Advancements in Automated Assessment and Diagnosis of Autism Spectrum Disorder Through Multimodality Sensing Technologies: Survey of the Last Decade

Athmar N. M. Shamhan, Marwa Qaraqe, and Dena Al-Thani, *Member, IEEE*

***Abstract*—Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder characterized by difficulties in social interaction, communication, and repetitive behavior patterns. Traditional research approaches have primarily focused on studying autism using single-modal data analysis, such as relying solely on audio, video, and neuro signals. However, recent advancements in technology, cognitive science, and artificial intelligence (AI) have provided opportunities to explore the potential benefits of multisensory integration and fusion of modalities in understanding autism patterns. This survey makes three key contributions to advancing the future of ASD diagnosis and intervention. First, it provides a comprehensive review of recent advancements in multimodal sensing technologies, detailing primary modalities, data cleaning and synchronization techniques, feature extraction, and fusion methodologies to integrate diverse sensory data. Second, it classifies assistive technologies into three major categories: 1) computer-based systems; 2) virtual reality simulations; and 3) robotic interactions, analyzing their applications for cross-referencing symptoms and enabling real-time interventions in skills assessment and therapy. Third, it identifies critical challenges related to data collection, sensor synchronization, standardizing assessment paradigms, and real-time processing demands, proposing actionable future directions to improve diagnostic precision, scalability, and adaptability. These contributions underscore the transformative potential of multimodal sensing systems to revolutionize ASD assessment and diagnosis by enabling comprehensive, objective, and tailored solutions for diverse individuals across the autism spectrum.**

***Index Terms*—Autism, multimodal sensing, virtual reality (VR).**

Received 12 February 2025; revised 24 April 2025; accepted 21 May 2025. Date of publication 26 May 2025; date of current version 13 August 2025. Recommended for acceptance by Associate Editor Min Wu. *(Corresponding author: Athmar N. M. Shamhan.)*

The authors are with the A-Sense Center of Excellence, College of Computer Science and Engineering, Hamad Bin Khalifa University, Doha 52, Qatar (e-mail: atsh51045@hbku.edu.qa).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCDS.2025.3574145, provided by the authors.

Digital Object Identifier 10.1109/TCDS.2025.3574145

I. Introduction

AUTISM spectrum disorder (ASD) is a neurodevelopmental condition characterized by difficulties in social communication, repetitive behaviors, and narrow interests [1]. Recent findings from the Centers for Disease Control (CDC) indicate a high prevalence of autism in the USA, with 27.6 cases per 1000 children aged eight years, translating to one in 36 children [2]. This prevalence represents an increase compared to estimates from the autism and developmental disabilities monitoring (ADDM) Network spanning 2000 to 2018. While the precise cause of autism remains elusive, the research underscores its heritability at 0.82, with nonshared genetic factors contributing 0.18 [3].

ASD symptoms usually appear between the ages of two and three and continue into adulthood [4]. The diagnostic criteria of autism according to the Diagnostic and Statistical Manual of Disorders—DSM-TR [7] include: 1) deficits in social communication and interaction; and 2) restricted and repetitive patterns of behaviors. Common social difficulties include reduced eye contact, challenges in interpreting social cues, and difficulty engaging in reciprocal conversations. Repetitive behaviors often include hand-flapping, insistence on routines, or intense focus on specific interests.

The methods used for diagnosing ASD can vary significantly, with some countries implementing prediagnostic screenings to spot and refer possible cases of ASD for specialized care. However, inconsistencies in training can lead to delays in diagnosis [13], [14]. The diagnostic process itself heavily relies on clinical judgment, involving evaluations of a child's developmental stages based on behaviors, communication abilities, and social skills. This evaluation uses data from standardized tools, direct observations, and feedback from third parties [16]. Despite these comprehensive evaluations, the diagnosis of ASD still lacks definitive biological markers, making it dependent on the clinical expertise and subjective assessments of healthcare professionals. Recent advancements in technology, particularly in automated and assistive domains such as computer tasks, virtual reality (VR), and robotics, have opened new avenues for assessing the diagnostic criteria of autism. Essentially, machine learning led to the development of complex models capable of automatically analyzing labeled data to identify potential biomarkers for ASD [15].

The importance of automatic tools in autism research and diagnosis is underscored by their ability to objectively measure external behaviors such as joint attention, body motion, and facial expressions. For example, a study in [17] utilized eye-tracking (ET) technology to identify visual processing differences in adults with high-functioning autism, achieving a diagnostic accuracy rate of approximately 74% during web page navigation tasks. This demonstrates a significant improvement over traditional observational methods, which often rely heavily on subjective interpretation. Another study by [18] aimed at early identification of infants at high risk for autism (HR-ASD) by utilizing the still-face paradigm (SFP) to induce social stress, observing behavioral cues such as head movements, facial expressions, and vocal characteristics in both HR-ASD and typically developing (TD) infants. The results showed over 90% accuracy, specificity, and sensitivity, highlighting the potential for more efficient and objective screening methods in HR-ASD compared to conventional approaches.

Brain development in infancy with autism differs from that of TD children, even before behavioral signs appear [19], [20]. Studies of high-risk infant siblings consistently show that the first noticeable signs of social communication difficulties linked to autism do not appear until late in the first year of life [21]. Thus, using behavioral data alone to detect autism might miss the crucial window for early intervention. This has prompted researchers to explore the use of neurophysiological signals, such as electroencephalography (EEG) imagine technique, to identify autism in its early stages [22], [24]. Moreover, monitoring physiological changes associated with negative emotions in people with autism can provide valuable insights into their internal states, particularly for those with Alexithymia, who may have difficulty identifying and expressing their emotions [25]. Thus, this can help caregivers better understand their emotional states and provide timely support. However, these studies were able to make advancements in the automated assessment of autism. However, relying solely on single modality data such as EEG and eye tracking in the context of autism still has some challenges for three main reasons summarized in the following sections.

A. Heterogeneity of Autism

Based on DSM-TR [7], the term "spectrum" reflects the broad range of symptoms and severity levels that individuals with autism may exhibit. The spectrum concept recognizes the diversity of presentations and the unique characteristics of each person with autism. This diversity in autism symptoms poses challenges when we only rely solely on single modality data in different dimensions. First, it influences diagnosis, as focusing on a single aspect, such as attention and facial expression, may result in incomplete diagnoses due to the wide range of autism variations. Additionally, this diversity impacts treatment, as unimodal research may not cover all the different factors contributing to autism, making it harder to develop effective therapy. Moreover, given the wide variations in how autism presents in individuals, personalized interventions are necessary. However, unimodal research, with its single focus, may not provide the insights needed to tailor treatments to each person's specific needs. To this end, understanding autism as a spectrum allows for a more comprehensive and individualized approach to diagnosis, intervention, and support.

B. Standard Diagnostic Complexity

The Autism Diagnostic Observation Schedule (ADOS) [26] is internationally recognized as the gold standard for diagnosing ASD [27]. ADOS utilizes a combination of structured and semistructured tasks that engage individuals in social and communicative interactions with an examiner. These tasks are designed to assess key areas such as communication skills, social interactions, play, and the imaginative use of materials, providing essential insights crucial for diagnosing autism in both children and adults. By integrating the results from ADOS, clinicians can precisely determine the presence and severity of ASD traits. Given the comprehensive scope of ADOS, which examines multiple behavioral dimensions, it clearly illustrates the limitations of single-modal data in capturing the full spectrum of autism. This complexity highlights the necessity for researchers and clinicians to adopt multimodal approaches. In clinical settings, therapists consider a variety of factors to make well-informed decisions, underscoring that a singular focus may not sufficiently address the nuanced understanding or effective assessment of ASD.

C. Data Collection and Validation Challenges

In autism research, data collection is notably challenging, especially when dealing with specific subgroups such as individuals feeling uncomfortable with wearable devices, individuals with cooccurring hyperactivity or infants. Capturing data from these groups can be particularly difficult due to factors such as heightened activity levels and inherent communication limitations in infants [43], [60]. As a consequence, the collected data may be less reliable and robust. Researchers often face the need to exclude certain samples from analysis, especially those where the experimental procedures were not executed with precision or completeness [43]. This exclusion is necessary to maintain the integrity of the study, but it comes at the cost of reducing the sample size and creates a notable gap in validating results. Validation becomes challenging when dealing with unreliable data from a single source. The lack of cross-modality validation limits the confidence in the findings.

The need for multimodality in autism research arises from the inherent diversity of the autism spectrum and the challenges posed by current diagnostic and data collection methods. By integrating multiple modalities—such as behavioral data (e.g., facial expressions and body gestures), physiological signals [e.g., heart rate variability (HRV) and skin conductance], and neurophysiological metrics (e.g., EEG patterns)—multimodal systems provide a holistic view of an individual's responses, capturing both subtle and pronounced autism-related traits. This approach mirrors the multidimensional assessments of tools such as ADOS, enabling personalized diagnostics and interventions while reducing reliance on subjective clinical expertise. Furthermore, multimodality addresses challenges such as incomplete and noisy data by enabling cross validation and leveraging complementary data streams, ensuring robustness and reliability in research findings. By bridging gaps and accommodating diverse presentations, multimodal data pave the way for scalable, precise, and individualized care strategies in autism assessment and intervention.

This review delves into the utilization of multimodal sensor technologies for the automatic diagnosis, assessment, and intervention of autism, outlining three primary research questions.

1) Which sensing modalities and feature extraction techniques are predominant in this research area, and how are they utilized, processed, and fused?
2) How do various multimodal sensing technologies, including computer-based tasks, VR simulations, and robotic interactions, contribute to understanding ASD?
3) What are the main challenges and future direction in the research of multimodality in the context of autism?

Our survey shares common goals with several previous reviews referenced in [8], [9], [10], [11], and [12], but it stands out in three key ways. First, unlike these studies that often focus on a single modality such as physiological data from electrocardiography (ECG), photoplethysmography (PPG), electromyography (EMG), polysomnography (PSG), and EEG for mental health analysis [8], our approach is broader, incorporating a wide range of sensor technologies and their applications in both behavioral and physiological signals. Second, while other surveys concentrate on specific settings such as VR [9], only discussing design and guidelines for individuals with autism, or on robotics for emotion recognition in interventions for children with ASD [11], [12], our survey encompasses these diverse environments collectively. We examine their integration and utility in creating more dynamic and effective tools for diagnosis, assessment, and therapy. Third, we provide a detailed analysis of critical technical aspects related to multimodality, including preprocessing, synchronization, and data fusion techniques. This focus highlights the challenges and methodologies for aligning and integrating data from diverse modalities to ensure robustness and accuracy. Finally, our survey critically identifies existing gaps and challenges within current research, offering strategic insights and proposing actionable directions to guide future advancements in this field.

The remainder of this review is organized as follows. Section II sheds light on the main sensing modalities, technologies utilized, and the main features extracted from each modality. The manner in which these features are utilized, preprocessed, and fused is presented in Section III. Section IV briefly summarizes the applications of multimodality for diagnosis. Section V summarizes advanced intervention and therapeutic solutions toward real-time applications, with the help of robotics and VR-based environments. The challenges and future directions are discussed in Section VI, and this review is concluded in Section VII.

II. Sensor's Modality in Autism Research

In autism research, the categorization of signals used can broadly be divided into physiological and behavioral categories [101], each offering unique insights into the condition. A breakdown of these categories and their importance in autism research are given as follows.

A. *Physiological Signals*

Physiological signals, including eye gaze, heart rate (HR), skin conductance, and brain activity measured through methods [such as EEG and galvanic skin response (GSR)] give researchers objective, quantifiable data that reflect the internal states of individuals with autism in real time, highlighting their importance in evidence-based autism care. Physiological responses, originating from the autonomic nervous system (ANS), offer distinct benefits compared to other behavior indicators such as facial expressions, body gestures, and vocal cues. These responses are largely involuntary, accurately mirroring the participant's genuine condition [82]. This characteristic is especially valuable in autism research, where individuals with ASD frequently exhibit unconventional social behaviors, including atypical facial expressions and body movements [81].

1) Eye: Eye gaze is based on recording and analyzing where and how individuals direct their gaze, providing a window into their internal cognitive processes and emotional states. Individuals with autism often exhibit impairments in movement and gaze patterns, such as atypical fixation durations on social versus nonsocial stimuli [33], [44], [45], altered saccadic movements indicating differences in information processing [43], unusual pupil dilation responses to emotional content [48], [49], and unique scan paths that diverge from neurotypical patterns [19], [28]. These differences are crucial for understanding the distinct ways individuals with autism perceive and interact with the world around them. Tobii and Mirametrix S2 are cutting-edge ET technologies, combining precision and adaptability for diverse applications, from autism diagnostics to academic research [19], [31], [33], [43], [44], [49], [50]. Tobii's high-frequency infrared-based systems (up to 1200 Hz) [65] enable precise tracking of saccades and microsaccades, supported by dynamic calibration algorithms that ensure robustness against lighting variations, participant movement, and obstructions. Meanwhile, Mirametrix S2 [https://www.mirametrix.com/] utilizes stereo infrared cameras for binocular tracking, excelling in real-time gaze overlay and portability. Its lightweight design simplifies deployment in naturalistic environments, making it ideal for usability testing and applications requiring immediate feedback, albeit with lower sampling rates compared to Tobii. However, common eye trackers, such as the Tobii, require users to be at a close distance with limited head movement. This can be restrictive and not ideal for naturalistic observation settings [30]. Advancements in computer vision have significantly enhanced the accessibility of ET technologies. A common approach involves identifying the face and key landmarks, such as the eyes, nose, and mouth, which are essential for estimating head orientation and calculating yaw, pitch, and roll angles as illustrated in Fig. 1. The authors [37], [57] employ advanced systems combining RGB and Kinect depth data to achieve robust head pose estimation. In [57], the pose from orthography and scaling with iterations (POSITs) algorithm is

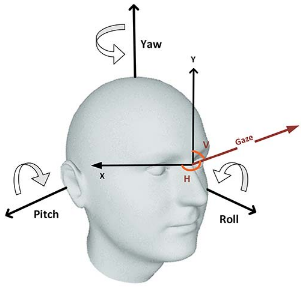

Fig. 1. Representations of gaze and head pose (Yaw refers to horizontal rotation (shaking head "no"), pitch to vertical rotation (nodding "yes"), and roll to lateral tilting (ear to shoulder)) - regenerated from [30].

utilized to fit a 3-D geometric model to 2-D facial landmarks, identified through boosted cascade detectors and supervised descent methods (SDMs). This approach enables precise estimation of yaw, pitch, and roll angles, maintaining accuracy even in the presence of occlusions or dynamic head movements. Similarly, the work done by [37] enhances horizontal orientation (yaw) estimation by integrating Kinect depth data with template matching. The depth data constructs a detailed 3-D head representation, while template matching refines alignment, resulting in improved accuracy for gaze direction tracking. Some other research like in [60] focuses exclusively on Kinect depth data, constructing a 3-D convex mesh of the participant's head to track head orientation in real time. By assuming gaze aligns with head orientation, this method simplifies the estimation process, prioritizing computational efficiency and real-time application but sacrificing finer accuracy in complex scenarios. In contrast, Cheng et al. [30] utilize a machine learning-based framework with convolutional neural networks (CNNs), such as ResNet-50, trained on labeled RGB datasets containing yaw, pitch, and roll annotations. The CNN extracts feature vectors from face images, and fully connected layers predict head pose angles. This joint learning framework integrates head pose into gaze estimation, achieving high precision and adaptability to varying conditions due to the model's data-driven nature.

2) Brain: There are several techniques for measuring brain activity, including brain-imaging methods that provide detailed visualizations of brain structure and function. However, the high costs of equipment and the need for specialized medical expertise have limited their application in human–computer interaction (HCI) research generally, and in autism research specifically [109]. In this survey, we focus on EEG signals, which have been widely utilized in ASD research due to their accessibility, affordability, and ability to capture neural dynamics in real time [19]. EEG is a technique that measures brain activity in the cerebral cortex using electrodes placed on the scalp. This process typically involves fitting participants with a cap containing 128 to 256 electrodes, which detect electrical signals across various regions of the brain. The activity is analyzed by examining differences in electrical signals between electrode sites or relative to an average baseline, providing insights into distinct patterns of brain activity [108]. Evoked responses, which are specific to auditory or visual stimuli, are captured by measuring the electrical differentials between electrodes placed at two locations, such as the scalp and earlobe [108]. EEG has proven instrumental in identifying atypical brain development, serving as a valuable early biomarker to differentiate children with autism from their neurotypical peers [66], [67]. High-density EEG systems, such as the 128 Ag/AgCl electrode array (Electrical Geodesics Inc., EGI), offer broad spatial coverage and detailed recordings with a sampling rate of 1000 Hz, further enhancing its utility in ASD research. This survey discusses the extensive use of EEG in autism studies, highlighting three primary neurophysiological features extracted from processed signals [19], [28], [43]. First, spectral features obtained through power spectral density (PSD) analysis reveal atypical activity across frequency bands (delta, theta, alpha, beta, and gamma), reflecting altered brain dynamics in autism [69], [70]. Second, multiscale entropy (MSE) quantifies the complexity of brain signals across temporal scales, uncovering unique neural complexity patterns in autistic individuals compared to neurotypical ones [24], [68]. Third, EEG data can be analyzed using graph theory to construct complex network models, providing insights into functional connectivity disruptions between brain regions—a hallmark of autism [22], [71].

3) Peripheral Physiology: Peripheral Physiological signals serve as a critical tool for decoding the intricate relationship between emotional and cognitive states, particularly in individuals with autism [84], [85]. With advancements in wearable and noninvasive sensor technologies, such as the E4 wristband [90] and BioNomadix systems, researchers can capture detailed real-time data on the ANS. These technologies facilitate the measurement of key metrics, including electrodermal activity (EDA), HRV, and blood volume pressure (BVP), which have been extensively linked to stress, emotional arousal, and cognitive load [86], [87], [88], [89]. EDA measures variations in skin conductivity resulting from changes in sweat gland activity, particularly in the hands and feet, triggered by emotional or cognitive stimuli. The salty composition of sweat enhances the skin's conductivity, which is monitored using electrodes typically placed on the fingers to track electrical flow between two points [108], [106]. Studies have shown that the intensity of EDA responses varies by emotion; for example, sadness often produces a greater increase in conductivity compared to fear. These distinctions make EDA a valuable tool for capturing nuanced emotional and cognitive experiences [107]. Cardiovascular signals add another layer of depth to physiological analysis. HRV, which quantifies the time intervals between consecutive heartbeats, provides insights into stress levels, cognitive effort, and emotional responses, such as fear, happiness, and anger [107], [84], [85]. Commonly used techniques for measuring cardiovascular activity include BVP monitoring and electrocardiography (EKG) [44], [47], [62]. BVP sensors, often placed on the fingers, measure fluctuations in blood flow by detecting changes in reflected light. These sensors offer an indirect but reliable method for assessing physiological arousal [109], as BVP is closely linked to the activity of the cardiovascular system. Additionally, HRV can be inferred from BVP data, providing a multifaceted understanding of heart activity [110]. EKG records the heart's electrical activity responsible for pumping blood.

Sensors placed on various points of the body measure HR, the intervals between beats, and variability, offering detailed insights into both physiological and emotional states [106].

B. Behavioral Signals

Behavioral signals in the context of autism research refer to observable actions or reactions that provide insights into an individual's emotional state, social interactions, cognitive processes, and adaptive behaviors. These signals include verbal such as vocal patterns, and nonverbal cues such as facial expressions and body pose.

1) Facial: Individuals with autism often exhibit atypical facial responses in social settings, indicating social communication and emotional expression difficulties. These responses, including delayed or exaggerated expressions, suggest emotional regulation challenges [74]. Recognizing these nuances helps researchers create advanced methods for early autism detection, especially in young children. Notably, specific patterns such as a reduction in social smiles during interactions have been identified as strong predictors of autism risk [75], emphasizing the importance of facial expression analysis in autism research. The utilization of EMG sensors to measure spontaneous facial expressions in individuals with autism exemplifies the technological advancements in this area. EMG sensors, capable of capturing subtle movements of facial muscles, represent a step forward in quantifying emotional states through physiological measurements. However, the technology's invasive nature and its focus on a limited set of facial muscles (typically cheek and eye-brow) [76], [77], [78] hinder the accurate representation of a subject's emotional state and may cause discomfort, especially in children with autism [33]. To address these limitations, nonintrusive video cameras coupled with computer vision methods have been proposed in some other research [30], [33], [37], [57], [59], [73]. Such methods minimize physical constraints and maximize subject engagement, enabling a more natural elicitation of facial responses across various contexts, including facial imitation robotic-based therapy sessions [60], or exposure to emotional video stimuli [73]. The computer vision methods for facial expression typically comprise three main components, including face landmark detection, multiface tracking, and Facial Action Unit extraction [30], [59]. Other research utilizes deep learning for nuanced analysis of facial expressions beyond manual feature extraction [61], [63]. This approach enhances traditional manual methods by automating the process to achieve a level of efficiency comparable to human observation or EMG measurements, as detailed in a recent review by [79].

2) Body Gestures: Motor abnormalities are becoming recognized as key indicators of ASD, affecting individuals across all ages and varying degrees of symptom severity [91]. These abnormalities, which include both gross movements, such as whole-body coordination [92] and fine movements such as dexterity [93], make those with ASD 22 times more prone to motor challenges compared to individuals with TD. The likelihood of these motor issues increases with the severity of repetitive and social behavior symptoms [94]. Kinematic studies show that movements in individuals with ASD are more variable and less organized than those in TD children. This implies that their movements may be characterized by repetitive or stereotyped behaviors, where certain motions are repeated excessively or in a simplified manner compared to the diverse and purposeful movements observed in TD children [95], [96]. Advances in technology, such as wearable sensors and 3-D motion capture [97], have significantly enhanced the objective measurement of these potential biomarkers, offering a window into the motor coordination and manual skills of individuals with ASD through comprehensive recording of whole-body actions using real-time computer vision algorithms. Gesture recognition in imitation tasks [57], [60] not only assesses the capacity for motor coordination and control but also reflects the ability to engage in social learning and interaction, which are often areas of difficulty for those on the spectrum. Similarly, the movement of body joints, such as the head, neck, and arms in [36], provides valuable information about the overall motor function and body awareness in individuals with ASD. Additionally, examining the speed and acceleration of these movements [29] sheds light on how individuals with ASD respond to different stimuli and navigate their physical and social environments, potentially indicating levels of arousal or anxiety. A deeper understanding of these motor abnormalities can aid clinicians in assessing ASD more accurately and pave the way for the development of automated, biomarker-based diagnostics, enhancing early detection and intervention strategies for ASD.

3) Vocal Patterns: Vocal features collectively offer a comprehensive toolkit for dissecting the complex vocal expressions associated with ASD. For instance, variations in pitch, frequency, and intensity can reveal subtle emotional nuances and social intentions [63] that might not be easily discernible through behavioral observation alone. Harmonicity and the nuanced analysis of MFCC provide a deeper understanding of speech quality and articulation, which are often unique in individuals with ASD [61]. The analysis of speech dynamics, such as interruptions in dialog transitions and the structure of sentences [44], illuminates the challenges individuals with ASD face in social interactions and conversational flow. Similarly, zero cross rate (ZCR) and spectral features [38] shed light on the texture and clarity of speech, which can influence how individuals with ASD are understood in social contexts. Prosodic and formant analysis, alongside energy metrics [37], delve into the rhythm and melodies of speech that carry significant emotional and communicative weight. Acoustic low-level descriptors [36] and the examination of speaking contents [30] further enrich the understanding of how individuals with ASD convey information and emotion, paving the way for tailored therapeutic interventions.

III. Multimodality Preprocessing and Integration

The analysis of multimodal data requires a comprehensive framework to ensure data consistency, accuracy, and relevance. This involves not only preparing the raw data for analysis but also leveraging advanced integration techniques to maximize the richness and dimensionality of the information collected.

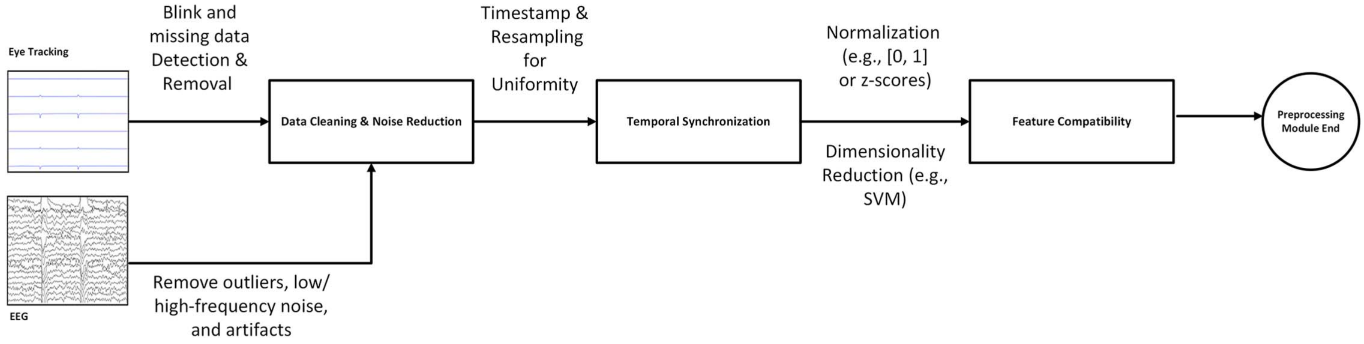
Eye Tracking
EEG
Blink and missing data Detection & Removal
Remove outliers, low/ high-frequency noise, and artifacts
Data Cleaning & Noise Reduction
Timestamp & Resampling for Uniformity
Temporal Synchronization
Normalization (e.g., [0, 1] or z-scores)
Dimensionality Reduction (e.g., SVM)
Feature Compatibility
Preprocessing Module End

Fig. 2. General methodology of preprocessing module for multimodality data.

To achieve this, preprocessing and sensor integration play pivotal roles, enabling efficient data preparation and fusion strategies tailored to application-specific requirements.

A. Preprocessing Module

The preprocessing module as illustrated in Fig. 2 (as an example of EEG and eye tracking data) in this section will cover the steps involved in preparing the multimodal data for analysis. This includes data cleaning and noise reduction techniques, methods for data synchronization to align data from different modalities, as well as normalization and feature reduction processes to ensure that the data are consistent, comparable, and reduced to essential features for efficient analysis.

1) Data Cleaning and Noise Reduction: ET data cleaning begins with a calibration process to ensure that the gaze measurements are accurately aligned with the visual stimuli. Calibration protocols, such as five-point and nine-point calibrations, are commonly used and repeated if necessary to ensure accuracy [19], [28], [33], [48]. Invalid data caused by factors such as occlusions, head movements, and tracking interruptions are excluded. Continuous tracking losses exceeding a threshold, such as 1000 ms, are removed [43], while shorter losses (typically under 100 ms) are handled using interpolation to maintain data continuity [19], [43]. Additionally, fixation thresholds are applied to exclude unintended or brief glances, ensuring that only meaningful fixations are considered in the analysis [28], [33], [48]. Advanced techniques, such as the isolation forest algorithm, are employed to identify and remove outlier data points resulting from sensor noise or extreme movements, and noise reduction methods, such as median filtering, are applied to further refine the data [30], [43]. These steps are crucial for ensuring the integrity and reliability of ET data, allowing for accurate analysis of visual attention and gaze behavior.

EEG data cleaning typically involves several key preprocessing steps to ensure the reliability and accuracy of the signals. Artifacts such as muscle movements, eye blinks, and power-line noise are removed using techniques such as independent component analysis (ICA), which decomposes the signal to isolate and exclude artifact-related components, followed by visual inspection to confirm the reliability of the cleaned data [28]. Power-line noise is further eliminated with a notch filter at 50 Hz, and bandpass filtering (0.5–45 Hz) is applied to retain relevant neural activity [19]. The data are then segmented into nonoverlapping epochs, typically 4 s in length, to facilitate analysis [19], [28]. Key electrodes are selected for analysis, such as eight specific electrodes (F3, F4, T3, C3, C4, T4, O1, O2) and 62 electrodes representing major brain regions [19], [28], and any channels with voltages exceeding ±200 μV are identified as bad and interpolated using neighboring electrodes to maintain data continuity [19]. Additionally, the data are downsampled to reduce computational complexity without losing essential information [19]. These steps collectively ensure clean and accurate EEG data for further analysis.

The general approach for cleaning physiological data typically involves several standardized steps to ensure the signals are reliable and suitable for analysis. First, outliers and artifacts are removed from the physiological signals, including ECG, PPG, skin temperature (SKT), and GSR signals, to eliminate extreme values that could distort the analysis. This step often includes filtering to remove high-frequency noise, baseline wandering, and other signal distortions [44], [43]. The signals are then smoothed to reduce short-term fluctuations and ensure consistency [48]. Baseline corrections are performed by subtracting the baseline mean from each signal to normalize for individual differences, followed by standardization to zero mean and unit variance to ensure comparability across participants [48]. For slowly changing signals, such as SKT, respiratory signals (RSP), and GSR, subsampling is often applied to reduce the computational load while maintaining the integrity of the data [43]. After cleaning, key physiological features such as HR, HRV, and skin conductance response rate are extracted, and the data are segmented into intervals (e.g., 1-minute intervals) to prepare the data for meaningful analysis [44], [48].

Facial data cleaning involves addressing issues such as posed faces, facial occlusions by hands, and partial faces that may occur during the recording process. Facial images with out-of-plane head rotation or occlusions are discarded to avoid introducing errors into the analysis. For faces with in-plane rotations, a landmark-based registration process is applied to adjust the alignment of the images, ensuring consistency in facial feature detection [33].

2) Data Synchronization: In multimodal data collection, various data sources are often recorded simultaneously but at slightly different rates, times, and space, or using distinct time-tracking methods. For example, ET data, video footage, and physiological measurements such as HR may each have different timestamps, reflecting when events were recorded relative to their respective internal clocks. To ensure these data align with one another and correspond to the same time events, synchronization is crucial. A widely used approach for synchronizing multimodal data is the timestamped method, which aligns the data by referencing the timestamps from each modality, bringing them to a common time base (e.g., [43], [30], [43], [44], and [47]). For example, in a system [43] where eye gaze data are recorded at 100 Hz (every 10 ms) and physiological data at 1 Hz (every second), the timestamps from both modalities must be mapped to a unified time base to ensure alignment of events. Since ET data are recorded more frequently, it needs to be interpolated to match the lower sampling frequency of the physiological data, or the physiological data must be adjusted to align with the higher frequency time intervals of the ET data. Interpolation or adjustment ensures timestamps from both modalities align, enabling meaningful analysis of eye movements and physiological responses.

Machine learning-based synchronization techniques aim to synchronize multimodal data by leveraging the power of machine learning models to discover patterns and relationships in the data rather than relying on traditional methods such as explicit temporal alignment or manual synchronization. For instance, synchronization in [28] employs a graph convolutional network (GCN) to analyze EEG and ET features without explicit temporal alignment. GCN captures the covariance relationships between features, enabling feature-based synchronization by modeling interrelationships through graph-based methods and aligning the data in a shared space for improved analysis and classification. Similarly, Han et al. [19] utilize a stacked denoising autoencoder (SDAE), a deep learning model that learns implicit synchronization by uncovering shared high-level representations of EEG and ET data in a latent feature space. Unlike the feature-based approach of GCN utilized in [28], the SDAE focuses on nonlinear temporal dependencies, compensating for asynchronous data and sampling rate differences, ensuring robust synchronization across modalities.

While temporal synchronization aligns events recorded at different timestamps, spatial/space synchronization such as eye centers, object locations, and skeleton data deal with merging data captured by various sensors or cameras, each with its own local spatial reference frame, into a unified global coordinate system. To efficiently combine this data from various sources, a coordinate transformation module is proposed in [30] and [57], which converts the local coordinates of each sensor into a unified global coordinate system. This process involves two key calibration steps: Kinect-camera calibration, where the relative poses of the Kinect and cameras are aligned through a joint calibration framework to ensure accurate data fusion; and Kinect–Kinect calibration, where the relative positions of two Kinects are calibrated using the iterative closest point (ICP) algorithm to optimize data alignment [30]. After calibration, the data from all sensors are mapped to a global coordinate system, with the world origin set at the base of the primary Kinect (e.g., with zero coordinate). The transformation is performed using a rotation matrix and translation vector using rigid transformations [30] and [57].

As an example, let $\mathbf{p_c} \in \mathbb{R}^3$ denote the head position within a specific camera's coordinate system. This position can be mapped to the world coordinate system using the following transformation:

$$\mathbf{p_w} = R \times \mathbf{p_c} + \mathbf{T}$$

where $R \in \mathbb{R}^{3\times3}$ represents the rotation matrix, and $\mathbf{T} \in \mathbb{R}^3$ is the translation vector, both describing the rigid transformation between the two coordinate systems. These values can be determined through the chessboard calibration technique [104].

Real-time synchronization of multimodal data is essential for applications where multiple data streams, such as gaze, speech, task input, and physiological signals, need to be processed simultaneously and interactively. In such systems, the data collected from various sensors and devices must be aligned in real time to ensure a seamless experience for users. This is particularly crucial in environments, such as VR and collaborative robotics, where interactions occur in dynamic and time-sensitive contexts. For instance, the work in [45] emphasizes the importance of the data synchronization channel to ensure seamless interaction between participants in real-time collaborative tasks. Specifically, it uses TCP for task-related data transmission, ensuring that data are delivered in the correct order, which helps maintain a synchronized virtual environment between participants. Additionally, WebRTC is utilized for real-time audio and video streaming, which ensures that participants' communications are synchronized with minimal delay. The Mirror plugin is responsible for virtual object synchronization within Unity, maintaining consistency across the shared virtual space. All of these components work together to handle the multimodal data, ensuring that the communication, task input, and virtual interactions are seamlessly aligned in real-time, which is crucial for effective collaboration and teamwork in the system.

3) Normalization and Feature Reduction: Normalization and feature reduction are essential techniques in handling multimodal data, as they ensure consistency, improve the efficiency of data processing, and enhance the interpretability of the combined features. In multimodal analysis, normalization typically involves transforming features into a common scale or format to account for differences in units or scales across the modalities. For instance, all features are normalized by [43] into the range [0, 1] to ensure compatibility between different data sources with varying units and scales. Similarly, Bekele et al. [48] standardize physiological data by computing z-scores, subtracting the individual baseline mean, and dividing by the standard deviation, to ensure comparability across physiological features. In [61], log-transformation and z-score normalization are applied to physiological signals, enhancing the comparability across different data sources and preparing them for fusion.

On the other hand, feature reduction is employed to manage the complexity and dimensionality of the data, ensuring that the analysis remains computationally feasible without sacrificing important information. Principal component analysis (PCA) is commonly used in these studies as a dimensionality reduction technique. For example, Zhang et al. [43] apply PCA during feature-level fusion to reduce the dimensionality of the

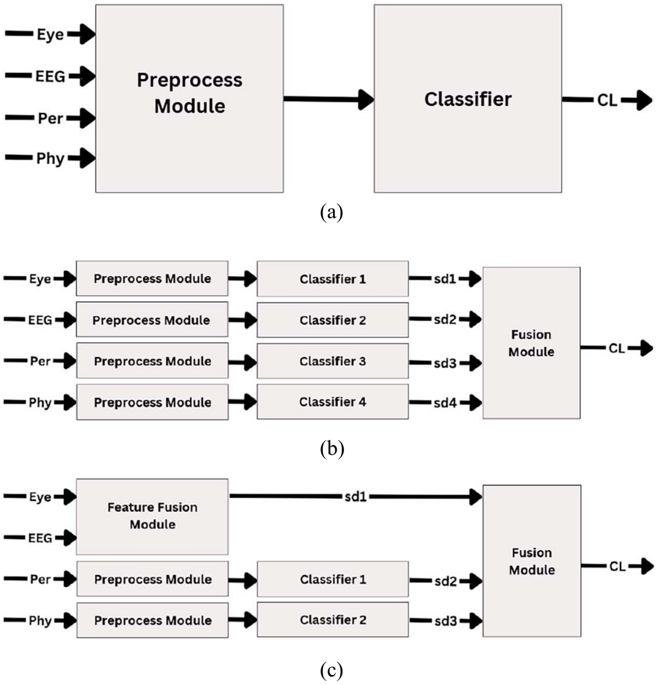

Fig. 3. (a) Early fusion framework. (b) Late-level fusion framework. (c) Hybrid fusion framework—regenerated from [43].

combined feature vector while retaining essential information. Bekele et al. [48] use PCA to reduce the dimensionality of 16 physiological features, selecting the first two components that capture over 80% of the variance. PCA is also applied by [59] to high-dimensional feature vectors derived from action unit intensities and appearance-based features, reducing the feature size from 4464 to 1391 dimensions. Additionally, adaptive thresholding in [59] further reduces data complexity by segmenting meaningful intervals in time-series data, preserving the key behavioral patterns while reducing redundancy.

B. Fusion Approaches

Integrating various sensors into a cohesive system enhances the dimensionality and completeness of the data captured, enabling a more accurate representation of objects within the data stream. Sensors contribute diverse data types, which are characterized by variations in sampling rates, input quantities, and the underlying information content. The selection of a multisensory fusion approach is guided by the specific requirements of the application, including the types of sensors used and the structure of the data involved. In the analysis of multimodal data, especially when dealing with complex biological signals lacking straightforward intermodal correlations, three principal fusion strategies emerge in the context of autism research (early, late, and hybrid, as illustrated in Fig. 3).

1) Early Fusion: Early fusion (also known as feature-level fusion) combines early-stage extracted features from multiple modalities to a single vector, enabling the identification of intricate patterns across different data sources. This approach requires the assumption that features from various modalities are directly comparable, a condition not always met due to disparities in data characteristics. An example of this is the assessment of learning abilities in children with autism, where feature-level fusion integrates data from facial expressions, body movements, and verbal communications along with emotional states into an elaborate feature vector [62]. This fusion method underscores the linkage between internal mental states and external behaviors in autism, providing a layered understanding of the children's learning capabilities.

The studies using early fusion benefit from several key advantages. First, they use tailored initial modality-specific layers, where each modality's unique feature is processed independently, preserving modality-specific characteristics. They also possibly find correlations between modalities, as seen in [28] and [19], where the fusion captures complementary relationships between EEG and ET data, enhancing model accuracy. Additionally, early fusion can work with different feature structures, as demonstrated in these studies [30] and [36], where various modalities such as gaze, gestures, and speech are integrated effectively, preserving their distinct features while allowing for meaningful fusion.

While the benefits of early fusion are clear, they come with certain challenges. In [28], the fusion of modalities with very different structures (e.g., EEG and ET) requires significant preprocessing to align these data types, which can complicate the integration process. Similarly, Han et al. [19] face the challenge of advanced neural network engineering, as the complex SDAE architecture necessitates careful design and sufficient data to optimize performance. Furthermore, Cheng et al. [30] highlight the difficulty of processing heterogeneous modalities within a unified framework, particularly when those modalities have different feature structures that must be harmonized before fusion.

The premise behind using early fusion is that it works best when there is a high likelihood of important modality interactions, as seen in the studies by [19], [28], and [30]. These interactions—such as the relationship between neural activity and ET behavior, or between vocal and gestural cues—become more meaningful when processed together at an early stage, allowing the model to leverage the full complement of data to improve classification accuracy. Additionally, the use of early fusion in these studies assumes that the data can be preprocessed or aligned to ensure compatibility, either through synchronization or adjusting sampling rates as discussed earlier in Section III-A. The approach is particularly advantageous when sufficient data are available to train complex models, as evidenced in [36], where early fusion is used to effectively combine speech and motion features after independent processing.

2) Late Fusion: Late fusion or decision level derives separate decisions from each modality and then aggregates these to reach a final conclusion, thus leveraging the independent predictive power of each data stream. An illustrative use case is in emotion recognition tasks that employ deep learning models to separately analyze audio and visual data, later fusing these insights at the decision level [61].

The study in [37] employs late fusion due to its suitability for integrating heterogeneous data structures and preserving modality-specific features. Unlike early fusion, which requires extensive preprocessing, late fusion processes video, audio, and physiological signals independently, ensuring that critical behavioral markers such as gaze direction, vocal turn-taking, and EDA are retained without forced alignment. Additionally, late fusion is robust to missing data, as predictions from available modalities can still contribute to the final decision, ensuring reliable analysis despite occasional sensor failures (e.g., physiological signals) or incomplete data. Compared to early fusion, which demands complex neural network architectures and large datasets to model feature interactions, late fusion is simpler and more practical for tasks where intermodality dependencies are minimal [83]. For instance, in engagement prediction, gaze, and speech features independently provide complementary insights without requiring intricate interaction modeling. However, a key disadvantage of late fusion is the difficulty in finding a suitable decision rule or algorithm that can effectively combine the outputs from different modalities [37], [61], [62]. The fusion decision relies heavily on the quality of the insights used to integrate data, which requires careful tuning and expertise [62]. Furthermore, while late fusion works best in cases with low intermodality interaction, its effectiveness diminishes when there is a strong correlation or interaction between modalities. This is particularly important when analyzing social interactions, where modalities such as gaze, facial expressions, and vocal behavior are deeply interconnected. Despite these challenges, the flexibility of late fusion, especially when enough data are available to train the decision model, makes it an ideal choice for multimodal analysis in dynamic, complex environments such as the study of child–adult interactions [61].

In the papers [37], [61], [62], the decision fusion process is implemented by combining the outputs of multiple classifiers using a weighted average method. Each classifier produces a binary decision indicating whether the cognitive load is low or high, which is stored in a subdecision vector $D = (d_1, d_2, d_3, d_4)$, with each element representing a decision from a different modality, such as eye gaze, EEG, or physiological data. A weight vector $W = (w_1, w_2, w_3, w_4)$ is assigned to each subdecision, where the weights are within the range [0, 1] and their sum equals 1, determining the importance of each modality in the final classification. The weighted average y is calculated by summing the products of each subdecision and its corresponding weight as follows:

$$y = W\mathbf{D}^T = \sum_{i=1}^{4} w_i d_i.$$

The final cognitive load classification is determined by applying a threshold to the weighted sum, where a value of y less than 0.5 indicates low cognitive load (0) and a value greater than or equal to 0.5 indicates high cognitive load (1). The optimal weight vector is determined by maximizing classification accuracy, ensuring that the most relevant information from each modality is considered in the final decision

$$\mathbf{W}_{\text{optimal}} = \arg\max_{\mathbf{W}} \left(\text{Accuracy}(\mathbf{W})\right).$$

Finding the optimal weight vector is typically found using an exhaustive search. In this method, all possible combinations of weights are tested to determine which one gives the highest accuracy. This process is usually computationally expensive, especially when the number of classifiers or modalities (and hence the number of weights) is large. To this end, Chen et al. [62] propose a technique to reduce the computational burden of exhaustive search by narrowing the search space. The authors demonstrate that by partitioning the weight vector set into subsets based on the weights' characteristics, a smaller number of weight vectors can be tested while still identifying the optimal one. To determine the optimal weight vector, a weight vector is randomly selected from each of the identified subsets, and the accuracy of decision-level fusion is computed and compared across these selections. The weight vector yielding the highest accuracy is identified as the optimal one, which leads to the same optimal decision as considering all possible vectors. In our survey, we have summarized the optimize weight vectors Algorithm 1 for decision-level fusion, as proposed in [62], and presented it as a pseudocode for better clarity and accessibility.

3) Hybrid Fusion: Hybrid fusion combines the strengths of both early fusion and late fusion, making it particularly valuable for integrating multimodal data. This approach allows for the preservation of modality-specific features through early fusion, where data from each modality are processed independently while also benefiting from the robust decision-making capabilities of late fusion, where the outputs of each modality are combined at the decision level. The main advantage of hybrid fusion is that it can capture both the individual and interactive effects of modalities, enabling a more comprehensive understanding of complex data. For instance, the research in [30] adopts a hybrid approach, combining feature-level and decision-level fusion strategies to optimize the analysis of human behaviors from synchronized multiview camera data. This demonstrates the potential of using various fusion techniques to navigate the complexities of multimodal and multiview data analysis. Table I presents three general fusion approaches, along with their respective advantages, challenges, and preconditions.

IV. Multimodality Sensing for Diagnosis

A. Computer-Based Tasks

One of the primary objectives of diagnostic models in computer-aided diagnostic (CAD) systems for autism is to enhance sensitivity (the ability to correctly identify individuals with autism) and specificity (the ability to correctly identify individuals without autism) [19], [28]. This, in turn, improves the overall accuracy and reliability of the classification process. Incorporating data from multiple modalities into these systems can provide complementary features, resulting in a richer dataset for machine learning models or statistical analyses.

EEG and ET modalities provide complementary strengths that enhance diagnostic accuracy for ASD when integrated. EEG captures internal neurophysiological activity, including neural oscillations, connectivity, and signal complexity, offering insights into cognitive processes and irregularities often associated with ASD. In contrast, ET focuses on external behavioral metrics, such as gaze patterns and attention allocation, providing direct observations of how individuals interact with

Algorithm 1: Optimize Weight Vectors for Decision-Level Fusion as Proposed in [62].

```
Require: Universal set of weight vectors $U = \{(w_1, w_2, w_3, w_4) \mid w_1 + w_2 + w_3 + w_4 = 1, \quad 0 \leq w_i \leq 1\}$
Require: Training data $(X, Y)$
Ensure: Optimal weight vector $w^*$
Step 1: Partition Universal Set $U$
Define $w_{\text{max}} = \max(w_1, w_2, w_3, w_4)$ and $w_{\text{min}} = \min(w_1, w_2, w_3, w_4)$
Partition $U$ into:
    $O = \{w \in U \mid w_{\text{max}} > 0.5\}$
    $P = \{w \in U \mid w_{\text{max}} = 0.5\}$
    $Q = \{w \in U \mid w_{\text{max}} < 0.5\}$
Exclude subset $P$ (boundary condition $y = 0.5$)
Step 2: Refine Subset $Q$
Partition $Q$ into:
    $Q_A = \{w \in Q \mid w_{\text{max}} + w_{\text{min}} > 0.5\}$
    $Q_B = \{w \in Q \mid w_{\text{max}} + w_{\text{min}} < 0.5\}$
    $Q_C = \{w \in Q \mid w_{\text{max}} + w_{\text{min}} = 0.5\}$
Exclude subset $Q_C$ due to low accuracy
Step 3: Further Partitioning
for $k \in \{1, 2, 3, 4\}$ do
    Partition $Q_A$ based on the index of $w_{\text{max}}$:
        $Q_A^k = \{w \in Q_A \mid w_k = w_{\text{max}}\}$
    Partition $Q_B$ based on the index of $w_{\text{min}}$:
        $Q_B^k = \{w \in Q_B \mid w_k = w_{\text{min}}\}$
end for
Step 4: Compute Decision Rules subset $Q_A^k$ or $Q_B^k$
Compute final decision $y$:
    $y = \sum_{i=1}^{4} w_i d_i$
if $y \geq w_{\text{max}} + w_{\text{min}} > 0.5$ then
    Assign $d_{\text{final}} = 1$ {High confidence decision}
else if $y = w_i < 0.5$ then
    Assign $d_{\text{final}} = 0$ {Low confidence decision}
else
    Assign $d_{\text{final}} = 1 - (w_i + w_{\text{min}})$ {Intermediate confidence}
end if

Step 5: Optimize Decision-Level Fusion
Select one representative weight vector $w_k$ from each subset
Compute accuracy for each selected weight vector:
    $A(w_k) = \frac{\sum_{j=1}^{m} \nVdash(D(w_k, x_j) = y_j)}{m}$
Identify the optimal weight vector:
    $w^* = \arg\max_{w_k} A(w_k)$
Return $w^*$ =0
```

their environment. The SDAE framework in [19] fuses these two modalities to utilize their complementary strengths. EEG excels in classifying ASD children by revealing atypical brain activity, while ET is more effective in distinguishing TD children due to its behavioral specificity. The fusion approach in this study captures the intermodality relationships, creating a robust feature representation that enhances classification performance compared to unimodal methods; thus, increasing in sensitivity and specificity of the final fused analysis.

The research outlined in [28] explores the functional connection between EEG and ET to examine both intramodality and intermodality information within and across these two modalities. Findings indicate that while EEG features do not all correlate with one another, ET features tend to show mutual correlations. For intramodality correlations, the study highlights robust connections between these two bio-signals, even when gathered asynchronously, suggesting their potential to be integrated in a complementary fashion in autistic diagnostic systems. For example, permutation entropy (PermEn) correlates with Joint Attention, linking neural complexity to challenges in shared focus, while wavelet entropy (WaveEn) correlates with Social Interaction, illustrating how chaotic brain activity impacts engagement. This fusion of modalities provides a comprehensive understanding of ASD, improving diagnostic accuracy by addressing both the internal neural processes and external behavioral dimensions of the disorder.

The correlation analysis method was also used in [29] to establish the relationship between patients' etiology and brain function changes. This method promotes data exchange between doctors and patients by monitoring both the patients' skeletal movements and brain activity. The multimodal quantitative output results serve as auxiliary diagnostic standards. Multimodality fusion approaches proposed in [28] and [19] significantly boost accuracy compared to relying on a single-modality approach. Additionally, both studies highlight that ET, in particular, outperforms the EEG modality in distinguishing between autistic children and those with typical development. Other research tends to compare the diagnostic scores from the multimodality systems with human evaluation outcomes as in [30]. This study shows that using multimodal data (including speech, facial, joint attention, and hand gestures) with different paradigms in an unconstrained testing studio provides diverse reference points, reinforcing assessments and minimizing single-source biases, thus improving the consistency with human scoring. Given the crucial role of nonverbal expressions in social communication, accounting for about 80% of the cues [32], researchers are investigating their potential to differentiate between autism and TD individuals. The research in [33], simultaneous monitoring of facial expressions, eye contact, and hand movements during tasks involving visual stimuli. Findings revealed that participants exhibited uncontrolled smiling without appropriate visual engagement, alongside a minimal correlation between eye and hand movements. This suggests challenges in social communication and motor coordination among participants. Such insights emphasize the value of simultaneously capturing and analyzing multimodal responses to quantitatively assess autism, aiding in the early identification of symptoms and the development of targeted interventions. Studies on multimodality reveal that individuals with autism face challenges in multisensory temporal processing (MTP), affecting their ability to synchronize sensory events from different modalities, such as sound and vision. Research highlighted in [34] using audio-visual tasks showed that those with autism

TABLE I
SUMMARY OF FUSION APPROACHES: ADVANTAGES, CHALLENGES, AND PRECONDITIONS

Approach	Advantages (+)	Challenges (-)	Preconditions (*)
Early Fusion	• Facilitates the discovery of correlations between modalities • Allows modality-specific processing through tailored initial layers • Accommodates diverse feature structures from different modalities • Enables the discovery of inter-modality relationships	• Requires similar feature structures across modalities • Difficult to apply with heterogeneous data structures • Demands advanced neural network design and sufficient training data • Individual preprocessing and engineering for each modality can be challenging	• Data must have synchronized sampling rates or be preprocessed for alignment • Best suited for scenarios where strong interactions between modalities are expected • Requires enough data to support complex neural network structures • Effective when capturing meaningful interactions between modalities is essential
Late Fusion	• Supports independent feature processing with modality-specific designs • Robust to missing or incomplete data, as outputs from available modalities can still contribute	• Requires the development of suitable decision rules or combination algorithms • Decision-making may require domain knowledge or substantial data to train the fusion mechanism	• Best applied when interactions between modalities are minimal or when outputs can be combined easily
Hybrid Fusion	• Combines strengths of early and late fusion approaches • Captures both modality-specific and inter-modality interactions • Increases flexibility and accuracy	• Requires careful integration of early and late fusion stages • Can be computationally intensive • Complex decision rules may be needed	• Requires appropriate data preprocessing for both early and late fusion stages • Sufficient data to train both feature-level and decision-level components • May require advanced computational resources for handling complex interactions between modalities

struggle with integrating sensory information over time yet still manage some level of nonverbal synchrony. This suggests other factors might mitigate MTP difficulties, highlighting the complex relationship between sensory processing and social interactions in autism.

While extensive research has focused on differentiating autism from the control group, relatively few studies have explored the differentiation between autism subgroups: autistic disorder (AD), high-functioning autism (HFA), and Asperger's (AS). This shift emphasized autism as a spectrum, recognizing the wide range of symptoms and severity that individuals may exhibit. A recent computational study by [35] has demonstrated the potential of extracting low-level motion and vocal behavior descriptors from both the participant and the investigator during the ADOS-emotion section to differentiate between three subgroups of autism. Another Study by [36] incorporates speech and motion subnetworks into a fusion network for classifying these three subgroups. The results from this study show that anger and fear are key interaction segments for observing subtle behavioral differences between these subgroups. Participants' vocal behavior is more discriminative than investigators' for speech, while investigators' gestural behaviors provide more discriminative information than participants' motion behaviors. This could have important implications for the diagnosis and treatment of autism, as different subgroups may respond differently to different interventions. Research in [37] introduces a new multimodal dyadic behavior (MMDB) dataset containing video, audio, and physiological recordings of 160 play sessions involving 121 infants and toddlers between 15 and 30 months interacting with adults. These brief (3–5 min) sessions offer valuable insights into children's social-communicative development through detailed annotations and automated analysis of behavior (gaze shifts, smiling, and play gestures), and engagement ratings. These behaviors reflect key socio-communicative milestones in the first two years of life, and their diminished occurrence and qualitative difference in expression have been found to represent early markers of ASD.

V. MULTIMODALITY SENSING FOR INTERVENTION

In this section, we explore the practical applications of multisensing modalities, specifically focusing on VR and robotic-based systems. In VR system, our exploration spans two key areas: 1) personalized and adaptive training; and 2) quantitative skills assessment. We uncover how VR technologies can tailor training experiences to individual needs and objectively measure skill levels. Next, within robot-based systems, we highlight their crucial role in providing cognitive and affective interventions within therapeutic solutions. This investigation underscores the valuable integration of advanced technologies in real-world scenarios, showcasing their potential to significantly enhance training and assessment approaches.

A. *VR for Autism Skills Training and Assessments*

People with autism often have strong interests in specific topics or activities [7]. To make learning more interesting, therapists can use VR with gamelike features that can harness their interests. In the context of autism, VR provides a safe space to practice social skills without worrying about mistakes in real-life interactions [41], [48]. It is also flexible and scalable, allowing therapists to personalize interventions based on individual abilities in cognition, language, and social skills, promoting generalization across different contexts [42]. In traditional VR, users primarily engage through visual input, immersing themselves in 3-D environments. However, this approach has limitations in replicating real-world interactions. Advancements in VR incorporate multiple modalities beyond visuals, enhancing user interaction. This is particularly beneficial for individuals with autism, offering a versatile platform for personalized social skills development, communication practice, and inclusive learning experiences. The research in multimodality within the VR environment serves two main goals. First, it aims to develop personalized and adaptive training methods for closed-loop applications, highlighting the essential role of adaptive feedback. Second, the focus is on leveraging multimodal data for quantitative skills assessment.

Real-time data handling is crucial for applications targeting complex skills measurements (e.g., driving [43] and conversation management [44]) where immediate feedback or intervention might be required. In [43], a VR-based driving simulator was developed for adolescents with ASD to personalize task difficulty based on real-time cognitive load estimation. Built in Unity3D, the system adjusted parameters such as vehicle speed, steering, and lighting across six difficulty levels. It collected eye gaze (Tobii X120), EEG (Emotiv EPOC), peripheral physiological signals (ECG, GSR, PPG, etc.), and performance data. Cognitive load was classified using multiple machine learning models (SVM, ANN, and LDA) and fused across feature-, decision-, and hybrid-levels, with ground truth provided by trained raters. Results showed that multimodal fusion improved accuracy, enabling adaptive VR training tailored to individual cognitive states. The career interview readiness in virtual reality (CIRVR) system, as described in [44], is an adaptive, closed-loop VR training platform designed to support individuals with ASD through immersive, structured mock job interviews. CIRVR integrates Microsoft Azure's emotion recognition and face tracking APIs to monitor facial expressions, eye gaze, and stress levels in real time. A Conversation Management System dynamically adjusts the interviewer's behavior based on these live inputs, enabling personalized and responsive interaction. Key behavioral indicators such as gaze aversion, long response latency, and stress-related behaviors are automatically flagged and visualized through a dashboard interface for job coaches, facilitating tailored feedback and intervention planning. A pilot with nine participants showed CIRVR's ability to identify individual communication challenges and deliver customized interview readiness support, highlighting its technical feasibility and user acceptability.

In a distinct research domain, multimodality data were leveraged for a comprehensive assessment of qualitative skills from various perspectives. A notable example is found in [45], where the utilization of multimodal data enabled the provision of quantitative measures for nine dimensions of collaboration [46]. This analysis involved two distinct groups, ASD-NT and NT-NT, and captured eye gaze, speech, and hand movements. The system quantified specific collaboration dimensions, such as technical coordination and the type/frequency of collaborative actions, resulting in higher accuracy compared to uni-modal analysis. Notably, both groups exhibited similar collaboration patterns, indicating that the tasks adhered to universal design principles. Another study developed CheerBrush [47], a specialized coaching system to improve tooth-brushing skills in children with autism. Using multimodal data, the system gauges engagement through ROI attention, monitors stress levels with HR and skin conductance level (SCL), and automatically assesses performance based on brushing speed and position. Both autism and TD children exhibited enhanced brushing skills postintervention, highlighting the system's positive impact. Notably, physiological features such as HR and SCL showed near-statistically significant differences between baseline and training sessions for autism.

A study by [48] introduced a VR-based system for facial emotion recognition to assess how adolescents with ASD respond to animated expressions. Using unity-based scenes, participants were shown Ekman's seven emotions while eye gaze (via Tobii X120) and physiological signals (e.g., ECG, GSR, and temperature) were recorded. Metrics such as fixation duration, pupil dilation, and skin conductance were analyzed across facial ROIs (e.g., eyes and mouth). The ASD group showed reduced emotion recognition accuracy, particularly for negative emotions, along with shorter gaze durations and altered physiological responses. These findings highlight atypical emotional processing in ASD and suggest implications for targeted social interventions.

Researchers in [49] developed a multimodal adaptive social interaction VR environment (MASI-VR) that adapts to users' gaze patterns to provide individualized training for recognizing emotions. This system collected eye tracking, physiological, and EEG data while users completed tasks. While physiological and EEG signals are collected for offline analysis, eye tracking data were used to dynamically conceal facial features, revealing them only when users focused on relevant areas such as the eyes and mouth. This online feedback mechanism helped users improve their performance by 3% in recognizing emotions, closing the gap between them and a control group not receiving this adaptive training.

While most VR research has centered on interventions and learning in autism, the technology's immersive nature opens a new avenue for investigating atypical sensory processing. Study in [50] developed SAVR, a pilot VR system featuring a game environment for presenting visual and tactile stimuli, manipulated via a haptic robot. By recording eye gaze and performance within the game, SAVR aimed to capture detailed visual and touch processing patterns. A pilot experiment with six children with autism and six TD peers demonstrated SAVR's feasibility, revealing notable sensory differences between the groups and exhibiting strong correlations with traditional sensory processing assessments, suggesting promising new avenues for objective and sensitive sensory processing evaluation in autism.

B. Robot System for Autism Therapy Solution

Robot-assisted therapy (RAT) stands out as a potential solution for enhancing the social skills of children with autism as these children tend to prefer interactions with nonhuman entities [52], [53]. In contrast to humans, robots offer greater predictability in behavior and ease of engagement, serving as intermediaries for human–human interaction [54]. This predictability aligns with the comfort and stability that autistic individuals find in routine, contributing to emotional stability during interactions with robots [55].

In standard RAT approaches such as the Wizard of Oz (WoZ) model [56], a human operator controls the robot from behind the scenes. Requiring an extra operator not only increases intervention costs and complexity but also complicates understanding children's behavior because crucial cues such as facial expressions are often not visible to the operator. Additionally, extra efforts are needed for postintervention performance analysis.

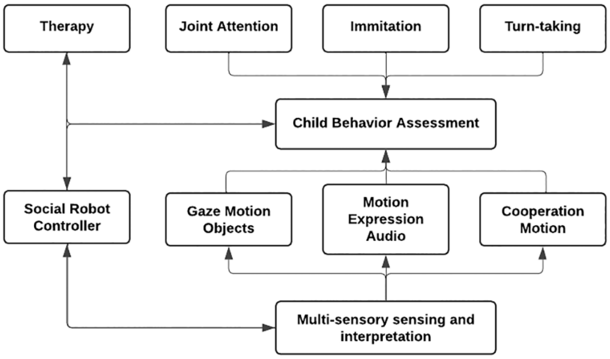

Fig. 4. General framework of the SET system—regenerated from [57].

To address these issues, Cai et al. [57] propose a new sensing-enhanced therapy (SET) system. This system uses multiple sensors to collect data about the child's behavior, which can then be analyzed to determine how the robot should respond. This closed-loop approach allows the robot to operate more autonomously, only requiring human intervention when necessary. This not only reduces the workload for therapists but also provides them with more data to improve their understanding of the child's behavior. The SET system prioritizes three key social interventions—imitation, joint attention, and turn-taking—which are fundamental in therapeutic approaches and commonly observed in children with autism [58]. In these interventions, children's behaviors are analyzed into various components such as gaze, expression, and motion using the multisensory sensing and interpretation module. The child behavior assessment module utilizes the outputs from each component to offer therapists valuable analyzed information on the behavior of children with autism, aiding in diagnosis, care, and treatment. Fig. 4 illustrates the general framework of the SET system as proposed in [57].

A system called robot-mediated imitation skill training architecture (RISTA) in [60] is designed to help children with autism learn new skills through imitation. The system includes a gesture recognition algorithm that can assess imitated gestures by tracking the person's skeleton and head pose in real-time and providing dynamic feedback. The results show that RISTA outperformed the human therapist in terms of both capturing the children's attention and imparting gesture learning. Study in [59] proposes a personalized detection of effective states in people with autism based on their facial expressions, eye movements, and head position. The system was able to accurately identify happiness-related behaviors in children with autism who were interacting with a humanoid robot.

Autism children's therapy systems pay little importance to their emotional perception and expression ability, and the timeliness and mobility of these systems are insufficient. To address these challenges, researchers have developed an innovative AI-based first-view paradigm [61], [62], a therapy system that utilizes wearable robots to provide autistic children with the ability to perceive and express emotions in their surroundings through a shared perspective with robots [51]. The system can also assist children in identifying and expressing their own emotions based on the facial and voice data of others while monitoring physiological signals to identify internal emotions. This real-time and long-distance emotional support system can enhance social interaction and emotional regulation and provide remote support, offering a promising new approach to autism treatment [61], [62]. While existing robotic-based studies have demonstrated the potential of personalization, their limitations in single-session experiments hinder long-term generalization. To address this, a long-term study using personalized models with a socially assistive robotics (SAR) tutor was proposed in [63]. Children with autism interacted with the SAR tutor over multiple sessions while playing educational math games. Based on each child's effective and cognitive performance, the SAR tutor provided both verbal and expressive feedback to promote the child's social and math skill development.

VI. Challenges and Directions

The exploration of multimodality advancement technologies in this work highlights a promising area in the diagnosis and assessment of ASD. These technologies, as structured in Fig. 5, delineate a future where nuanced diagnostics and personalized therapies are increasingly accessible through innovative platforms. The taxonomy classifies the current technological approaches into three main categories: computer-based, VR-based, and robotics-based technologies, each tailored to address specific aspects of ASD. Computer-based technologies enhance diagnostic precision by leveraging multimodal data to analyze cross symptoms, thus paving the way for early and accurate identification of ASD. VR-based technologies, on the other hand, transform therapeutic methods by providing immersive environments that enable real time, adaptive training and assessments. This not only aids in developing necessary life skills but also ensures that these interventions are engaging and tailored to individual needs. Robotics-based technologies introduce a tactile dimension to therapy, emphasizing interaction and engagement that are crucial for developing social skills and emotional responsiveness.

The convergence of these technologies marks a significant transformation toward a holistic approach in ASD care, creating a more integrated and interactive framework adaptable to individual needs and progressions. However, despite these advancements, substantial challenges remain, necessitating ongoing improvements and further exploration, as detailed in the following sections and in tables in Appendix A.

A. Challenges

Based on our previous discussions and the body of the literature, there are four main challenges related to diagnosis and intervention based on multimodality in the context of autism.

1) General Data Collection Challenges: Data collection in multimodal sensing faces significant challenges, including the limitations of small number of participants (small sample sizes) and the reliance on single-visit studies, which restrict generalizability [28], [29], [45], [47]. Discomfort with sensor-based technology and participant hyperactivity further exacerbate data loss [33], [37], [43]. For example, Zhang et al. [43] document

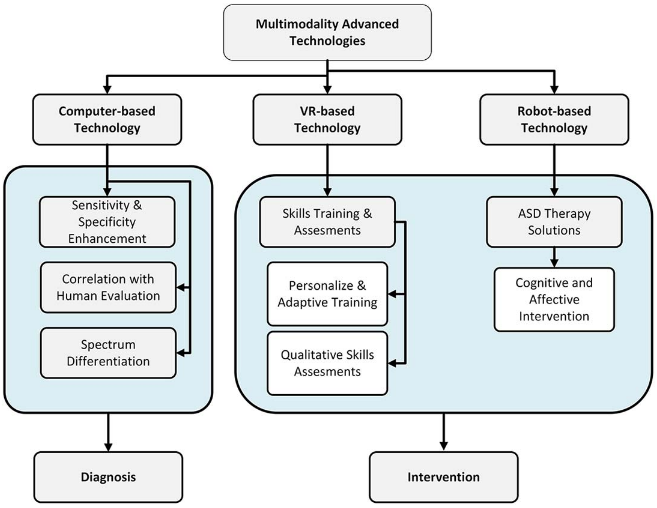
Multimodality Advanced Technologies
Computer-based Technology
VR-based Technology
Robot-based Technology
Sensitivity & Specificity Enhancement
Correlation with Human Evaluation
Spectrum Differentiation
Skills Training & Assesments
Personalize & Adaptive Training
Qualitative Skills Assesments
ASD Therapy Solutions
Cognitive and Affective Intervention
Diagnosis
Intervention

Fig. 5. General multi-modality advancement technologies.

a notable data reduction (20.56%) due to participant movement in VR driving tasks. Similarly, issues of discomfort and stress led to the withdrawal of participants in [60], affecting study completion rates. Despite some studies, such as [29], asserting that constraints posed by small sample sizes do not necessarily compromise the verification of system research methodologies' feasibility, the lack of large-scale data support poses potential risks to the durability and reliability of feedback mechanisms in longitudinal perspectives. Technical limitations in multimodal systems further impact the reliability of data collection and analysis. For instance, the ET systems face additional issues with validation, leading to reduced precision, especially when detecting gaze on small objects [45]. Furthermore, hardware constraints, such as the limited range and sensitivity of devices such as Kinect, impose strict requirements on participant positioning, such as maintaining a distance of 1.5 m for optimal performance [60], [47].

2) Synchronization Challenges in Multimodal Sensor Integration: A key challenge in sensor multimodality for autism interventions is achieving synchronization across diverse sensors such as EEG, eye tracking, and facial expressions. Asynchronous data collection with varying intervals and resolutions complicates integration, hindering real-time analysis and reducing the effectiveness of multimodal systems [19], [28]. The integration of data from various modalities—ranging from physiological sensors to environmental monitors—necessitates precise temporal alignment (as discussed in Section III-A2) to ensure that the data streams are accurately correlated. This synchronization is crucial for developing a cohesive understanding of the multimodal inputs and for the subsequent analysis to be meaningful and actionable. However, differences in sensor sampling rates, delays in data transmission, and processing speeds can complicate this alignment [64], potentially leading to inaccuracies in data interpretation and decision-making. Additionally, fusion of sensory data presents significant challenges, particularly when determining the most appropriate method to use, as highlighted in Table I. These challenges are especially pronounced in late fusion, where selecting an effective method for combining outputs from various modalities is crucial for achieving accurate and meaningful integration. For example, while most studies rely on weighted averages for data fusion, there is a pressing need for more advanced solutions to improve the flexibility and accuracy of the fusion process. Overcoming these synchronization and fusion challenges is essential for leveraging the full potential of multimodal sensor technologies in providing comprehensive and effective interventions.

3) Standardizing Assessment Paradigms: Multimodality-based assessment paradigms represent a significant advancement over single-modality approaches, demonstrating enhanced efficacy in detecting autism. However, standardizing these assessment paradigms across the diverse spectrum of autism remains a challenge, primarily due to the complexities involved in distinguishing between genuine skill deficits and atypical but nonpathological behaviors. Predefined difficulty levels often fail to account for the various factors affecting task difficulty, further complicating the creation of consistent assessment methods [43]. The False Diagnostics Dilemma, a significant problem in diagnostic practices, arises from inaccuracies in identifying conditions correctly. Specifically for autism, research [30] has found that a child's age and any early interventions they've received can skew diagnosis results. False-negative cases in their final model arise for those children with mild symptoms or who got prior interventions, while false positives occur in younger children due to developmental variations that mimic autism signs. Simulated social interactions in many assessments remain very limited, reducing the ecological validity of the paradigms and their ability to capture real-world social behavior [48]. Additionally, the unique characteristics of ASD profiles—including wide variations in behavior, sensory responses, and emotional expressions—further complicate the standardization of assessment paradigms [63]. These individual differences introduce high variance and noise into datasets, particularly when assessments are conducted in unconstrained, in-home settings. There is also significant difficulty in identifying which types of multimodal data are most effective for autism analysis [62], making it harder to create universal diagnostic standards. This highlights the need for personalized approaches that account for these complexities. This underscores the complexity of diagnosing autism accurately and the importance of refining diagnostic and intervention approaches.

4) Real-Time Multimodal Data Processing and Adaptive Interaction Challenges: Real-time data analysis poses significant challenges for effective interaction with ASD individuals. These challenges stem from the need to continuously process and interpret a diverse range of data streams in real time, including visual cues, speech patterns, and physiological responses. Systems like the one in [48] relied on static performance, lacking real-time adjustments or physiological integration, which limits their ability to adapt dynamically. To alleviate the computational demands, some research such as [44], [47], and [43] adopts analyzing parts of the collected data offline, avoiding the needs posed by processing all the data from each sensor simultaneously in real time. Despite this strategy, the system must promptly make sense of this multifaceted data to adapt its interactions to the individual's current state and

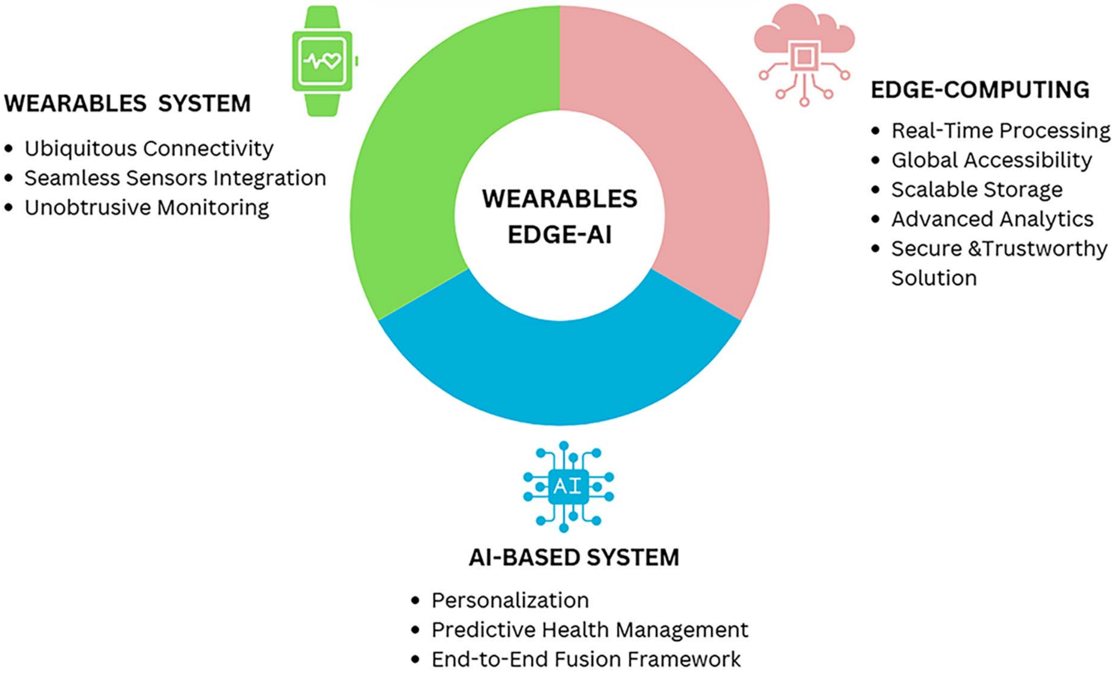

Fig. 6. General future improvement direction.

needs. This requires efficient algorithms and computational resources to handle the volume, storage, and variability of data, thereby enabling the support of real-time personalized feedback [19], [57].

B. Directions

In this section, we delve into the promising future directions in the context of autism research and intervention technologies as depicted in Fig. 6. Wearables with edge-AI provide a scalable and adaptive framework for addressing the complexities of data collection, synchronization, standardization, and real-time dynamics in ASD research and clinical practice.

1) Edge Computing for Real-Time and Adaptive Processing: Edge and cloud computing offer a complementary technological foundation to address critical challenges in multimodal autism research, particularly those related to data synchronization, latency, real-time feedback, storage capacity, and data privacy—as highlighted in Section VI. Edge computing plays a pivotal role in overcoming sensor misalignment and latency issues that arise in multisensor environments. By processing data locally—on edge devices such as embedded processors in wearable sensors or mobile platforms—researchers can perform real-time preprocessing, feature extraction, and data synchronization closer to the source. This localized handling allows for seamless alignment between modalities such as EEG, ET, and peripheral physiological signals, minimizing the need for post-hoc timestamp correction. Furthermore, edge computing supports closed-loop applications, including real-time feedback in VR environments or socially assistive robots, which rely on low-latency responses that other systems may not reliably provide. While edge computing optimizes real-time, low-latency interaction, cloud computing complements it by addressing computational scalability and long-term data management. Offloading heavier analytical tasks and model training to remote cloud servers enables the processing of larger datasets, which supports personalized adaptation and feedback delivery during assessment or intervention. Moreover, cloud storage provides a scalable infrastructure for securely managing the vast data collected in longitudinal autism studies as reported in [61] and [62]. This facilitates cross-institutional collaboration and opens opportunities for data sharing, which is essential for overcoming the limitation of small sample sizes and enhancing the generalizability of findings across diverse populations. Although edge computing holds immense potential for enhancing autism technology, safeguarding data security and user privacy is paramount. Researchers and developers may adopt a multilayered approach to building secure and trustworthy solutions. This includes prioritizing data anonymization and encryption using the FHIR standard and minimizing data collection to only essential elements. Robust secure connectivity measures are also vital, including in-transit encryption and trust management through digital certificates and secure boot processes. Additionally, security measures should consider the unique vulnerabilities of individuals with autism. Providing transparent data control options empowers users to manage their own information. Adhering to secure development practices and conducting regular audits will be crucial in collectively creating a secure and trustworthy ecosystem for technology powered by edge computing. Additionally, any research involving this data should be meticulously planned in collaboration with the appropriate institutional review boards to prevent any breaches of participant privacy and trust [100].

2) AI-Driven Personalization: The heterogeneity of autism symptoms poses a significant barrier to developing generalized diagnostic tools, as noted in the standardization and generalizability challenges. AI-based personalization offers a scalable solution by learning individual behavioral patterns over time. Machine learning models can adapt to each participant's unique sensory, emotional, and motor profiles, adjusting stimuli and task complexity accordingly. For instance, by analyzing the nuanced ways in which a child with autism responds to various stimuli, ML algorithms can predict which therapeutic activities will likely yield the best outcomes [102]. This helps to tailor the intervention in real time to suit the child's evolving preferences and responses during robot-assisted autism therapy sessions [103]. This individualized approach helps overcome the one-size-fits-all limitation and supports the development of flexible, inclusive, and personalized assessment systems—a key recommendation in response to the variability and spectrum nature of ASD discussed in Section VI. Moreover, the predictive power of AI and ML extends to foreseeing developmental trajectories and potential behavioral challenges, enabling preemptive adjustment of treatment plans. By anticipating future needs, these technologies enable the development of interventions that are proactive rather than merely reactive, enhancing the overall approach to care. Additionally, AI plays a transformative role in enhancing fusion approaches for multimodality data integration. Traditional methods, such as early or late fusion, often fall short of capturing the complex relationships across diverse modalities. In contrast, AI facilitates the use of end-to-end frameworks that seamlessly integrate data from multiple modalities, ensuring a more comprehensive and accurate analysis. For instance, advanced models such as CNNs and attention networks, as proposed in [43], can effectively process and fuse multimodal data by dynamically identifying the most relevant features from each modality. This capability enables a deeper understanding of the relationships between modalities, resulting in enhanced predictive accuracy and adaptability.

3) Advanced Data Collection With Wearables: Wearable technologies address multiple barriers in data collection, particularly those related to participant discomfort, dropouts, and limited monitoring duration. Unlike bulky or obtrusive lab-based setups, modern wearables (such as those embedded in clothes [61], [62], watches, hats, glasses, etc. [100]) offer lightweight, wireless, and noninvasive solutions that are well tolerated even by sensory-sensitive individuals. This reduces participant stress and improves compliance, allowing for longitudinal and ecologically valid data collection in naturalistic environments such as schools or homes. This promotes unobtrusive monitoring, reducing participant stress and improving compliance, which enables longitudinal and ecologically valid data collection in naturalistic environments such as schools and homes. Additionally, the ubiquitous connectivity of wearable systems allows for continuous data transmission across different settings, while seamless sensor integration ensures that multimodal data—including peripheral physiological signals (e.g., HR, EDA, and temperature)—can be reliably synchronized and analyzed in conjunction with behavioral metrics. As discussed in Section VI, these technologies can mitigate the challenge of small, fragmented datasets and increase the reliability and richness of collected signals, which are essential for building robust multimodal diagnostic systems.

VII. Conclusion

In conclusion, this survey highlights the significant advancements in multimodal sensing for the automated assessment and diagnosis of ASD. Traditional approaches have predominantly been based on single-modal data analysis, but recent progress in technology, cognitive science, and artificial intelligence has paved the way for exploring the potential benefits of integrating multiple sensory modalities. The survey categorizes key assistive technologies employed in autism research, delving into three main categories: 1) computer-based tasks; 2) vr simulations; and 3) robotic interactions. Within each technology, the exploration of utilized multimodalities, sensory data utilization, preprocessing and synchronization demanding, feature extraction methods, and fusion approaches is presented. The applications of multimodality are summarized across two main aspects: 1) diagnosis; and 2) intervention. While multimodal sensing offers immense promise for improving autism assessment, several challenges remain. Data collection faces limitations due to small sample sizes and the need for extensive data collection periods. Additionally, the utilization of multimodal data presents computational challenges related to synchronization demands and high-dimensional data processing. Future directions highlight the adoption of wearable with edge-AI in autism research. Edge computing facilitates real-time analysis of large datasets, AI and ML personalize interventions based on extensive data learning, and wearables enable continuous, unobtrusive monitoring. These advancements collectively promise significant improvements in autism intervention and understanding.

The current survey has several limitations. First, it does not delve into the technical aspects of multimodal data in detail. Rather, it focuses on providing an overview of the multimodality sensing technologies employed, the challenges of synchronizing and fusing these data, and their application in automated assessment and diagnosis in autism research. Second, the survey's literature search and organization were conducted from a behavioral perspective, primarily focusing on multimodal techniques for autism diagnosis and intervention based on sensory behavior data. Thus, the survey does not cover the pathogenesis of autism, including genetic studies or studies based on brain imaging techniques such as MRI. Despite these limitations, the survey provides a valuable overview of the current landscape of multimodal sensing in autism research. It highlights the potential of these technologies for improving autism assessment and intervention and identifies areas for further research and development.

References

[1] S. J. Rogers and B. Pennington, "A theoretical approach to the deficits in infantile autism," *Dev. Psychopathol.*, vol. 3, no. 2, pp. 137–162, 1991.

[2] D. L. Christensen et al., "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2012," *MMWR Surveill. Summ.*, vol. 65, no. 3, pp. 1–23, 2016.

[3] D. Bai et al., "Association of genetic and environmental factors with autism in a 5-country cohort," *JAMA Psychiat.*, vol. 76, no. 10, pp. 1035–1043, 2019.

[4] American Psychiatric Association, *The Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC, USA: American Psychiatric Association, 2013.

[5] X. Liu, Q. Wu, W. Zhao, and L. Xiong, "Technology-facilitated diagnosis and treatment of individuals with autism spectrum disorder: An engineering perspective," *Appl. Sci.*, vol. 7, no. 10, p. 1051, 2017.

[6] C. Li et al., "Improving the early screening procedure for autism spectrum disorder in young children: Experience from a community-based model in shanghai," *Autism Res.*, vol. 11, no. 9, pp. 1206–1217, 2018.

[7] R. G. Kent et al., "Diagnosing autism spectrum disorder: Who will get a DSM-5 diagnosis?" *J. Child Psychol. Psychiat.*, vol. 54, no. 11, pp. 1242–1250, 2013.

[8] S. K. Khare, S. March, P. D. Barua, V. M. Gadre, and U. R. Acharya, "Application of data fusion for automated detection of children with developmental and mental disorders: A systematic review of the last decade," *Inf. Fusion*, vol. 99, 2023, Art. no. 101898.

[9] N. Glaser and M. Schmidt, "Systematic literature review of virtual reality intervention design patterns for individuals with autism spectrum disorders," *Int. J. Human–Comput. Interact.*, vol. 38, no. 8, pp. 753–788, 2022.

[10] W. Farzana, F. Sarker, T. Chau, and K. A. Mamun, "Technological evolvement in AAC modalities to foster communications of verbally challenged ASD children: A systematic review," *IEEE Access*, early access, Jan. 2021, doi: 10.1109/ACCESS.2021.3055195.

[11] 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.

[12] M. Kohli, A. K. Kar, and S. Sinha, "Robot facilitated rehabilitation of children with autism spectrum disorder: A 10 year scoping review," *Expert Syst.*, vol. 40, no. 5, 2023, Art. no. e13204.

[13] F. Thabtah and D. Peebles, "Early autism screening: A comprehensive review," *Int. J. Environ. Res. Public Health*, vol. 16, no. 18, 2019, Art. no. 3502.

[14] L. Vllasaliu et al., "Diagnostic instruments for autism spectrum disorder (ASD)," *Cochrane Database System. Rev.*, vol. 2016, no. 1, 1996.

[15] A. Valizadeh et al., "Automated diagnosis of autism with artificial intelligence: State of the art," *Rev. Neurosci.*, vol. 35, no. 2, pp. 141–163, 2024.

[16] G. Baird, H. R. Douglas, and M. S. Murphy, "Recognising and diagnosing autism in children and young people: Summary of NICE guidance," *BMJ*, vol. 343, 2011, Art. no. d6360.

[17] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, "Detecting high-functioning autism in adults using eye tracking and machine learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1254–1261, Jun. 2020.

[18] C. Tang et al., "Automatic identification of high-risk autism spectrum disorder: A feasibility study using video and audio data under the still-face paradigm," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2401–2410, Nov. 2020.

[19] J. Han, G. Jiang, G. Ouyang, and X. Li, "A multimodal approach for identifying autism spectrum disorders in children," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2003–2011, 2022.

[20] A. R. Levin et al., "EEG power at 3 months in infants at high familial risk for autism," *J. Neurodevelop. Disord.*, vol. 9, no. 1, pp. 1–13, 2017.

[21] E. J. H. Jones et al., "Developmental pathways to autism: A review of prospective studies of infants at risk," *Neurosci. Biobehav. Rev.*, vol. 39, pp. 1–33, 2014.

[22] J. Han et al., "Development of brain network in children with autism from early childhood to late childhood," *Neuroscience*, vol. 367, pp. 134–146, Dec. 2017.

[23] T.-M. Heunis, C. Aldrich, and P. J. De Vries, "Recent advances in resting-state electroencephalography biomarkers for autism spectrum disorder—A review of methodological and clinical challenges," *Pediatric Neurol.*, vol. 61, pp. 28–37, Aug. 2016.

[24] T. Takahashi et al., "Enhanced brain signal variability in children with autism spectrum disorder during early childhood," *Hum. Brain Mapp.*, vol. 37, no. 3, pp. 1038–1050, Mar. 2016.

[25] G. Bird and R. Cook, "Mixed emotions: The contribution of alexithymia to the emotional symptoms of autism," *Transl. Psychiatry*, vol. 3, no. 7, p. e285, 2013.

[26] C. Lord, M. Rutter, P. S. DiLavore, and S. Risi, 1999, *Autism Diagnostic Observation Schedule: Manual.* Los Angeles, LA, USA: Western Psychological Services

[27] C. Lord et al., "Autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Develop. Disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[28] S. Zhang, D. Chen, Y. Tang, and L. Zhang, "Children autism evaluation through joint analysis of EEG and eye-tracking recordings with graph convolution network," *Front. Hum. Neurosci.*, vol. 15, p. 651349, 2021.

[29] L. Zhao et al., "A multimodal data driven rehabilitation strategy Auxiliary feedback method: A case study," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1181–1190, 2022.

[30] M. Cheng et al., "Computer-aided autism spectrum disorder diagnosis with behavior signal processing," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2982–3000, Oct./Dec. 2023.

[31] D. Bian et al., "A novel multisensory stimulation and data capture system (MADCAP) for investigating sensory trajectories in infancy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1526–1534, Aug. 2018.

[32] J. J. Thompson, *Beyond Words: Nonverbal Communication in the Classroom.* New York, NU, USA: MacMillan Publishing Company, 1973.

[33] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, "A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 353–361, Feb. 2018.

[34] J. P. Noel, M. A. De Niear, N. S. Lazzara, and M. T. Wallace, "Uncoupling between multisensory temporal function and nonverbal turn-taking in autism spectrum disorder," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 4, pp. 973–982, Dec. 2018.

[35] C.-P. Chen, X.-H. Tseng, S. S.-F. Gau, and C.-C. Lee, "Computing multimodal dyadic behaviors during spontaneous diagnosis interviews toward automatic categorization of autism spectrum disorder," in *Proc. Interspeech*, 2017, pp. 2361–2365.

[36] Lin, Y. S. Gau, S. S. F, and Lee, C. C., "A multimodal interlocutor-modulated attentional BLSTM for classifying autism subgroups during clinical interviews," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 299–311, Feb. 2020.

[37] J. Rehg et al., "Decoding children's social behavior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3414–3421.

[38] C. D. Heath, H. Venkateswara, T. McDaniel, and S. Panchanathan, "Using multimodal data for automated fidelity evaluation in pivotal response treatment videos," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Piscataway, NJ, USA: IEEE, Nov. 2019, pp. 1–5.

[39] O. Rudovic, M. Zhang, B. Schuller, and R. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 6–15.

[40] L. Deng, P. Rattadilok, and R. Xiong, "A machine learning-based monitoring system for attention and stress detection for children with autism spectrum disorders," in *Proc. 2021 Int. Conf. Intell. Med. Health*, Aug. 2021, pp. 23–29.

[41] A. Alcorn et al., "Social communication between virtual characters and children with autism," in *Proc. Int. Conf. Artif. Intell. Educ.,* Berlin, Heidelberg, Germany: Springer, 2011, pp. 7–14.

[42] P. Mesa-Gresa, H. Gil-Gómez, J.-A. Lozano-Quilis, and J.-A. Gil-Gómez, "Effectiveness of virtual reality for children and adolescents with autism spectrum disorder: An evidence-based systematic review," *Sensors*, vol. 18, no. 8, p. 2486, 2018.

[43] L. Zhang et al., "Cognitive load measurement in a virtual reality-based driving system for autism intervention," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 176–189, Apr.–Jun. 2017.

[44] D. Adiani et al., "Career interview readiness in virtual reality (CIRVR): A platform for simulated interview training for autistic individuals and their employers," *ACM Trans. Accessible Comput. (TACCESS)*, vol. 15, no. 1, pp. 1–28, 2022.

[45] A. Z. Amat et al., "Design of a desktop virtual reality-based collaborative activities simulator (ViRCAS) to support teamwork in workplace settings for autistic adults," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2184–2194, 2023.

[46] A. Meier, H. Spada, and N. Rummel, “A rating scheme for assessing the quality of computer-supported collaboration processes,” *Int. J. Comput.-Supported Collab. Learn.*, vol. 2, no. 1, pp. 63–86, Mar. 2007.

[47] Z. K. Zheng, N. Sarkar, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, “CheerBrush: A novel interactive augmented reality coaching system for toothbrushing skills in children with autism spectrum disorder,” *ACM Trans. Access. Comput.*, vol. 14, no. 4, p. 20, 2021, doi: 10.1145/3481642.

[48] E. Bekele, Z. Zheng, A. Swanson, J. Crittendon, Z. Warren, and N. Sarkar, “Understanding how adolescents with autism respond to facial expressions in virtual reality environments,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 4, pp. 711–720, Apr. 2013, doi: 10.1109/TVCG.2013.42.

[49] E. Bekele et al., “Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (autism),” in *Proc. IEEE Virtual Reality (VR)*, Piscataway, NJ, USA: IEEE, Mar. 2016, pp. 121–130.

[50] A. Koirala, Z. Yu, H. Schiltz, A. Van Hecke, K. A. Koth, and Z. Zheng, “An exploration of using virtual reality to assess the sensory abnormalities in children with autism spectrum disorder,” in *Proc. 18th ACM Int. Conf. Interaction Des. Children*, Jun. 2019, pp. 293–300.

[51] C. Cheroni, N. Caporale, and G. Testa, “Autism spectrum disorder at the crossroad between genes and environment: Contributions, convergences, and interactions in autism developmental pathophysiology,” *Mol. Autism.*, vol. 11, no. 1, p. 69, 2020.

[52] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, “The clinical use of robots for individuals with autism spectrum disorders: A critical review,” *Res. Autism Spectr. Disorders*, vol. 6, no. 1, pp. 249–262, 2012.

[53] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, “Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with autism,” *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 2, pp. 125–135, Apr. 2018.

[54] D. François, S. Powell, and K. Dautenhahn, “A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children’s proactivity and initiative-taking,” *Interact. Stud.*, vol. 10, no. 3, pp. 324–373, 2009.

[55] K. Dautenhahn, I. Werry, T. Salter, and I. R. J. A. Te Boekhorst, “Towards adaptive autonomous robots in autism therapy: Varieties of interactions,” in *Proc. IEEE Int. Symp. Comput. Intell. Robot. Autom.*, 2003, pp. 577–582.

[56] B. Scassellati, H. Admoni, and M. Mataric, “Robots for use in autism research,” *Annu. Rev. Biomed. Eng.*, vol. 14, pp. 275–294, May 2012.

[57] H. Cai et al., “Sensing-enhanced therapy system for assessing children with autism spectrum disorders: A feasibility study,” *IEEE Sensors J.*, vol. 19, no. 4, pp. 1508–1518, Feb. 2019.

[58] C. Wong et al., “Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review,” *J. Autism Develop. Disorders*, vol. 45, no. 7, pp. 1951–1966, 2015.

[59] M. Del Coco et al., “Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 4, pp. 993–1004, Dec. 2018.

[60] Z. Zheng, E. M. Young, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, “Robot-mediated imitation skill training for children with autism,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 6, pp. 682–691, Jun. 2016.

[61] W. Xiao, M. Li, M. Chen, and A. Barnawi, “Deep interaction: Wearable robot-assisted emotion communication for enhancing perception and expression ability of children with Autism Spectrum Disorders,” *Future Gener. Comput. Syst.*, vol. 108, pp. 709–716, 2020.

[62] M. Chen, W. Xiao, L. Hu, Y. Ma, Y. Zhang, and G. Tao, “Cognitive wearable robotics for autism perception enhancement,” *ACM Trans. Internet Technol. (TOIT)*, vol. 21, no. 4, pp. 1–16, 2021.

[63] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Mataric, “Toward personalized affect-aware socially assistive robot tutors in long-term interventions for children with autism,” 2021, *arXiv:2101.10580*.

[64] K. A. Funes Mora, F. Monay, and J. M. Odobez, “EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras,” in *Proc. Symp. Eye Tracking Res. IEEE Int. Symp. Spread Spectr. Tech. Appl.*, Mar. 2014, pp. 255–258.

[65] A. Olsen, “The Tobii I-VT fixation filter,” *Tobii Technol.*, vol. 21, pp. 4–19, 2012.

[66] W. Bosl, A. Tierney, H. Tager-Flusberg, and C. Nelson, “EEG complexity as a biomarker for autism spectrum disorder risk,” *BMC Med.*, vol. 9, p. 18, 2011, doi: 10.1186/1741-7015-9-18.

[67] W. J. Bosl, H. Tager-Flusberg, and C. A. Nelson, “EEG analytics for early detection of autism spectrum disorder: A data-driven approach,” *Sci. Rep.*, vol. 8, pp. 1–20, 2018, doi: 10.1038/s41598-018-24318-x.

[68] A. Catarino, O. Churches, S. Baron-Cohen, A. Andrade, and H. Ring, “Atypical EEG complexity in autism spectrum conditions: A multiscale entropy analysis,” *Clin. Neurophysiol.*, vol. 122, no. 12, pp. 2375–2383, Dec. 2011.

[69] S. Matlis, K. Boric, C. J. Chu, and M. A. Kramer, “Robust disruptions in electroencephalogram cortical oscillations and large-scale functional networks in autism,” *BMC Neurol.*, vol. 15, no. 1, p. 97, Dec. 2015.

[70] A. R. Levin, K. J. Varcin, H. M. O’Leary, H. Tager-Flusberg, and C. A. Nelson, “EEG power at 3 months in infants at high familial risk for autism,” *J. Neurodevelop. Disorders*, vol. 9, no. 1, p. 34, Dec. 2017.

[71] K. Zeng et al., “Disrupted brain network in children with autism spectrum disorder,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, Dec. 2017.

[72] N. Qiu et al., “Application of the still-face paradigm in early screening for high-risk autism spectrum disorder in infants and toddlers,” *Front. Pediatr.*, vol. 8, p. 290, 2020.

[73] M. Del Coco et al., “A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW),* 2017, pp. 1401–1407.

[74] R. Y. Cai, A. L. Richdale, M. Uljarević, C. Dissanayake, and A. C. Samson, “Emotion regulation in autism spectrum disorder: Where we are and where we need to go,” *Autism Res.*, vol. 11, no. 7, pp. 962–978, 2018.

[75] N. Qiu et al., “Application of the stillface paradigm in early screening for high-risk autism spectrum disorder in infants and toddlers,” *Front. Pediatr.*, vol. 8, p. 290, 2020.

[76] P. M. Beall, E. J. Moody, D. N. McIntosh, S. L. Hepburn, and C. L. Reed, “Rapid facial reactions to emotional facial expressions in typically developing children and children with autism spectrum disorder,” *J. Exp. Child Psychol.*, vol. 101, no. 3, pp. 206–223, Nov. 2008.

[77] C. Liu, K. Conn, N. Sarkar, and W. Stone, “Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder,” *Int. J Human-Comput. Stud.*, vol. 66, no. 9, pp. 662–677, 2008.

[78] A. Rozga, T. Z. King, R. W. Vuduc, and D. L. Robins, “Undifferentiated facial electromyography responses to dynamic, audio-visual emotion displays in individuals with autism spectrum disorders,” *Dev. Sci.*, vol. 16, no. 4, pp. 499–514, 2013.

[79] K. Briot, A. Pizano, M. Bouvard, and A. Amestoy, “New technologies as promising tools for assessing facial emotion expressions impairments in autism: A systematic review,” *Front. Psychiatry*, vol. 12, 2021, Art. no. 634756.

[80] E. Covi et al., “Adaptive extreme edge computing for wearable devices,” *Front. Neurosci.*, vol. 15, 2021, Art. no. 611300.

[81] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th ed. Arlington, VA, USA: American Psychiatric Association, 2013, p. xliv, 947p.

[82] J. R. Bergstrom, S. Duda, D. Hawkins, and M. McGill, “Physiological response measurements,” in *Eye Tracking in User Experience Design.* San Mateo, CA, USA: Morgan Kaufmann, 2014, pp. 81–108.

[83] L. M. Vortmann, S. Ceh, and F. Putze, “Multimodal EEG and eye tracking feature fusion approaches for attention classification in hybrid BCIs,” *Front. Comput. Sci.*, vol. 4, 2022, Art. no. 780580.

[84] J. A. Healey, “Wearable and automotive systems for affect recognition from physiology,” Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2000. Advisor(s) Rosalind W. Picard.

[85] N. Sarkar, “Psychophysiological control architecture for human-robot coordination-concepts and initial experiments,” in *Proc. IEEE Int. Conf. Robot. Automat.*, Washington, DC, USA, vol. 4. Piscataway, NJ, USA: IEEE, 2002, pp. 3719–3724.

[86] J. Kim and E. André, “Emotion recognition based on physiological changes in music listening,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[87] A. Pecchinenda, “The affective significance of skin conductance activity during a difficult problem-solving task,” *Cognition Emotion*, vol. 10, no. 5, pp. 481–504, 1996.

[88] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *Int. J. Psychophysiol.*, vol. 61, no. 1, pp. 5–18, 2006.

[89] C. M. A. van Ravenswaaij-Arts, L. A. A. Kollee, J. C. W. Hopman, G. B. A. Stoelinga, and H. P. van Geijn, "Heart rate variability," *Ann. Internal Med.*, vol. 118, pp. 436–447, 1993.

[90] Empatica, "E4 Wristband: Real-time physiological data streaming and visualization," 2021. Accessed: Sep. 11, 2021. [Online]. Available: https://www.empatica.com/research/e4/

[91] K. A. Fournier, C. J. Hass, S. K. Naik, N. Lodha, and J. H. Cauraugh, "Motor coordination in autism spectrum disorders: A synthesis and meta-analysis," *J. Autism Dev. Disord.*, vol. 40, pp. 1227–1240, 2010.

[92] M. McPhillips, J. Finlay, S. Bejerot, and M. Hanley, "Motor deficits in children with autism spectrum disorder: A cross-syndrome study," *Autism Res.*, vol. 7, pp. 664–676, 2014.

[93] E. Khoury, L. Carment, P. Lindberg, R. Gaillard, M. O. Krebs, and I. Amado, "Sensorimotor aspects and manual dexterity in autism spectrum disorders: A literature review," *L'encephale*, vol. 46, pp. 135–145, 2020.

[94] A. N. Bhat, "Motor impairment increases in children with autism spectrum disorder as a function of social communication, cognitive and functional impairment, repetitive behavior severity, and comorbid diagnoses: A SPARK study report," *Autism Res.*, vol. 14, pp. 202–219, 2021, doi: 10.1002/aur.2453.

[95] A. Ardalan, A. H. Assadi, O. J. Surgent, and B. G. Travers, "Whole-body movement during videogame play distinguishes youth with autism from youth with typical development," *Sci. Rep.*, vol. 9, no. 1, p. 11, 2019, doi: 10.1038/s41598-019-56362-6.

[96] Z. Zhao et al., "Excessive and less complex body movement in children with autism during face-to-face conversation: An objective approach to behavioral quantification," *Autism Res.*, vol. 15, pp. 305–316, 2022, doi: 10.1002/aur.2646.

[97] R. B. Wilson, P. G. Enticott, and N. J. Rinehart, "Motor development and delay: Advances in assessment of motor skills in autism spectrum disorders," *Curr. Opin. Neurol.*, vol. 31, pp. 134–139, 2018.

[98] E. B. Varghese, M. Qaraqe, D. Al Thani, and H. K. Ekenel, "Attention assessment in children with autism using head pose and motion parameters from real videos," in *Proc. IEEE Global Commun. Conf. GLOBECOM*, Piscataway, NJ, USA: IEEE, Dec. 2023, pp. 6462–6468.

[99] J. Lazar, J. H. Feng, and H. Hochheiser, "Chapter 15—Working with human subjects," in *Research Methods in Human Computer Interaction*, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2017, pp. 455–491, doi: 10.1016/B978-0-12-805390-4.00015-7.

[100] J. Lazar, J. H. Feng, and H. Hochheiser, "Chapter 12—Automated data collection methods," in *Research Methods in Human Computer Interaction*, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2017, pp. 329–368, doi: 10.1016/B978-0-12-805390-4.00012-1.

[101] D. Bone, T. Chaspari, and S. Narayanan, "Behavioral signal processing and autism: Learning from multimodal behavioral signals," in *Autism Imaging and Devices.* Boca Raton, FL, USA: CRC Press, 2017, pp. 335–360.

[102] L. J. Marcos-Zambrano et al., "Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment," *Front. Microbiol.*, vol. 12, 2021, Art. no. 634511.

[103] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Sci. Robot.*, vol. 3, no. 19, 2018, Art. no. eaao6760.

[104] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[105] R. M. Stern, *Psychophysiological Recording.* Oxford, U.K.: Oxford Univ. Press, 2001.

[106] R. L. Mandryk and K. M. Inkpen, "Physiological indicators for the evaluation of co-located collaborative play," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, Nov. 2004, pp. 102–111.

[107] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, and T. A. Ito, "The psychophysiology of emotion," in *Handbook of Emotions*, M. Lewis and J. M. Haviland-Jones, Eds., 2nd ed. New York, NY, USA: Guilford Press, 2000, pp. 173–191.

[108] R. M. Stern, W. J. Ray, and K. S. Quigley, *Psychophysiological Recording.* Oxford, U.K.: Oxford Univ. Press, 2001.

[109] J. Lazar, J. H. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction.* San Mateo, CA, USA: Morgan Kaufmann, 2017.

[110] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard, "Frustrating the user on purpose: A step toward building an affective computer," *Interact. Comput.*, vol. 14, no. 2, pp. 93–118, 2002.

Athmar N. M. Shamhan received the B.Tech. degree in information technology from Taiz University, Taiz, Yemen, in 2018, and the M.A. degree in artificial intelligence from Warsaw University of Technology, Warsaw, Poland, in 2023. She is currently working toward the Ph.D. degree in computer science and engineering with the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.

Prior to her Ph.D. studies, she was a Software Engineer with Rockwell Automation, Katowice, Poland, where she developed industrial control and automation solutions. Her research focuses on computer vision, machine learning, and human–computer interaction, with a particular emphasis on computational genomics and de Bruijn graph-based genome assembly.

Marwa Qaraqe received the bachelor's degree from Texas A&M University, Doha, Qatar, in 2010, and the M.Sc. and Ph.D. degrees from Texas A&M University, College Station, TX, USA, in 2012 and 2016, respectively, all in electrical engineering.

Currently, she is an Associate Professor with the Division of Information and Communication Technology, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar. She is also a Co-Founder of the Autism Sensing Center, College of Science and Engineering, HBKU. Her research focuses on wireless communication, signal processing, and machine learning, with multidisciplinary applications spanning security, the Internet of Things (IoT), and health. She has a particular interest in physical layer security, reconfigurable intelligent surfaces, and machine learning techniques for enhancing wireless communication, security, and health. Passionate about predictive health analytics and personalized learning, she actively explores innovative approaches to managing health disorders through artificial intelligence (AI)-driven solutions. In addition, she is deeply dedicated to research that leverages AI, sensing technology, and computer vision for the nonsubjective assessment of behavior in children with autism, aiming to improve intervention strategies, outcomes, and learning.

Dena Al-Thani (Member, IEEE) received the Ph.D. degree in computer science from Queen Mary University of London, London, UK, in 2013.

Currently, she is an Associate Professor, a Co-Founder of the A-sense Center of Excellence, and the Head of the Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar. She actively publishes in top journals and presents at international conferences. Her research interests include accessibility, inclusive design, and eHealth. Her research on inclusion aims to make a global impact.

Dr. Al-Thani is a member of the WHO's technical advisory group on assistive technology and the Arab ICT Accessibility Expert Group, led by Mada.