analysis of brain regions and neurons and vocal characteristics.

Functional Magnetic Resonance Imaging (fMRI) serves as a non-invasive technology for studying the brain, and research employing fMRI images has yielded valuable insights into functional connections (FCs) within the brain. A study conducted by Shao et al. (2021) revealed that FCs in individuals with ASD exhibited weakened connections in the cerebral hemisphere. Through the utilization of Deep Feature Selection (DFS), ASD-specific FCs were identified, and the application of a Graphical Convolutional Network (GCN) resulted in an ASD diagnosis accuracy of 79.5%. It is important to note, however, that the recording of fMRI data is considerably more expensive and inconvenient compared to the collection of home videos that can achieve higher accuracies in ASD diagnosis.

Audio data has also been harnessed for ASD diagnosis through the implementation of AI algorithms. Researchers employed Long Short-term Memory (LSTM) and Synthetic Random Forest to construct a framework for predicting the Autism Diagnostic Observation Schedule (ADOS-2) Calibrated Severity Score (CSS) of Social Affect (SA) using audio data, achieving an R^2 value of 0.402 (Sadiq et al., 2019). This outcome, considering the sole utilization of audio data as input, indicates promising potential for future research endeavors, and can be incorporated as additional features derived from home videos, such as the use of vocal rate seen in Wu et al., (2021).

Several characteristics of ASD are anticipated to be integrated into the diagnostic process. However, the challenge associated with the utilization of AI lies in its opacity. While the AI approach can attain a heightened level of diagnostic accuracy and effectively manage extensive datasets, the complexity of neural networks renders it challenging for researchers to elucidate the rationale behind diagnostic decisions. Consequently, the adoption of the latest technology known as explainable AI (XAI) is envisaged in clinical settings. XAI is a technological advancement capable of highlighting crucial regions, parameters and neurons within the architecture and input. A prior investigation (Alam et al., 2023) employed DL methods in conjunction with XAI for ASD diagnosis based on images. This study not only achieved a commendable predictive accuracy of 98.9% but also implemented the XAI technology referred to as Grad-CAM (Selvaraju et al., 2020). Grad-CAM is applicable to CNN architectures and, during image processing, visualizes noteworthy features within the images as a heatmap. Regions such as the forehead, the center of the eyes, the nose, and the lips were identified as areas frequently utilized in the learning process. Similar explainable artificial intelligence (XAI) techniques, including Grad-CAM, as well as other XAI methods, were

employed in an activity recognition study utilizing video data (Hiley et al., 2020). In this study, XAI methods elucidated the contribution of each pixel in the input frame. For example, relevant background information is explained as unnecessary information for activity recognition. The application of such XAI methods holds potential for future studies in ASD diagnosis in selecting the best features and modalities.

The prospect of integrating these methods with a video-based ASD diagnosis is anticipated.

4. Conclusion

This review paper delves into AI approaches employed in the diagnosis of ASD using video data. While ASD has been extensively studied, traditional diagnostic methods, reliant on interviews and questionnaires, exhibit certain limitations, such as prolonged diagnostic durations and questionable accuracy at very early ages. In response to these challenges, there has been an exploration of a video-based approach incorporating AI applications. The nine examined papers focused on facial, pose, or gait, as well as multimodal features within videos for ASD detection, showcasing high accuracy in predicting diagnosis. However, there remains potential for further development to achieve practical clinical applications.

It is noteworthy that many studies have collected video data in controlled environments, such as medical institutions or laboratories. Yet, for leveraging video data effectively, the future development of home video-based approaches is anticipated. To realize this objective, researchers must establish a robust protocol with clear instructions for parents to capture semi-structured home videos of their children for diagnostic purposes. From an AI perspective, numerous challenges exist, including sample biases and the black box problem in training. As discussed in the review, XAI holds promise in addressing these issues, although the technology is still in its developmental stages. Consequently, additional investigations involving larger training sample sizes and out-of-sample replication, or using alternative AI technologies for video processing are deemed necessary for the continued advancement of this research field.

5. References

1. Alam, M. S. *et al.* Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum Disorder Diagnosis from Facial Images Using Explainable AI. *Technologies* **11**, 115 (2023).

2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®), 5th ed.; American Psychiatric Association: Arlington, VA, USA, 2013; pp. 1–947.

3. Awatramani, J. & Hasteer, N. Facial Expression Recognition using Deep Learning for Children with Autism Spectrum Disorder. in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)* 35–39 (2020). doi:10.1109/ICCCA49541.2020.9250768.

4. Baltrusaitis, T., Zadeh, A., Lim, Y. C. & Morency, L.-P. OpenFace 2.0: Facial Behavior Analysis Toolkit. in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* 59–66 (2018). doi:10.1109/FG.2018.00019.

5. Brentani, H. *et al.* Autism spectrum disorders: an overview on diagnosis and treatment. *Braz. J. Psychiatry* **35**, S62–S72 (2013).

6. Cai, M. *et al.* An Advanced Deep Learning Framework for Video-Based Diagnosis of ASD. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (eds. Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) 434–444 (Springer Nature Switzerland, 2022). doi:10.1007/978-3-031-16440-8_42.

7. Calhoun, M., Longworth, M. & Chester, V. L. Gait patterns in children with autism. *Clin Biomech (Bristol, Avon)* **26**, 200–206 (2011).

8. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. Preprint at https://doi.org/10.48550/arXiv.1812.08008 (2019).

9. Carreira, J. & Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Preprint at https://doi.org/10.48550/arXiv.1705.07750 (2018).

10. Chanyoung, K. *et al.* AI-assisted Initiation to Joint Attention Evaluation for Autism Spectrum Disorder Detection. in *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)* 1–6 (2022). doi:10.1109/ICHMS56717.2022.9980778.

11. Charman, T. *et al.* Outcome at 7 years of children diagnosed with autism at age 2: predictive validity of assessments conducted at 2 and 3 years of age and pattern of symptom change over time. *J Child Psychol Psychiatry* **46**, 500–513 (2005).

12. Chlebowski, C., Green, J. A., Barton, M. L. & Fein, D. Using the Childhood Autism Rating Scale to Diagnose Autism Spectrum Disorders. *J Autism Dev Disord* **40**, 787–799 (2010).

13. Christiansz, J. A., Gray, K. M., Taffe, J. & Tonge, B. J. Autism Spectrum Disorder in the DSM-5: Diagnostic Sensitivity and Specificity in Early Childhood. *J Autism Dev Disord* **46**, 2054–2063 (2016).

14. Dawson, G. *et al.* Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder. *Sci Rep* **8**, 17008 (2018).

15. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/CVPR.2009.5206848.

16. Gepner, B., Deruelle, C. & Grynfeltt, S. Motion and Emotion: A Novel Approach to the Study of Face Processing by Young Autistic Children. *J Autism Dev Disord* **31**, 37–45 (2001).

17. Girshick, R. Fast R-CNN. *arXiv.org* https://arxiv.org/abs/1504.08083v2 (2015).

18. Goodfellow, I. J. *et al.* Challenges in Representation Learning: A Report on Three Machine Learning Contests. in *Neural Information Processing* (eds. Lee, M., Hirose, A., Hou, Z.-G. & Kil, R. M.) 117–124 (Springer, 2013). doi:10.1007/978-3-642-42051-1_16.

19. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in 770–778 (2016).

20. Henderson, B., Yogarajah, P., Gardiner, B. & McGinnity, T. M. Encoding Kinematic and Temporal Gait Data in an Appearance-Based Feature for the Automatic Classification of Autism Spectrum Disorder. *IEEE Access* **11**, 134100–134117 (2023).

21. Hiley, L. *et al.* Explaining Motion Relevance for Activity Recognition in Video Deep Learning Models. Preprint at https://doi.org/10.48550/arXiv.2003.14285 (2020).

22. Huang, G.-B. & Chen, L. Convex incremental extreme learning machine. *Neurocomputing* **70**, 3056–3062 (2007).

23. Kleinman, J. M. *et al.* The Modified Checklist for Autism in Toddlers: A Follow-up Study Investigating the Early Detection of Autism Spectrum Disorders. *J Autism Dev Disord* **38**, 827–839 (2008).

24. Kojovic, N., Natraj, S., Mohanty, S. P., Maillart, T. & Schaer, M. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Sci Rep* **11**, 15069 (2021).

25. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

26. Li, J., Zhong, Y. & Ouyang, G. Identification of ASD Children based on Video Data. in *2018 24th International Conference on Pattern Recognition (ICPR)* 367–372 (2018). doi:10.1109/ICPR.2018.8545113.

27. Luyster, R. *et al.* The Autism Diagnostic Observation Schedule – Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders. *J Autism Dev Disord* **39**, 1305–1320 (2009).

28. Martin, K. B. *et al.* Objective measurement of head movement differences in children with and without autism spectrum disorder. *Molecular Autism* **9**, 14 (2018).

29. Matson, J. L. & Kozlowski, A. M. The increasing prevalence of autism spectrum disorders. *Research in Autism Spectrum Disorders* **5**, 418–425 (2011).

30. Matson, J. L. *et al.* A multinational study examining the cross cultural differences in reported symptoms of autism spectrum disorders: Israel, South Korea, the United Kingdom, and the United States of America. *Research in Autism Spectrum Disorders* **5**, 1598–1604 (2011).

31. Okoye, C. *et al.* Early Diagnosis of Autism Spectrum Disorder: A Review and Analysis of the Risks and Benefits. *Cureus* **15**, e43226.

32. Patankar, R., Vedpathak, S., Thakre, V., Sethi, P. & Sawarkar, S. AutiScan: Screening of Autism Spectrum Disorder Specific to Indian Region. in *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)* 1–8 (2022). doi:10.1109/GCAT55367.2022.9972038.

33. Perkowitz, S. The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing* (2021) doi:10.21428/2c646de5.62272586.

34. Prakash, V. G. *et al.* Video-based real-time assessment and diagnosis of autism spectrum disorder using deep neural networks. *Expert Systems* **n/a**, e13253.

35. Robins, D. L., Fein, D., Barton, M. L. & Green, J. A. The Modified Checklist for Autism in Toddlers: An Initial Study Investigating the Early Detection of Autism and Pervasive Developmental Disorders. *J Autism Dev Disord* **31**, 131–144 (2001).

36. Sadiq, S. *et al.* Deep Learning Based Multimedia Data Mining for Autism Spectrum Disorder (ASD) Diagnosis. in *2019 International Conference on Data Mining Workshops (ICDMW)* 847–854 (2019). doi:10.1109/ICDMW.2019.00124.

37. Saranya, A. & Anandan, R. FIGS-DEAF: an novel implementation of hybrid deep learning algorithm to predict autism spectrum disorders using facial fused gait features. *Distrib Parallel Databases* **40**, 753–778 (2022).

38. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis* **128**, 336–359 (2020).

39. Shao, L., Fu, C., You, Y. & Fu, D. Classification of ASD based on fMRI data with deep learning. *Cogn Neurodyn* **15**, 961–974 (2021).

40. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556 (2015).

41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. Preprint at https://doi.org/10.48550/arXiv.1512.00567 (2015).

42. Tang, C. *et al.* Automatic Identification of High-Risk Autism Spectrum Disorder: A Feasibility Study Using Video and Audio Data Under the Still-Face Paradigm. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **28**, 2401–2410 (2020).

43. Tang, J., Xia, J., Mu, X., Pang, B. & Lu, C. Asynchronous Interaction Aggregation for Action Detection. Preprint at https://doi.org/10.48550/arXiv.2004.07485 (2020).

44. Tariq, Q. *et al.* Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLOS Medicine* **15**, e1002705 (2018).

45. Toshev, A. & Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1653–1660 (2014). doi:10.1109/CVPR.2014.214.

46. WHO Homepage. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders. Last accessed 14 January 2023.

47. Wu, C. *et al.* Machine Learning Based Autism Spectrum Disorder Detection from Videos. in *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)* 1–6 (2021). doi:10.1109/HEALTHCOM49281.2021.9398924.

48. Zeidan, J. *et al.* Global prevalence of autism: A systematic review update. *Autism Res* **15**, 778–790 (2022).

49. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **23**, 1499–1503 (2016).