mathematics

MDPI

Article

IMITASD: Imitation Assessment Model for Children with Autism Based on Human Pose Estimation

Hany Said [1,†], Khaled Mahar [2,3], Shaymaa E. Sorour [4,5,*], Ahmed Elsheshai [6], Ramy Shaaban [7], Mohamed Hesham [6], Mustafa Khadr [8], Youssef A. Mehanna [8], Ammar Basha [8] and Fahima A. Maghraby [9,†]

[1] College of Artificial Intelligence, Arab Academy for Science, Technology, and Maritime Transport, El Alamein 51718, Egypt; hanysaid2000@aast.edu

[2] Arab Center for Artificial Intelligence, Arab Academy for Science, Technology, and Maritime Transport, Alexandria 21532, Egypt; khmahar@aast.edu

[3] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Alexandria 21532, Egypt

[4] Department of Management Information Systems, School of Business, King Faisal University, Alhufof 31982, Saudi Arabia

[5] Faculty of Specific Education, Kafrelsheikh University, Kafrelsheikh 33511, Egypt

[6] College of Medicine, Arab Academy for Science, Technology, and Maritime Transport, El Alamein 51718, Egypt; ahmed.elsheshai@aast.edu (A.E.); mhesham@aast.edu (M.H.)

[7] Department of Instructional Technology and Learning Sciences, Utah State University, Salt Lake City, UT 84322, USA; ramy.shaaban@usu.edu

[8] Research and Innovation Center, Arab Academy for Science, Technology, and Maritime Transport, El Alamein 51718, Egypt; mostafakhedr4674@gmail.com (M.K.); yousefmohana@student.aast.edu (Y.A.M.); ammarbasha@student.aast.edu (A.B.)

[9] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Cairo 2033, Egypt; fahima@aast.edu

[*] Correspondence: ssorour@kfu.edu.sa

[†] These authors contributed equally to this work.

check for updates

Citation: Said, H.; Mahar, K.; Sorour, S.E.; Elsheshai, A.; Shaaban, R.; Hesham, M.; Khadr, M.; Mehanna, Y.A.; Basha A.; Maghraby, F.A. IMITASD: Imitation Assessment Model for Children with Autism Based on Human Pose Estimation. *Mathematics* **2024**, *12*, 3438. https://doi.org/10.3390/math12213438

Academic Editors: Huawen Liu, Chengyuan Zhang and Chunwei Tian

Received: 9 October 2024
Revised: 31 October 2024
Accepted: 1 November 2024
Published: 3 November 2024

Abstract: Autism is a challenging brain disorder affecting children at global and national scales. Applied behavior analysis is commonly conducted as an efficient medical therapy for children. This paper focused on one paradigm of applied behavior analysis, imitation, where children mimic certain lessons to enhance children's social behavior and play skills. This paper introduces IMITASD, a practical monitoring assessment model designed to evaluate autistic children's behaviors efficiently. The proposed model provides an efficient solution for clinics and homes equipped with mid-specification computers attached to webcams. IMITASD automates the scoring of autistic children's videos while they imitate a series of lessons. The model integrates two core modules: attention estimation and imitation assessment. The attention module monitors the child's position by tracking the child's face and determining the head pose. The imitation module extracts a set of crucial key points from both the child's head and arms to measure the similarity with a reference imitation lesson using dynamic time warping. The model was validated using a refined dataset of 268 videos collected from 11 Egyptian autistic children during conducting six imitation lessons. The analysis demonstrated that IMITASD provides fast scoring, takes less than three seconds, and shows a robust measure as it has a high correlation with scores given by medical therapists, about 0.9, highlighting its effectiveness for children's training applications.

Keywords: autism; imitation; attention; MediaPipe; dynamic time warping; human pose estimation

MSC: 68T05

1. Introduction

Autism spectrum disorder (ASD) is a complex neurodevelopmental condition characterized by difficulties in communication and repetitive behaviors. These challenges significantly impact the daily lives of individuals with ASD and their families, specifically in developing and acquiring social and cognitive skills. According to the Centers for Disease Control and Prevention (CDC) statement in 2023, the rate of autistic children has risen from 1 in 44 to 1 in 36. According to their statistics, that covered eleven states in the USA, boys are four times more likely to be autistic than girls, while children are commonly diagnosed with autism at the age of four years [1,2]. Globally, there are about 75 million people who have an autism disorder [3]. According to the autism rate by country, Egypt ranks twentieth on the list of countries suffering from the spread of autism, where the prevalence per ten thousand children is 89.40 [3]. Imitation, joint attention, and turn-taking are used in the context of applied behavior analysis (ABA) [4,5], which provides one of the most effective therapies for children in clinical settings, along with other recent studies such as deep pressure therapy that provides interesting intervention to reduce anxiety by measuring electroencephalograms for participants while wearing inflatable vests. Winarni et al. [6–8] proposed sensory integration therapists who focused on improving the behavioral responses of autistic children using their sensory inputs such as touch and body movement. Particularly, they studied the effect of deep pressure on children, using the portable hug machine. They confirmed that this type of deep pressure minimizes the children's stress response during children's travels. These studies provide interesting directions to tackle children's stress, where ABA seeks to enhance children's skills by deploying various interventions to ease children's coping with their communities. Imitation is mimicking behaviors which are crucial for social interaction and learning; it is an essential component for children with ASD, who often struggle with learning and developing social engagement [9–11]. It allows the autistic child to acquire new skills and interact meaningfully with their environment [12]. Joint attention fosters shared focus, it reflects the child's ability to focus on other activities, supporting cognitive development [9,13,14]. Lastly, turn-taking promotes social interaction, it supports an advanced social skill that helps children conduct normal interactions and relationships with their surroundings [15–17].

Autism is a natural learning disorder that is usually identified within the first two years of autistic children's lives [10,18]. Children with autism lack learning capabilities for behavioral patterns that non-disorder learning children usually acquire. Imitation is categorized into motor, sound, and verbal imitation. One of the major benefits of imitation is improving social functionality for autistic children, which can improve children's integration in the surrounding society [18]. Although normal imitation sessions in clinics provide a suitable environment for health therapy, they raise significant challenges, especially in developing countries. First, the autistic child tends to feel bored when the clinic is filled with children with autism. This increases the difficulty of guiding the child during the session therapy. Regular therapy typically requires two sessions per week to be aligned with a typical therapy plan. This demands great effort, especially for families who live in distant areas. The ultimate goal of this research is to enable an efficient framework to conduct imitation remotely that does not require on-site sessions. Proposing a model that provides a score for the children's imitation is necessary to develop such a framework.

This study utilized imitation as a foundational component of ABA therapy. It plays a fundamental role in acquiring new skills and fostering social connections. However, children with autism may exhibit difficulties in imitating actions, which impacts their ability to learn and engage with their surroundings. Advancements in technology have sparked a surge in innovative approaches for training and assessment of autistic patients. From primitive methods to advanced methods represented in machine learning and deep learning, these methodologies have exhibited promise in aiding the assessment of children with ASD. However, a critical evaluation of these studies reveals a common limitation that hinders their widespread applicability in real-world scenarios. Previous investigations in this domain have often relied on sophisticated equipment such as high-end cameras or specialized devices like Kinect for data collection and analysis. While these tools yield valuable insights, their practicality in everyday environments remains restricted, thereby limiting their utility in routine assessments [19–21].

This paper proposes Imitation Assessment for Children with Autism Spectrum Disorder (IMITASD). It emphasizes seamless integration of the used modules and prioritizes practical techniques applicable in everyday settings such as mid-specification computers with a webcam. Central to our approach is the utilization of human pose estimation with an efficient time-series measure. Accordingly, the proposed method relies on deploying MediaPipe and dynamic time warping (DTW) [22,23] as key elements for IMITASD. Therefore, it extends the scope beyond gait behavior assessment to encompass a comprehensive evaluation of imitation movements crucial for ABA in children with ASD. Our proposed tool not only facilitates practical assessments but also endeavors to pave the way for more comprehensive and practical therapeutic strategies for children diagnosed with ASD.

In this study, we present IMITASD, a novel tool designed to assess the behaviors of children with autism spectrum disorder (ASD) through imitation interventions. IMITASD is specifically developed to provide accurate scoring that closely aligns with therapist evaluations, enhancing the assessment process. The dataset for this tool is gathered from Egyptian autistic children, fulfilling the need for culturally sensitive and contextually relevant tools for the region. Furthermore, IMITASD addresses key limitations of prior research, which often relied on expensive equipment such as high-end cameras, Kinect devices, and specialized setups. By offering a practical, low-cost, and fast solution, IMITASD ensures high-validity assessments, making it an invaluable resource for both clinical and educational settings.

This article is organized as follows. The literature survey is presented in Section 2, followed by Section 3, which provides a brief description of the set of algorithms and tools used in this study. After that, the methodology and the IMITASD system architecture are demonstrated in Sections 4 and 5, respectively. Lastly, the results, followed by a detailed discussion, are articulated in Section 6. Conclusions drawn from the findings are presented in Section 7, and potential directions for future work are discussed in Section 8.

2. Related Works

Artificial intelligence's (AI) application reshapes different fields, such as health monitoring, energy optimization, and machining. In health [24], AI assists in predictive diagnostics and personalized treatment plans. In energy [25], it optimizes resource consumption and improves system reliability, while in machining [26], AI automates processes and enhances precision. These advancements demonstrate AI's capacity to revolutionize various sectors by improving efficiency and decision-making capabilities.

Regarding ASD detection, AI tools have investigated various novel techniques for estimating ASD scores. The authors in [19] explored the potential of digital biomarkers, such as eye gaze, tracked through wearable devices like smartphones, in aiding the early diagnosis and intervention of ASD in preschool children. The study's limitations include the absence of information about monitoring activities, potentially hindering understanding of children's attention prompts, and the challenge of losing temporal behavior nuances. In [27], Farooq, M.s. et al. focus on utilizing federated learning as a promising approach for ASD detection. The study employed support vector machine (SVM) [28] and logistic regression (LR) [29] models, showing their effectiveness in detecting ASD in diverse age groups. The authors acknowledge limitations, including constrained model complexity due to decentralized training on devices with limited resources. In [30], Suman R. and Sarfaraz M. used SVM, naïve Bayes [31], k-nearest neighbor [32], artificial neural networks, and convolutional neural networks (CNNs) [33] to identify ASD across diverse age demographics. The evaluation metrics reveal that the performance of CNNs achieves accuracy levels ranging from 95.75% to 99.53% for the UCI Repository datasets. While the study achieved notable accuracy in ASD detection, limitations include reliance on publicly available datasets and the absence of a standardized medical test for ASD.

The following studies present various techniques for the estimation of ASD scores and the classification of children as either neurotypical or on the autism spectrum. M. Wang and, N. Yang have proposed a model named Observational Therapy-Assistance

Neural Network (OTA-NN) [34]. It is based on two components: the first is based on a spatial–temporal Transformer and the second consists of multiple-instance learning (MIL). The two components are responsible for extracting the 3D skeleton and deploying a set of multiple learners that can score a child's training state during a medical treatment session. They proposed another study that replaced their spatial–temporal Transformer with a graph convolutional network (GCN) with MIL [35,36]. Their model is tested using videos available in the Dream dataset collected by five cameras, and it obtains an area under the curve (AUC) score of 0.824. The main restriction of the studies is the setting requirements, which means it cannot be deployed with only a single camera. S. Zahan et al. [37] proposed a model based on gesture analysis with movement patterns for normal and autistic children. Their model predicts the ASD severity score using the Autism Diagnostic Observation Schedule (ADOS) [38], where they developed a model based on a hypothesis that refers to disparity gesture patterns: asymmetric movement and gait. The model uses a graph convolution network (GCN) to extract gait posture, while the Vision Transformer is used to process skeleton frames in terms of patches from various perspectives. Although the authors proved a strong potential factor in differentiating normal from autistic children, they relied on using a camera with Kinect v2, which is crucial for their model to capture the human skeleton from different angles.

Varun G. et al. [39] developed three models that assess ABA that focus on activity comprehension, joint attention, and emotion from facial expression. They deployed a spatiotemporal Transformer to assess interaction between children and their therapist. They compiled about 51 K images to train the emotion and facial expression model using the ResNet-34 deep learning model. For joint attention, they implemented R-CNN with ResNet-50 (v1) to assess the children's performance while interacting with therapists to look or point in a certain direction. Their results indicate acceptable performance for activity comprehension, while high scores are obtained for joint attention and facial expression modules. Their study has several limitations. First, they used an adjusted camera on a stable tripod to fit the clinic area. Their model did not extend to integrating these modules into a single pipeline. Second, since they focused on activity comprehension, they did not focus on assessing the imitation, as the covered actions were running, sitting, and walking. Ahmed A et al., in [20], provided a study that classified children according to the level of autism severity. They used Samsung Note 9 and Kinect v2 to record videos of children while they walked for 1.5 m in front of the camera. They used seven Transformers for data augmentation while computing a set of distances between different joints. A multi-layer perceptron (MLP) network was used during the model training, resulting in an accuracy of 95%. Although their study provides a practical approach to classify children into normal or autistic, they cover only gait behavior, which does not align with imitation, especially when the target is to enhance primary children's behavior for ABA. Furthermore, they used Kinect v2, which might not be available at the patient's home.

Table 1 summarizes the previous studies. The limitations across these studies include challenges in real-world applicability due to reliance on high-end cameras, Kinect devices, or specialized setups. Further work should address the integration of modules, focusing on practical techniques applicable in everyday settings. Additionally, advancements should extend beyond gait behavior to encompass comprehensive imitation movements for ABA. It is crucial to tackle these challenges to provide a low-cost, fast, and practical solution. This tool would provide a high-validity assessment for movement when applied to ASD children, where the results should have a significant degree of correlation with a medical expert's evaluation.

Table 1. Summary of studies on ASD diagnosis and treatment.

#	Study	Models/Techniques	Contributions	Limitations
1	Sandhu, G et al., 2022 [19]	Eye gaze tracked through wearable devices like smartphones for aiding in the early diagnosis of ASD children	Use digital biomarkers to monitor ASD children's performance	Absence of information about training activities
2	Ahmed A et al., 2020 [20]	Principal component analysis; multi-layer perceptron network	Model's accuracy is 95% in classifying videos of children according to the level of autism	Covers only gait behavior, which does not align with imitation
3	Farooq, M. S. et al., 2023 [27]	Support vector machine and logistic regression models	Federated learning model shows its effectiveness in detecting ASD	Certain measures must be fed manually into the system, e.g., sensory processing, repetitive behavior, and other parameters
4	Suman Raj and Sarfaraz Masood, 2020 [30]	Support vector machine, naive Bayes, k-nearest neighbor, artificial neural network, and convolutional neural network	Obtained high performance, with accuracy levels ranging from 95.75% to 99.53%	Absence of a standardized medical test for ASD
5	M. Wang and N. Yang, 2023 [34,35]	Spatial–temporal Transformer; multiple-instance learning; graph convolutional networks	Potential tool could predict child's training state during therapy	Framework is not suitable to be deployed with a single-camera system
6	S. Zahan et al., 2023 [37]	Graph convolutional networks and Vision Transformer	Model predicts ADOS for children with ASD, having high correlation with the true ADOS	Relies on camera with Kinect v2 to capture human skeleton
7	Varun G. et al., 2023 [39]	Spatiotemporal Transformer; ResNet-34 deep learning model; R-convolutional network with ResNet-50	Results indicate acceptable performance for activity comprehension	Did not focus on acting as a stand-alone model that could interact with children

Human pose estimation is a critical factor for the proposed model. It has been integrated into several ASD studies: Kojovic N. et al. have integrated human pose estimation into ASD prediction [40]. They used OpenPose technology to extract a child skeleton with a total of 18 key points and deployed an integration of a CNN network using VGG16 in conjunction with LSTM networks. They recorded about 68 videos for children while playing with their parents and their framework was able to predict ASD children with 80.9% accuracy. Song C. et al. proposed an interesting method for automatic name detection, where ASD children usually have a lack of response to their name in early screening for ASD [41]. They collected a dataset named *Response to Name* from 30 children. They focused on face detection and head pose estimation in conjunction with computer vision techniques. Sternum J. et al. presented a survey for the applications of pose estimation in various human health aspects [42]. One of these domains is the usage of human pose in clinical diagnosis for children with neurodevelopmental and movement-based disorders. One part of this study were cerebral palsy diagnosis using magnetic resonance imaging and diagnosing autistic children by monitoring their gait patterns during their walking activity. Prakash V. G. et al. integrated three deep neural network models to learn activity recognition and estimate joint attention based on pose estimation for head and hands, and recognize emotion and facial expressions for ASD children [39]. Based on their collected 300-video dataset, their models achieved about 72%, 95%, and 95%, respectively. Vallee L.

N. et al. developed an imitation game for autistic children based on OpenPose to extract key points [43]. The Gaussian mixture model is used to compare the child-imitated video with 14 key points in the target video. Although previous studies have focused on various scenarios of children doing relatively complex tasks, they have not targeted children with autism while conducting a set of primitive imitation tasks.

3. The Techniques Employed in Implementing IMITASD

This section presents gross motor imitation, a basis behind autism spectrum disorder training. The second subsection outlines measures for time-series data that comply with the nature of the data processed in this research. Lastly, human pose estimation is briefly outlined where the MediaPipe library is highlighted.

3.1. Gross Motor Imitation

This research accentuates the critical significance of gross motor imitation abilities in the developmental trajectory of children diagnosed with ASD and intellectual disabilities. Emphasizing the pivotal role of these skills in daily functioning, social integration, and physical coordination, the study specifically employs the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP) at level 1 [44,45]. Within this assessment framework, six distinct gross motor imitation tasks have been carefully selected for evaluation among participating children. These tasks encompass actions such as *"wave by hand, side by side"*, *"arm up"*, *"hands fold together"*, *"thumbs up"*, *"fold hands together over head"*, and *"arms up"*. These tasks serve as benchmarks to address the developmental progression of gross motor imitation abilities of children with ASD and intellectual disability, aiming to improve overall functional capabilities in everyday life, as shown in Figure 1.

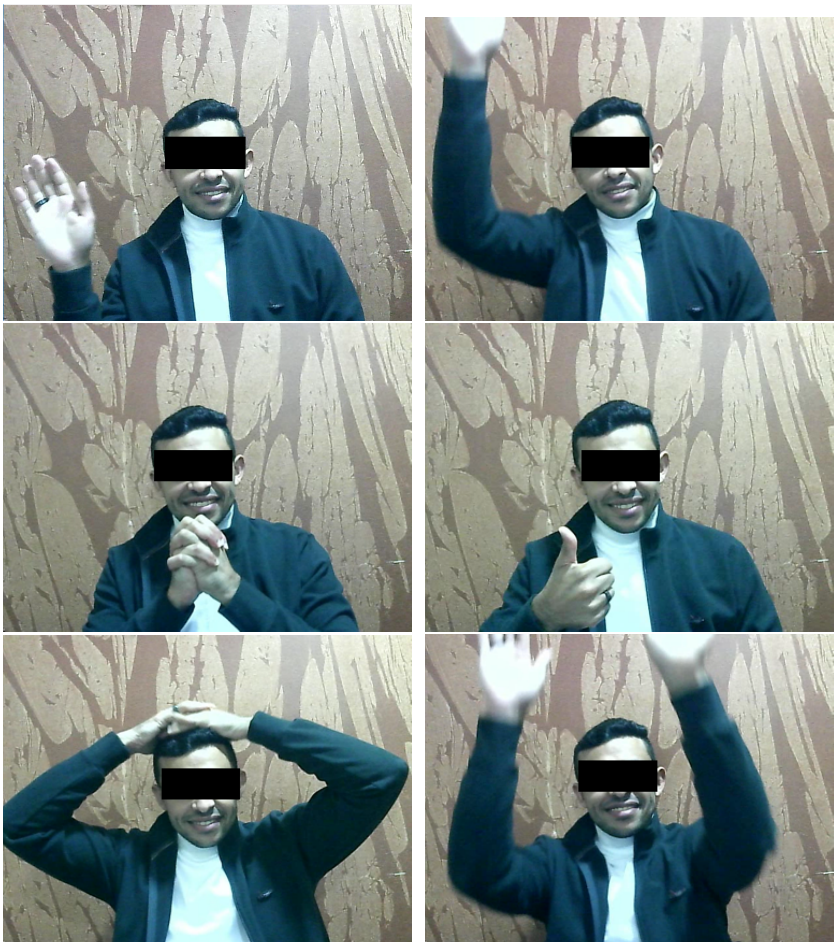


Figure 1. List of imitation tasks.

The six imitation tasks were chosen to assess fundamental gross motor imitation skills, which are crucial for physical coordination, social interaction, and overall functioning in children with ASD. Medical therapists validated these tasks to ensure their appropriateness for assessing children with moderate ASD, providing a reliable evaluation of gross motor imitation abilities. Therapist ratings were used as a benchmark to score imitation behaviors in the IMITASD. Each child's imitation video was evaluated by expert medical therapists, who provided scores based on their assessment of the child's performance using a four-level rating scale: poor, good, very good, and excellent. While these ratings serve as a critical validation mechanism, we acknowledge the potential for subjective bias inherent in human assessments. Therefore, multiple therapists assessed each video, and their combined ratings were averaged to provide a more reliable score.

3.2. Time-Series Measures

Time series contain a set of points that are taken over an equal space–time period. There are well-defined measures that can compute the distance or similarity between two time-series sequences. These include Euclidean distance, cosine similarity, Pearson correlation coefficient, and dynamic time warping (DTW) [46–48]. These time-series measures have been deployed as typical metrics in recent studies for finding the (dis)similarity among time-series data, especially for data augmentation [49–52]. A recent survey evaluated different metrics when assessing the quality of synthetic time series [53].

Euclidean distance [54–56] is the most common way to obtain the shortest distance between two sequences. It finds the straight line between two points for 2D or higher-dimensional space [57]. Equation (1) depicts the Euclidean distance formula, where n is number of dimensions [58]:

$$distance = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \tag{1}$$

Cosine similarity [59] measures the angle between two non-zero vectors, often used for high-dimensional data. It focuses on measuring the similarity orientation rather than its magnitude. The angle is computed through Equation (2), where θ is the angle between two vectors, A and B:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \tag{2}$$

The Pearson correlation quantifies the linear relationship between two sequences [57–60]. It measures the closeness between two sequences by finding the ratio between the covariance of two variables X and Y with the product of their standard deviations. The Pearson correlation coefficient, r, is computed by the following formula, Equation (3) [61]. It reflects a strong correlation between sequences X and Y when its value is higher than 0.8, while a low correlation is indicated through an r value less than 0.2.

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \tag{3}$$

Dynamic time warping is a powerful technique in time-series analysis, particularly when dealing with sequences that exhibit variations in timing [62]. It measures the similarity between two sequences by allowing for flexible alignments and accommodating temporal distortions between sequences X and Y. Let $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_M\}$ be sequences of lengths N and M, respectively. DTW involves the computation of a cost matrix C, where each element $C(i, j)$ represents the cumulative distance of aligning x_i with y_j [63]. This is formulated by the following recurrence Equation (4):

$$C(i, j) = d(x_i, y_i) + \min\{C(i-1, j), C(i, j-1), C(i-1, j-1)\} \tag{4}$$

where $d(x_i, y_j)$ is a local distance measure, often the Euclidean distance between x_j and y_j. The dynamic programming approach efficiently computes C, minimizing the overall alignment cost. The DTW distance, is then obtained as the minimum cumulative cost in the last row of the matrix according to Equation (5):

$$DTW = \min\{C(N, j)\} \tag{5}$$

This distance represents the minimum cost of aligning sequence X with sequence Y. To delve into the specifics, the optimal alignment path is determined by backtracking from $C(N, M)$ to $C(1, 1)$ based on certain rules: diagonal movement signifies a match, upward movement implies an insertion, and leftward movement corresponds to a deletion. This path delineates the optimal alignment between the sequences and effectively minimizes the cumulative cost.

3.3. Human Pose Estimation

Human pose estimation (HPE) [64] seeks to find all human parts relevant to the video. The output is a structure of edges that connects key points (such as ball–socket joint, hinge joint, gliding joint, etc.). HPE is vital to the computer vision discipline, where the rapid development of deep learning encourages researchers to develop efficient HPE open libraries. This explains the recent studies based on HPE in surveillance, medical supportive applications, and sport-related research. Recent surveys provide great sources for demonstrating the basic concepts behind HPE [65]. A recent study presented a taxonomy for key point extraction [65]. It classifies these libraries into top-down, and bottom-up approaches. The former locates a person in different frames in the video, and then, seeks to estimate the locations for key points, while the latter applies the opposite approach. CNN and its related models (R-CNN and fast R-CNN) are deployed as two-stage detectors [65,66].

A comparative survey has compared four state-of-the-art HPE libraries. These libraries are OpenPose, PoseNet, MoveNet, and MediaPipe [67]. These libraries have been deployed in several medical assistance studies. PoseNet and MoveNet extract 17 key points while OpenPose, and MediaPipe extract 135 and 33 points, respectively. According to the study results by Jen-Li C. and Meng-Chew L., on a benchmark with different action datasets, MediaPipe shows a superior percentage of detected joints (PDJ), 71.4%, compared to other libraries [67]. Self-occlusion and inaccurate camera positions are usually the main factors that decrease the accuracy of PDJ for these HPE libraries.

Google's MediaPipe excels in intricately tracing the positions of hands, facial landmarks, and the overall body pose, as shown in Figure 2 [68]. The MediaPipe module stands out for its robust and efficient estimation of both hand and body poses, allowing for monitoring of real-time movements [69]. Through hand tracking, a comprehensive set of 21 landmarks for each hand can be extracted, while body tracking allows the extraction of up to 33 landmarks [70]. MediaPipe is sensitive to pose identification failures. The child's chair position is fixed at a convenient position toward the display while recording the child's imitation. This increases the potential of successfully extracting the landmark. While recording the child's imitation, a fallback mechanism is implemented to skip any frame whose landmarks are not detected and consider the following frames instead. This is essential during occlusions that can happen during hand and arm movements.

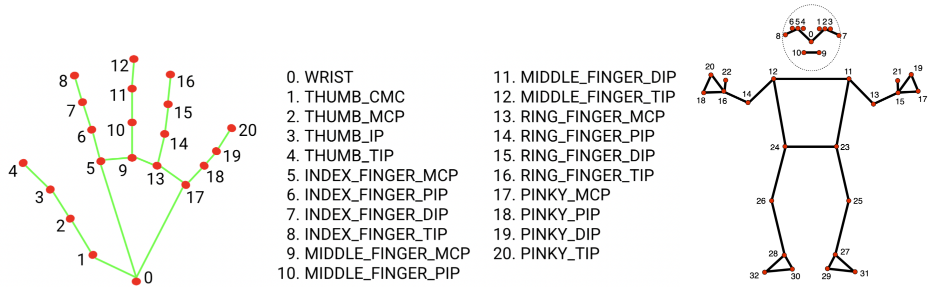


Figure 2. Landmarks from MediaPipe Hand and Body Pose Tracking module [69,70].

4. Methodology

This section describes the collected dataset. After that, the experimental setup is presented, and then, a graphical user interface is illustrated, that is designed specifically to facilitate recording the videos of the children.

4.1. Dataset Description

The imitation training in this study is provided by a PC attached to a webcam. To the best of the authors' knowledge, no public dataset is available to match the research requirements. Therefore, a set of autistic children's videos is recorded as test data to evaluate the performance of the proposed method. As mentioned earlier, six imitation movements are considered, as shown in Table 2. These movements were identified through several imitation lessons designed especially for training autistic children [44]. The movements are validated by a children's medical therapist, before conducting the video collection with children.

The proposed method assesses the video of a child by estimating the matching degree with the imitation lesson. Patients with high levels of autism may face challenges in engaging with the study experiment's setup. Their unique needs may require more specialized interventions tailored to their profiles. Conversely, individuals with lower levels of autism may already possess higher levels of skills, necessitating a different intervention approach that is not covered in this study. Therefore, the research has been directed towards children with a moderate degree of autism, specifically targeting those with autism scores ranging from 30 to 36 according to the Childhood Autism Rating Scale (CARS) [71].

The dataset originally had 302 videos, collected from 11 Egyptian children during two intervention sessions. Initially, data collection was planned to include 15 instead of 11 children. Three out of the fifteen autistic children suffered from attention-deficit/hyperactivity disorder (ADHD); so, they had difficulty sitting appropriately in front of the computer desk, while one child was tired and unable to continue when initiating the imitation session; thus, leaving eleven children. The children had ages ranging from 3 to 15 years; the number of boys was 9 (82%), with 2 girls (18%). During the data cleaning process, thirty-four videos were excluded due to anomalies such as the child's seat position in front of the camera being incorrect. As a result, the dataset was refined to consist of 268 videos. This dataset formed the basis for the evaluation and analysis of imitation behaviors in children with autism.

Table 2. List of imitation movements.

No.	Imitation Behavior	Number of Videos	Amount (%)
1	Wave by hand	49	18.3%
2	Arm up	43	16.0%
3	Hands fold together	55	20.5%
4	Thumbs up	41	15.3%
5	Fold hands together over head	38	14.2%
6	Arms up	42	15.7%

4.2. Experimental Setup

The videos of the children were recorded in a clinic where two adjacent rooms were used during the imitation sessions. Figure 3 illustrates the arrangement in the two rooms. The primary room, left figure, was equipped with a desk, a 21-inch LCD, and a 720 p webcam; mounted above the display, the LCD showed imitation lesson videos acted by the person of trust. This person may be a family member, teacher, or therapist, while the connected webcam records the child's actions during the mimicking process. Simultaneously, a technician (an engineering expert) was stationed in the secondary room, out of sight of the child. The technician ensured the smooth flow of playback and recording during imitation sessions, and additionally, initiated the lesson playback upon receiving notification from the system indicating that the child was fully attentive (this is discussed in the next section, attention module). As shown in Figure 3, two-room separation was implemented to minimize potential distractions for the child during the imitation sessions.

Children were accompanied by their parents, who were briefed on the video collection procedure by the medical therapist. Before conducting the experiments, relevant data on the child were stored in a log file containing the child's name, age, and degree of autism. Furthermore, the child's position and orientation were appropriately adjusted relative to the camera's location before initiating the imitation session.

During the imitation procedure, a medical therapist and a person of trust accompanied the child in the primary room. The medical therapist acted as an observer, ensuring the experiment progressed smoothly by preventing unexpected occurrences, such as the child suddenly leaving the room or interfering with the screen. The children were then guided through the imitation lessons to perform the specified movements.

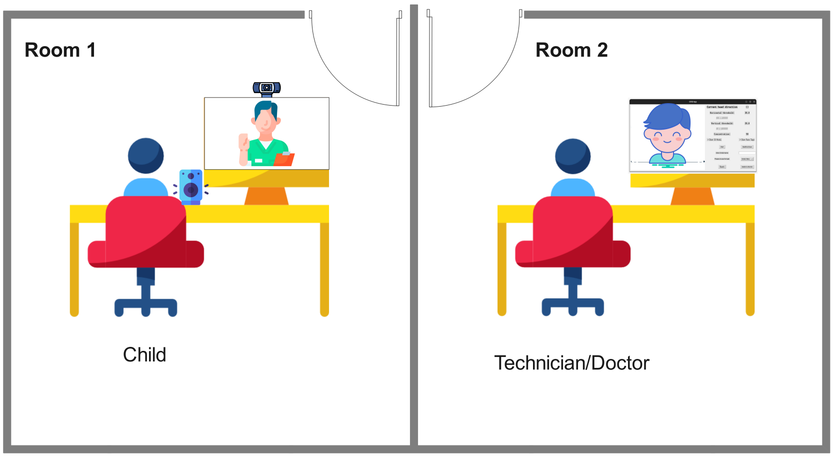
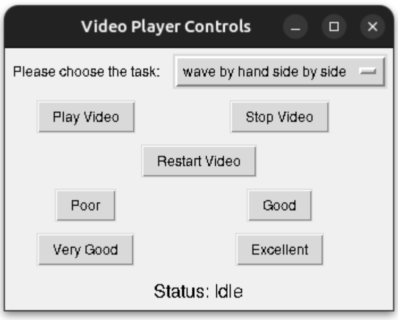


Figure 3. Room setting inside medical clinic.

4.3. Graphical User Interface Tool

There are two graphical user interfaces (GUIs), and both are designed to fulfill several functionalities needed by the research's requirements. Both are displayed on two monitors, connected to the technician's computer, which is located in room 2. The first GUI, Video Player Controls, is displayed on the primary monitor, located in room 2, as shown in Figure 4; while the second GUI, Child App, is displayed across two monitors, positioned in rooms 1 and 2, as depicted in Figure 5.

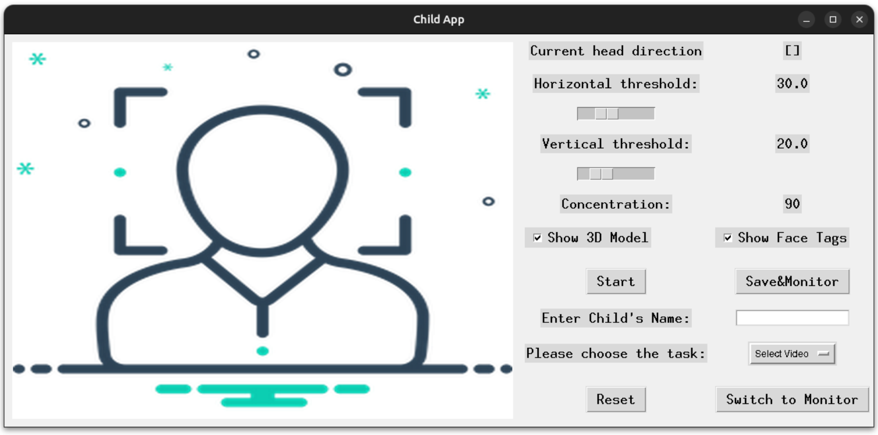


Please choose the task: wave by hand side by side

Status: Idle

Figure 4. GUI control available for the admin.

The Video Player Controls GUI provides the technician with the typical video controls. It allows for managing all operations related to the imitation lesson including playing, stopping, and restarting the lesson. Additionally, other GUI controls support the medical therapist in rating the video of the child after completing the imitation session. The score has 4 levels; poor, good, very good, and excellent. Finally, the label "status" shown at the bottom of Figure 4 clarifies the imitation lesson state. For instance, when it indicates "idle", the lesson has not been started yet.


Current head direction []

Horizontal threshold: 30.0

Vertical threshold: 20.0

Concentration: 90

☑ Show 3D Model

☑ Show Face Tags

Enter Child's Name:

Please choose the task: Select Video

Figure 5. GUI interface, where the left part (child preview) is visible on the child's screen.

The Child App GUI (left pane area) is displayed on the secondary monitor in room 1, where the child is sitting. It is the interface that shows the imitation lesson to the child. The GUI's right pane is displayed on the primary monitor in room 2, as it contains a set of controls. It allows the technician to enter the child's name and monitor the child's attention by displaying the child's estimated concentration. The latter is essential to support the technician in selecting a suitable time to play the imitation lesson, where the child's concentration ranges from 0 (no attention) to 100 (high attention level). A high concentration level indicates the trigger time to initiate the imitation lesson. Furthermore, the technician adjusts horizontal and vertical thresholds, which regulate the child's head movement direction and eye pose. The attention module uses these adjustments to find the child's gaze direction, which is discussed in the following subsection.

The Child App GUI includes other features such as "show face tags", which displays the detected facial landmarks. Additionally, the GUI has "show 3D model" that visualizes the child's perspective. The remaining controls, the "start" and "save/monitor" buttons, allow the technician to activate the camera and initiate recording of the video of the child or save the video of the child, respectively.

4.4. Subjective Assessment by Psychiatric Doctors

The subjective assessments were conducted under the direct supervision of psychiatric doctors, who have extensive experience in working with autistic children. Several pre-experiment meetings were held to standardize the criteria for evaluating imitation performance. During these discussions, they reached a consensus on how to categorize and distinguish the four levels of performance: "bad", "good", "very good", and "exceptional". It was determined that hand positioning and fine arm movements would be the primary factors in assessing quality, contrary to factors such as speed and facial expressions, that would not be considered in the evaluation. Multiple doctors independently assessed each child's performance, and the final grade was determined by averaging their scores to ensure evaluation consistency.

4.5. Parental Engagement and Bias Control

Parents were informed about their role during the video gathering sessions, and strict guidelines were implemented to control their engagement. Parents were instructed to observe without interacting with or guiding the child and were seated away from the immediate task area. A trained observer monitored all sessions to ensure compliance. Additionally, a behavioral calibration phase was included to help the child acclimate to the task environment, ensuring natural and unbiased responses.

4.6. Hardware Requirements

The IMITASD tool is based on deploying Yin Guobing's Facial Landmark Detector and fast dynamic time warping, which is designed to run efficiently on mid-range computers with standard webcams, ensuring accessibility in both clinical and home environments. The minimum technical requirements for consistent performance include an Intel i5 (4th generation or newer) or equivalent processor, 8 GB of RAM, and a 720 p or higher resolution webcam. While systems with higher-end configurations may deliver faster processing times, the accuracy of the tool's imitation assessments remains unaffected, as long as the minimum hardware specifications are met.

5. System Architecture Overview

The proposed method has three steps: imitation lesson preparation, checking the child's attention, and imitation assessment. The first step deploys activities related to imitation video lessons, extracting landmarks from the ground truth videos. The second step employs various metrics to gauge attention, encompassing gaze direction, head movements, and estimating concentration levels. This stage provides automatic insights into the child's engagement during the learning process, which is crucial to ensure the child's visual attention towards the LCD before displaying the imitation video. The final step measures the alignment between the recorded video of the child and the ground truth video. This comparison is essential for assessing how closely the child's actions replicate the behaviors demonstrated in the ground truth videos.

Imitation lesson preparation is the first step in the proposed method. Once the child is set in front of the monitor, an imitation lesson is selected, where the corresponding features related to the selected video are fetched to be used later during stage three, as detailed in Algorithm 1. These features are the head and hand landmarks extracted by MediaPipe that are tracked across the lesson's frames.

Step 2—Checking the child's attention: This focuses on measuring the child's attention by analyzing the facial features and head movements. Attention is a complex cognitive process that involves various visual and spatial cues. Attention is a complex cognitive state that is simplified in this study by checking whether the child's head is directed toward the display. Although it is a high level of approximation, it allows the development of a model that approaches real time. Attention estimation is linked to frequency and magnitude. Both combinations provide an attention estimation. The frequency reflects the child's head movements while magnitude extends head movements by measuring head movement distance over time. The weighted integration provides a means to prioritize movement frequency over magnitude as the head movement frequency is correlated more with the child's engagement during imitation sessions. In future studies, we will consider more robust measures such as eye and gaze tracking to estimate the child's attention during imitation sessions. The following procedure summarizes the steps to quantify the attention levels, as shown in Figure 6.

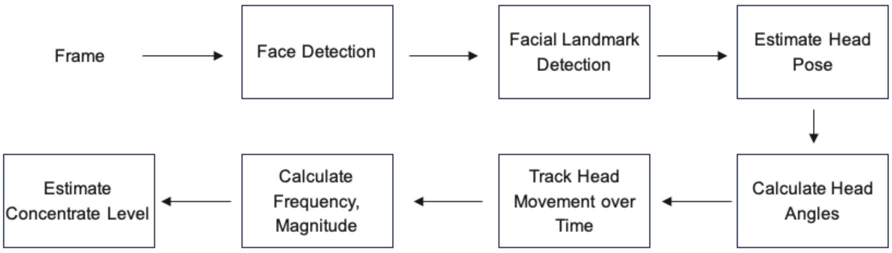

Frame
Face Detection
Facial Landmark Detection
Estimate Head Pose
Calculate Head Angles
Track Head Movement over Time
Calculate Frequency, Magnitude
Estimate Concentrate Level

Figure 6. Child attention module.

Algorithm 1 Imitation Lesson Preparation

```
Input: Set of lesson videos $V = \{v_1, v_2, \ldots, v_n\}$, Child $C$
Output: Corresponding features $F$ related to selected video
Place the child $C$ in front of the LCD
A lesson video $v_i$ is selected from the set $V$ ($v_i \in V$)
Fetch the corresponding features $F_i$ related to the selected video $v_i$
    $F_i = \text{fetch\_features}(v_i)$
Store these features $F_i$ to be used in stage three for similarity assessment
```

- Face detection: A frontal face detector is deployed to identify and locate faces within an input image.
- Facial landmark detection: This is responsible for extracting spatial information about key facial points through applying Yin Guobing's Facial Landmark Detector [72,73]. It goes beyond simple face detection, capturing the nuances of facial expressions and features. The Facial Landmark Detector model operates on square boxes of size 128 × 128, each containing a face. Upon analysis, it returns 68 facial landmarks, with the resulting landmarks serving as a critical input for subsequent phases. Yin Guobing's Facial Landmark Detector was employed for tracking children's attention based on facial orientation and head pose estimation. While this detector is accurate under normal conditions, autistic children may exhibit atypical facial expressions and frequent gaze aversion, which could impact the precision of attention tracking. To mitigate these limitations, the IMITASD tool includes a calibration step to ensure the child's face is properly aligned with the camera, reducing the likelihood of tracking errors during the task.
- Head pose estimation: This obtains the head's pose relative to the camera. Both the rotation and translation vectors are computed to provide a robust representation of the head's spatial orientation and position.
- Head pose angle calculation: This calculates specific head angles: yaw, pitch, and roll. These angles depict the head's orientation in three-dimensional space, capturing horizontal and vertical rotations and tilt motion. These angles are crucial for understanding the user's head position. Thirty-four videos were excluded due to anomalies such as incorrect child's seat position in front of the camera.
- Head movement tracking: To assess continual attention levels, the script simulates the head angle over time. The aim is to capture the changes in head orientation over time. The frequency and magnitude of head movements are quantified based on the differences between consecutive head angles.
- Attention measurement: This uses facial analysis and head movements to measure a child's attention. The frequency of head movements represents changes over time, while the magnitude of head movements indicates the angular displacement. These metrics feed into the computation of the concentration level, a weighted combination of frequency and magnitude, as detailed in Algorithm 2.

Step 3—Imitation assessment: This step deploys a low-complexity method to measure the similarity between the child's behavior and the ground truth video (imitation lesson), as shown in Algorithm 3. As soon as the child looks toward the display screen, the procedures depicted in Figure 7 are conducted. The following elaborates on the main process in detail.

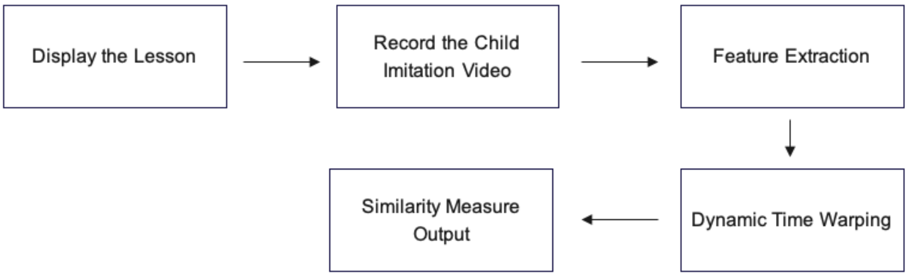


Figure 7. Imitation Assessment Block module.

Algorithm 2 Checking Child Attention

```
**Input:** Video stream $V_s$, Child $C$
**Output:** Attention level $A$
**Face Detection:**
Deploy Frontal Face Detector to identify and locate faces within input image $I$
$D = \text{FaceDetector}(I)$
**return** $\{d_1, d_2, \ldots, d_m\}$ where $d_i$ is a detected face
**Facial Landmark Detection:**
Extract detailed spatial information about key facial points using Yin Guobing's Facial Landmark Detector
$L = \text{FacialLandmarkDetector}(d_i)$
**return** $\{l_1, l_2, \ldots, l_{68}\}$ where $l_i$ are the 68 facial landmarks
**Head Pose Estimation:**
Estimate head's pose to the camera by computing rotation and translation vectors
$(R, t) = \text{HeadPoseEstimator}(L)$
**Head Pose Angles Calculation:**
Calculate specific head angles: yaw ($\theta_y$), pitch ($\theta_p$), and roll ($\theta_r$)
$\theta_y, \theta_p, \theta_r = \text{HeadPoseAngles}(R, t)$
**return** $\theta_y, \theta_p, \theta_r$
**Head Movements Tracking:**
Track head movements over time to assess attention levels
Compute differences between consecutive head angles $\Delta\theta_t = \theta_t - \theta_{t-1}$
Calculate frequency and magnitude of head movements $F, M$
**Attention Measurement:**
Measure attention using facial analysis and head movements
Calculate concentration level $C$ as a weighted combination of frequency and magnitude
$A = w_f \cdot F + w_m \cdot M$
**return** $A$
```

- Display the lesson video: After selecting the imitation lesson and ensuring the appropriate attention for the child toward the display screen (steps 1 and 2), the selected video is played through the LCD in front of the child.
- Record the child's movement: While playing the lesson video, the child instantly begins imitating the lesson. Therefore, the recording is started once the child begins the imitation. It stops the recording once the child finishes performing the imitation.
- Feature extraction: Features for imitation lessons and the videos of the child are obtained. The features from the former are extracted offline while their data are stored in a pickle data format. For the latter videos, the features are extracted while assessing the imitation behavior of autistic children. Features are based on pose and hand landmarks as predicted by MediaPipe. The connections between MediaPipe landmarks, articulated as pairs of indices, are transformed to vectors. These vectors serve as the foundation for subsequent angle calculations. The angles, meticulously computed using the dot product and vector norms, collectively contribute to the feature vectors. These vectors represent the trace and hand, where the former corresponds to the

child's arm and head positions. The latter supports hand tracking, where fine details are considered during the child's imitation. Both vectors are extracted from pose and hand landmarks, respectively. Note that color conversion is necessary as distinct color representation exists between OpenCV (BGR) and MediaPipe (RGB). After that, the system leverages the angles between hand parts, referred to as connections, incorporating all 21 connections intrinsic to MediaPipe's Hand Model. Given v_i is the video of the child, the detailed processes are depicted in Figure 8.

Algorithm 3 Imitation Assessment

```
Input: Child video $v_i$, Ground truth video $v_g$
Output: Similarity measure $S$
Display the Lesson Video:
Play the selected lesson video $v_g$.
Record the Child Movement:
Begin recording the child's movement as soon as imitation starts.
Stop recording when the child finishes the imitation.
Feature Extraction:
Extract features for both the ground truth lesson video $v_g$ (prerecorded) and the child's video $v_i$.
for each frame in $v_i$ do
    Extract pose and hand landmarks using MediaPipe.
    Convert color format from BGR to RGB if necessary.
    Initialize feature vectors for trace and hand vectors.
    for each frame $f_j$ in $v_i$ do
        Extract pose landmarks $\{p_{13}, p_{17}, p_{14}, p_{18}\}$ for trace vector.
        Extract all 21 hand landmarks for hand vector.
        Normalize the trace data by the maximum $x$ and $y$ coordinates.
        Update and store the trace and hand vectors for further processing.
    end for
end for
Dynamic Time Warping (DTW):
Apply DTW for the child's trace vector and the ground truth trace vector.
Apply DTW for the child's hand vector and the ground truth hand vector.
Compute the average distance $D$:
    $D = \frac{D_{\text{trace}} + D_{\text{hand}}}{2}$
Similarity Measure Output:
Map the distance $D$ to a similarity measure $S$ in the range of 0 to 10.
The similarity measure $S$ is defined as
    $S = 10\left(1 - \frac{D}{D_{\max}}\right)$
where $D_{\max}$ is the maximum possible distance.
return Similarity measure $S$
```

- The feature vectors, hand and trace vectors, are initialized for the given video, where the first frame that belongs to the video of the child is prepared for processing.
- It applies an iterative process over v_i's frames, where pose and hand landmarks for each frame are predicted. These landmarks are used to obtain trace and hand vectors for the current frame. These vectors are appended to the corresponding vectors representing the video of the child. Once, the video's vectors are updated, the next frame is fetched to be processed. The hand and pose landmarks predicted by MediaPipe are a set of 3D points, as depicted in Figure 2, where each point is

characterized by an (x, y, z) coordinate. For each frame, the trace vector focuses on four points extracted from the pose landmarks, they are points number 13 and 17 for the child's left arm and points 14 and 18 for the child's right arm. The hand vector uses all 21 points for each hand. Based on pose landmarks, the extraction of trace_left and trace_right unfolds as a process governed by precision and meticulousness. The initiation of these variables as lists of coordinate points paves the way for detailed scrutiny of detected landmarks.

- The normalization procedure focuses on normalizing the trace data. Identifying both maximum x and y coordinates for the trace and reference sets the stage for normalization. This process normalizes the trace coordinates through division by the respective maximum values, thereby laying the groundwork for meaningful distance calculations between traces.
- Saving trace and hand vectors: Both vectors are stored for further processing.

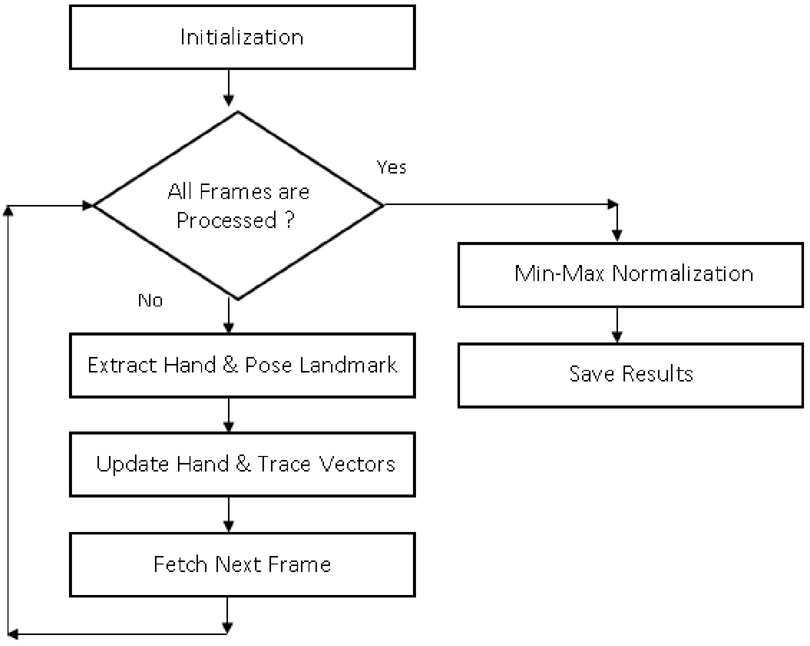

Initialization
All Frames are Processed ?
Yes
No
Min-Max Normalization
Save Results
Extract Hand & Pose Landmark
Update Hand & Trace Vectors
Fetch Next Frame

Figure 8. Feature extraction flowchart.

- Dynamic time warping algorithm (DTW): This measures the distance between features extracted from both the imitation lesson and the video of the child. Based on the videos' trace and hand vectors, DTW calculates the distances between these vectors. The output from both distances is averaged to obtain the final distance for the child's behavior to the given lesson. There are challenges when dealing with autistic children. The spectrum of variability when handling children's videos is large. They tend to begin the imitation process instantly when they begin watching the lessons and the children's imitation speed is varied. Here, IMITASD relies on the attention module and the DTW features. The former estimates the child's focus, therefore displaying the lesson and recording the child's imitation at convenient times. It supports DTW for better similarity estimation between the lesson and the child's imitation videos. Furthermore, the proposed model deploys fast DTW that features fast computation through processing the given inputs using their down-sampled sets to accelerate measuring the similarity between the given sets. DTW by default supports temporal alignment, as it can measure the similarity between sets with non-equal length.
- Similarity measure output: The resulting distance is mapped into a similarity measure in the range of 0 to 10. Therefore, the assessment module output does not require further processing. It should output 10 when the child's behavior matches the given

imitation lesson, while it obtains zero when the child's imitation does not match the lesson video.

6. Results and Discussion

Considering the nature of providing the imitation training program through a PC instead of physical training in the clinic, it is vital to predict the child's performance accurately and in a reasonable time. This will enable the development of a training program so that the autistic child interacts with a full imitation session provided by a mid-specification computer with a webcam. Therefore, this section presents an evaluation of IMITASD and answers the following questions.

- Is the proposed method suitable for scoring the children's performance, given the six imitation training lessons?
- What is the IMITASD performance when using different time-series measures?
- How long does the proposed method take to rate the child's performance?
- What are the limitations of IMITASD?

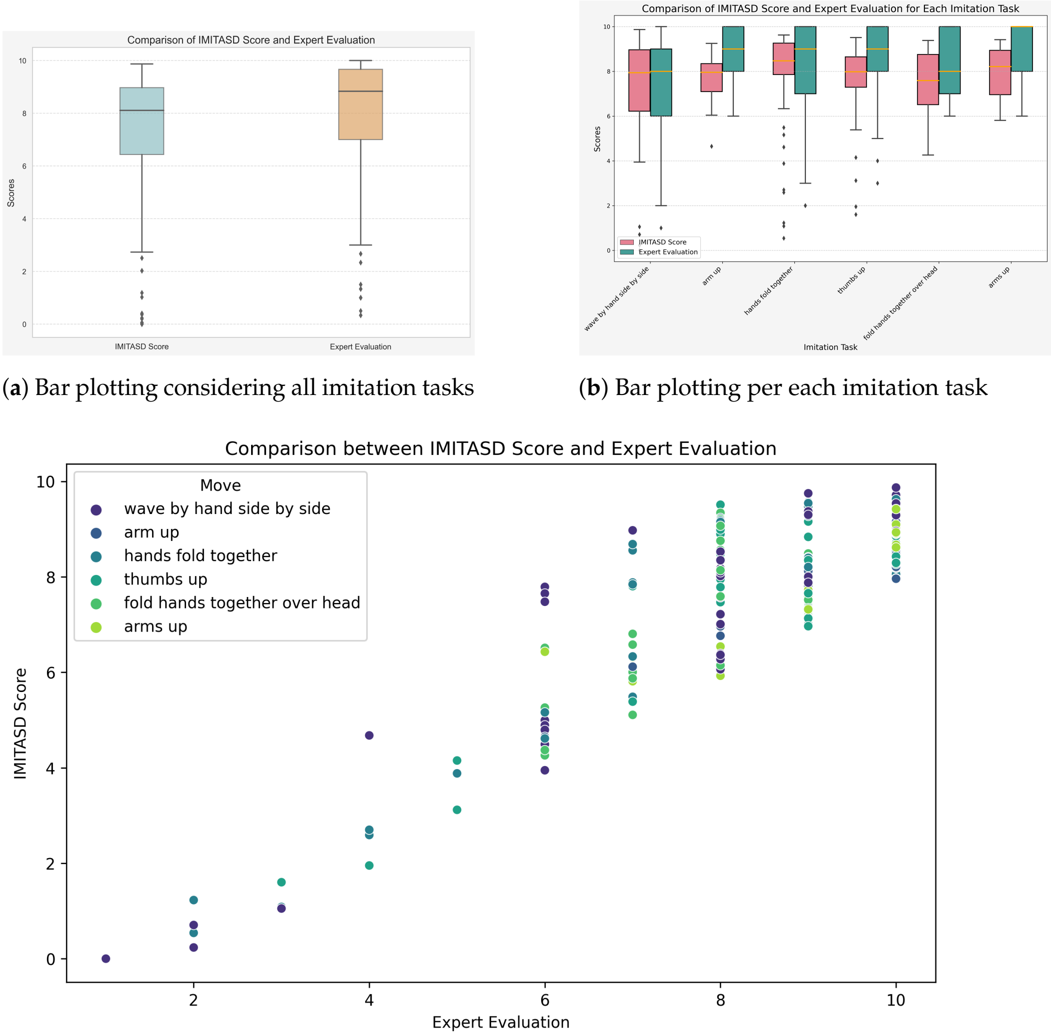
Regarding the first research question, the following Figure 9a–c reveal the answer for how the IMITASD prediction is relevant to therapist scoring. Figure 9a demonstrates how the IMITASD provides on average a close estimation of the autistic child's imitation video. It is on average 1 point lower than the medical therapist's assessment. Figure 9b,c further illustrate this estimation across all imitation tasks. The proposed assessment method provides close approximation with the therapist scoring, where imitation movements such as wave by hand and folding hand over head have very close matches, while movements such as single and both arms up are not perfectly aligned. Therefore, the IMITASD is an appropriate method for scoring the child's video for the given imitation lessons.

A comparative study of the proposed method using various time-series measures is conducted for the second research question. The children's videos have been used to measure the IMITASD prediction using the measures mentioned in Section 3.2. These measures are Euclidean distance, cosine similarity, Pearson Correlation, and dynamic time warping. Figure 10 shows the average scoring of these dis(similarity) measures for the collected dataset. Additionally, Figure 11 delves into a more detailed presentation, showing the scores for each imitation task individually. Together, these figures present a comprehensive overview of the comparative performance of the techniques across the entire spectrum of imitation tasks.

In Table 3, the correlation analysis across different metrics for specific imitation tasks reveals distinctive patterns in the model's performance. Notably, the IMITASD score computed based on DTW consistently exhibits high positive correlations, showcasing its proficiency in capturing temporal dynamics during imitation tasks. Particularly, there are exceptional correlations for imitation tasks like "thumbs up" (0.9912), and "hands fold together" (0.9689). In contrast, traditional metrics such as Euclidean distance, cosine similarity, and Pearson correlation obtain varied correlations across these imitation tasks. Upon scrutinizing the results, it is evident that DTW exhibits exceptional efficacy in capturing the children's behavior patterns.

Table 3. Comparative assessment scores for IMITASD using DTW with Pearson correlation, Euclidean distance, and cosine similarity metrics across imitation tasks using correlation coefficient.

	Wave by Hand	Arm Up	Hands Fold Together	Thumbs Up	Fold Hands Together	Arms Up	Overall Correlation
Euclidean distance	0.01	0.55	0.06	−0.05	0.22	−0.20	−0.04
Cosine similarity	0.09	−0.45	−0.08	−0.14	−0.03	−0.11	−0.10
Pearson correlation	−0.15	0.17	−0.17	0.09	0.01	−0.02	0.05
IMITASD score (DTW)	0.94	0.64	0.97	0.99	0.87	0.86	0.94



(**a**) Bar plotting considering all imitation tasks

(**b**) Bar plotting per each imitation task

(**c**) Scatter plotting per each imitation task

Figure 9. Comparison between IMITASD score and medical evaluation.

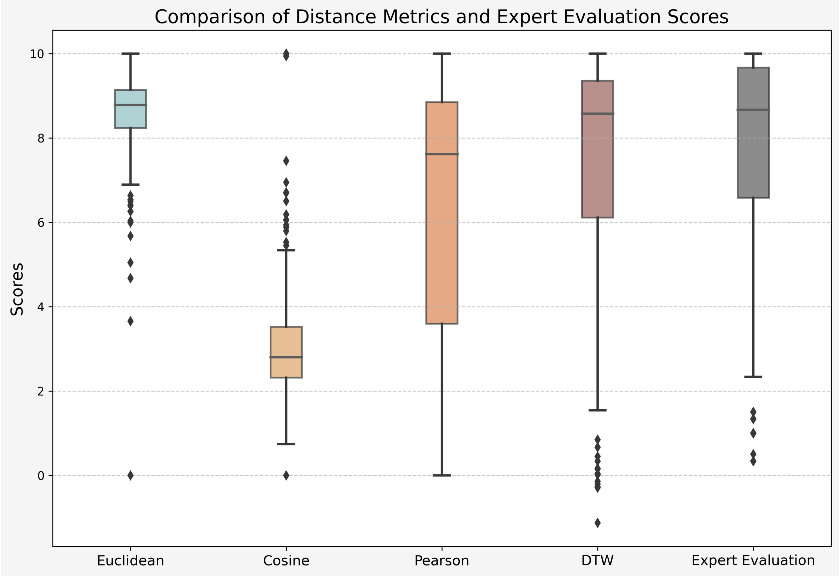


Figure 10. Detailed comparison of distance metrics and expert evaluation scores.

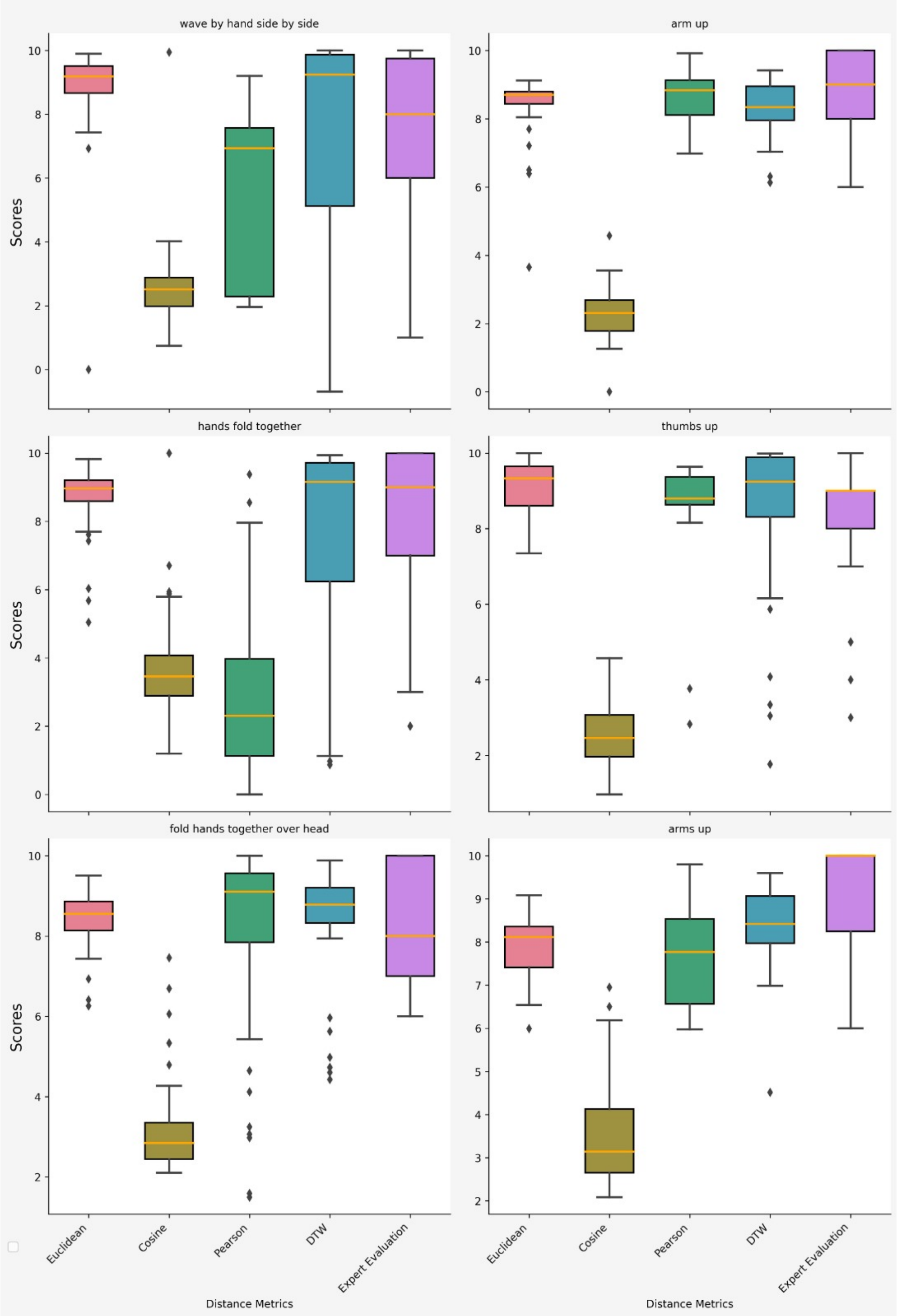


Figure 11. Comparison of distance metrics and expert evaluation scores for each imitation task.

To investigate the required average time for IMITASD to score a video of a child, the running time is measured across the children's videos considering all imitation lessons. Based on the measurements, a single frame is on average processed in 0.1 s, while it takes on average less than 3 s to rate a single video. Considering the lessons *"hands fold together"*, *"fold hands together over head"*, *"arm up"*, and *"arms up"* have 20, 18, 17, and 13 frames, their running times are 2.24, 1.62, 1.49, and 1.6 s, respectively. Figure 12 shows the running time to score a video containing 25 frames. Since the IMITASD has a relatively fast response when scoring a video, it could be integrated into an efficient training program suitable for imitation lessons.

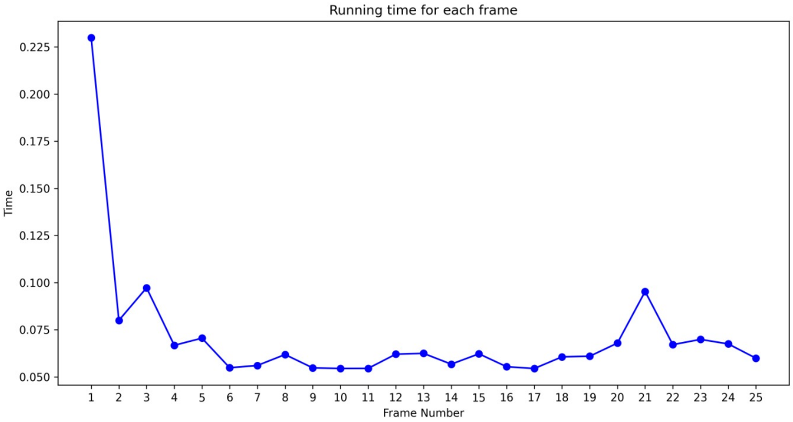


Figure 12. Running time to process a video segment.

Regarding the fourth research question, which investigates the limitations of IMITASD, several observations have been obtained from conducting the experiments. The main constraint of IMITASD is that is uses a simple experimental setup (relying on a single camera). The MediaPipe, in some scenarios, is unable to extract the key points for the child's body. This is the result of not including multiple cameras or using an additional Kinect camera in the experiment setup. This limitation is particularly evident in scenarios involving occlusion and complex poses. In these scenarios, MediaPipe may struggle to accurately detect and track skeletal key points, leading to incorrect representations of a child's body movements. About 11% of the children's videos belong to this type of limitation; it occurs frequently for three participants (numbers 3, 7, and 10). The remaining children's videos barely suffered from this limitation, as depicted in Figure 13. Looking closely at the three children's videos, there are 28 videos out of 33 videos suffer from this limitation. The common imitation tasks affected by this limitations were "arm up" and "arms up" (both account for 20 videos among all the 33 videos), as depicted in Figure 14. The overall statistics for undetected movement is depicted in Figure 15, where "arm up", and "arms up" account for about 66.7% of the entire undetected movements.

Based on investigating the four research questions mentioned earlier in this section, the significant remarks are summarized as follows:

- The proposed method rates the children's imitation videos very similar to the therapist's score. The closest match occurred in *"wave by hand"*, while the worst match was for the *"arms up"* task.
- The IMITASD results using different time-series measures highlight the superior performance of the IMITASD score based on dynamic time warping compared to Euclidean distance, cosine similarity, and Pearson correlation. Both the tasks *"thumbs up"*, and *"hands fold together"* attain high correlations of 0.9912 and 0.9689, respectively. These results confirm the IMITASD score's precision in aligning temporal dynamics during imitation, outshining traditional metrics.
- The proposed method's running time is on average less than three seconds to score a single video of a child. Therefore, the proposed method could be embedded in a training program that should be fast enough during the child imitation session.
- IMITASD faces challenges due to relying on a single camera. The landmarks based on a single camera are sensitive to occlusion. This affects IMITASD's capability to estimate the similarity accurately for the child imitation video.
- It is important to note that during the data cleaning, thirty-four videos were excluded due to anomalies. The exclusion of these videos ensured the integrity of the remaining dataset and prevented the introduction of bias due to sub-optimal video quality. To address challenges related to improper seat posture in future studies, measures such as seat markers, adjustable seating, or real-time posture feedback systems could be

implemented to ensure that participants are consistently positioned correctly in front of the camera. These measures would help reduce the number of unusable videos and enhance the overall quality of data collection.

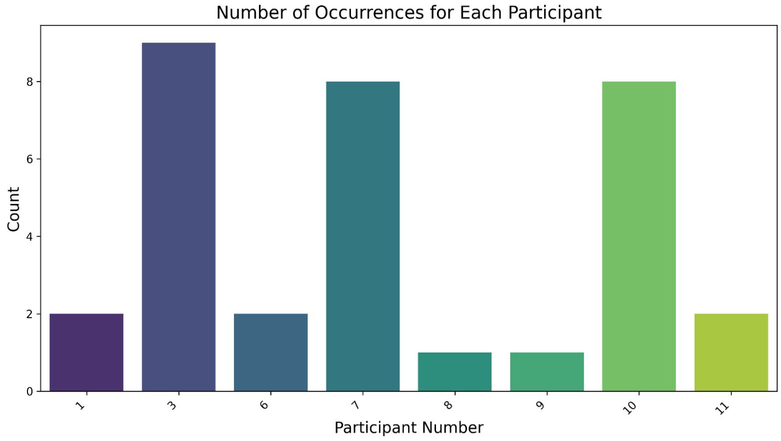


Figure 13. Number of videos that were unable to be processed by MediaPipe grouped by participant.

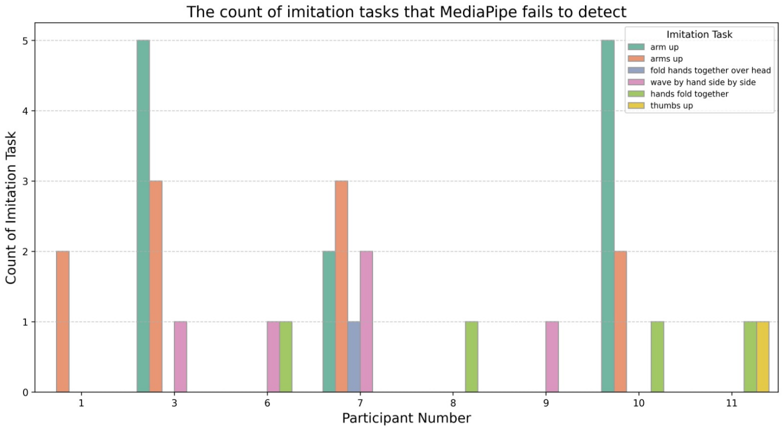


Figure 14. Number of videos that were unable to be processed by MediaPipe grouped by participant and task.

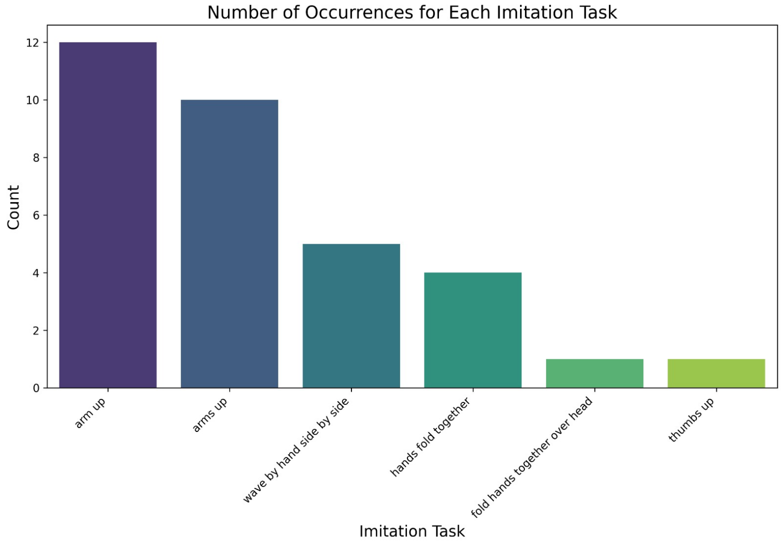


Figure 15. Number of videos that were unable to be processed by MediaPipe grouped by task.

7. Conclusions

In conclusion, this research proposes IMITASD, a novel tool designed for assessing children with autism spectrum disorder (ASD) behaviors using a set of imitation interventions. This is part of applied behavior analysis (ABA) therapy. The dataset used is collected from Egyptian autistic children. A graphical user interface is designed specifically to enable easy data collection. IMITASD addressed the challenges identified in previous studies, recognizing the limitations associated with high-end cameras, Kinect devices, and specialized setups. The proposed methodology aims to overcome these challenges, aiming for real-world applicability by employing available equipment based on a mid-specification PC with a webcam.

The experimental results demonstrate the tool's relevance to medical therapy assessments and its suitability across children with a mid-level degree of autism. Additionally, it provides an accurate score, closely aligned with therapist scoring. The dynamic time warping method used in IMITASD consistently outperforms traditional metrics such as Euclidean distance, cosine similarity, and Pearson correlation. IMITASD proves to be a practical solution for providing low-cost, fast, and high-validity assessments of imitation behaviors in children with ASD.

8. Future Work

Future works in this domain could involve extensions in various directions. First, more complex imitation interventions need to be examined. Secondly, other human pose estimation libraries could be investigated as potential alternatives to MediaPipa for extracting facial and body landmarks, such as OpenPose, PoseNet, and MoveNet. Third, research should refine the capabilities of machine learning and deep learning models in scoring the children's imitation videos efficiently. Fourth, IMITASD could be integrated with a blockchain-based platform to ensure data security and privacy. This would secure the children's video. Finally, the role of deep fakes on a child's behavior could be investigated, when a child watches lessons performed by persons of trust. This involves developing deep fake models within this framework to generate personalized scripts for patients, facilitating interactive therapeutic sessions. Fifth, future versions of the IMITASD tool may include additional imitation tasks to improve the comprehensiveness of the gross motor imitation assessment, allowing for a broader evaluation of motor abilities in children with ASD. Finally, to improve scalability in diverse clinical or therapeutic settings, the IMITASD system can be enhanced in several ways. Imitation lesson preparation can be automated through machine learning algorithms for personalized lesson selection. Multi-modal sensory inputs such as eye tracking and body posture analysis could integrate additional attentiveness parameters for a more comprehensive evaluation.

Author Contributions: K.M. and R.S. conceived the idea for the IMITASD method, designed the experimental framework, and analyzed the results. F.A.M. and H.S. contributed to the design and implementation of the attention and assessment modules, contributed to data analysis, provided insights into video processing techniques, and reviewed the drafting. M.K., Y.A.M. and A.B. contributed to the technical aspects of real-time video analysis, assisted in data interpretation, and provided critical feedback on the methodology. A.E., M.H. and R.S. participated in the design and execution of the study, collected and analyzed data, provided medical insights and expertise in child psychology, and contributed to the critical review and editing of the manuscript. S.E.S. contributed to the methodology, funding acquisition, and validation of the research. All authors contributed to the writing and revision of the manuscript, provided critical feedback on the methodology, and approved the final version of the paper for submission. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Information Technology Industry Development Agency (ITIDA)-Egypt, Information Technology Academia Collaboration (ITAC) program of collaborative funded projects under the category of Preliminary Research Projects (PRP), Grant PRP2022.R32.29. This work was also supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under Project Grant KFU242006.

Data Availability Statement: Videos for this study are not publicly available to preserve individuals' privacy, but the data analysis and findings are available upon request.

Acknowledgments: We extend our heartfelt thanks to Psychologist Somaya Hamdy for her invaluable support in implementing the IMITASD tool. Her expertise and dedication to the well-being of autistic children have greatly influenced our research. Her insights helped refine our assessment approach, improving our understanding of the children's needs. Somaya's contributions have been crucial to the success of this work.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this work.

References

1. Magazine, A.P. Autism Statistics: Facts and Figures. 2024. Available online: https://www.autismparentingmagazine.com/autism-statistics/ (accessed on 5 October 2024).
2. American Academy of Pediatrics. CDC: Autism Rate Rises to 1 in 36 Children. 2024. Available online: https://publications.aap.org/aapnews/news/23904/CDC-Autism-rate-rises-to-1-in-36-children?autologincheck=redirected#/ (accessed on 5 October 2024).
3. Treetop, T. Autism Prevalence Statistics. 2024. Available online: https://www.thetreetop.com/statistics/autism-prevalence/ (accessed on 5 October 2024).
4. Gitimoghaddam, M.; Chichkine, N.; McArthur, L.; Sangha, S.S.; Symington, V. Applied behavior analysis in children and youth with autism spectrum disorders: A scoping review. *Perspect. Behav. Sci.* **2022**, *45*, 521–557. [CrossRef] [PubMed]
5. Silva, A.P.d.; Bezerra, I.M.P.; Antunes, T.P.C.; Cavalcanti, M.P.E.; Abreu, L.C.d. Applied behavioral analysis for the skill performance of children with autism spectrum disorder. *Front. Psychiatry* **2023**, *14*, 1093252. [CrossRef] [PubMed]
6. Maula, M.I.; Ammarullah, M.I.; Fadhila, H.N.; Afif, I.Y.; Hardian, H.; Jamari, J.; Winarni, T.I. Comfort evaluation and physiological effects/autonomic nervous system response of inflatable deep pressure vest in reducing anxiety. *Heliyon* **2024**, *10*, e36065. [CrossRef]
7. Maula, M.I.; Afif, I.Y.; Ammarullah, M.I.; Lamura, M.D.P.; Jamari, J.; Winarni, T.I. Assessing the calming effects of a self-regulated inflatable vest: An evaluation based on Visual Analogue Scale and Electroencephalogram. *Cogent Eng.* **2024**, *11*, 2313891. [CrossRef]
8. Husaini, F.A.; Maula, M.I.; Ammarullah, M.I.; Afif, I.Y.; Lamura, M.D.P.; Jamari, J.; Winarni, T.I. Control design of vibrotactile stimulation on weighted vest for deep pressure therapy. *Bali Med. J.* **2024**, *13*, 860–865. [CrossRef]
9. Nielsen, M. The social glue of cumulative culture and ritual behavior. *Child Dev. Perspect.* **2018**, *12*, 264–268. [CrossRef]
10. Bravo, A.; Schwartz, I. Teaching imitation to young children with autism spectrum disorder using discrete trial training and contingent imitation. *J. Dev. Phys. Disabil.* **2022**, *34*, 655–672. [CrossRef]
11. Halbur, M.; Preas, E.; Carroll, R.; Judkins, M.; Rey, C.; Crawford, M. A comparison of fixed and repetitive models to teach object imitation to children with autism. *J. Appl. Behav. Anal.* **2023**, *56*, 674–686. [CrossRef]
12. Posar, A.; Visconti, P. Autism spectrum disorder in 2023: A challenge still open. *Turk. Arch. Pediatr.* **2023**, *58*, 566.
13. Chiappini, M.; Dei, C.; Micheletti, E.; Biffi, E.; Storm, F.A. High-Functioning Autism and Virtual Reality Applications: A Scoping Review. *Appl. Sci.* **2024**, *14*, 3132. [CrossRef]
14. Liu, L.; Li, S.; Tian, L.; Yao, X.; Ling, Y.; Chen, J.; Wang, G.; Yang, Y. The Impact of Cues on Joint Attention in Children with Autism Spectrum Disorder: An Eye-Tracking Study in Virtual Games. *Behav. Sci.* **2024**, *14*, 871. [CrossRef] [PubMed]
15. Cano, S.; Díaz-Arancibia, J.; Arango-López, J.; Libreros, J.E.; García, M. Design path for a social robot for emotional communication for children with autism spectrum disorder (ASD). *Sensors* **2023**, *23*, 5291. [CrossRef] [PubMed]
16. López-Florit, L.; García-Cuesta, E.; Gracia-Expósito, L.; García-García, G.; Iandolo, G. Physiological Reactions in the Therapist and Turn-Taking during Online Psychotherapy with Children and Adolescents with Autism Spectrum Disorder. *Brain Sci.* **2021**, *11*, 586. [CrossRef] [PubMed]
17. Nunez, E.; Matsuda, S.; Hirokawa, M.; Yamamoto, J.; Suzuki, K. Effect of sensory feedback on turn-taking using paired devices for children with ASD. *Multimodal Technol. Interact.* **2018**, *2*, 61. [CrossRef]
18. Jameson, J. Autism and Imitation Skills Importance. 2020. Available online: https://jewelautismcentre.com/jewel_blog/autism-and-imitation-skills-importance/ (accessed on 17 October 2024).
19. Sandhu, G.; Kilburg, A.; Martin, A.; Pande, C.; Witschel, H.F.; Laurenzi, E.; Billing, E. A learning tracker using digital biomarkers for autistic preschoolers. In Proceedings of the Society 5.0, Integrating Digital World and Real World to Resolve Challenges in Business and Society, 2nd Conference, Hybrid (Online and Physical) at the FHNW University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland, 20–22 June 2022; EasyChair, pp. 219–230.
20. Al-Jubouri, A.A.; Ali, I.H.; Rajihy, Y. Generating 3D dataset of Gait and Full body movement of children with Autism spectrum disorders collected by Kinect v2 camera. *Compusoft* **2020**, *9*, 3791–3797.
21. Liu, X.; Zhao, W.; Qi, Q.; Luo, X. A Survey on Autism Care, Diagnosis, and Intervention Based on Mobile Apps: Focusing on Usability and Software Design. *Sensors* **2023**, *23*, 6260. [CrossRef]

22. Zhang, W.; Sun, Z.; Lv, D.; Zuo, Y.; Wang, H.; Zhang, R. A Time Series Prediction-Based Method for Rotating Machinery Detection and Severity Assessment. *Aerospace* **2024**, *11*, 537. [CrossRef]
23. Sun, S.; Gu, M.; Liu, T. Adaptive Sliding Window–Dynamic Time Warping-Based Fluctuation Series Prediction for the Capacity of Lithium-Ion Batteries. *Electronics* **2024**, *13*, 2501. [CrossRef]
24. Isa, I.G.T.; Ammarullah, M.I.; Efendi, A.; Nugroho, Y.S.; Nasrullah, H.; Sari, M.P. Constructing an elderly health monitoring system using fuzzy rules and Internet of Things. *AIP Adv.* **2024**, *14*, 055317. [CrossRef]
25. Sen, B.; Bhowmik, A.; Prakash, C.; Ammarullah, M.I. Prediction of specific cutting energy consumption in eco-benign lubricating environment for biomedical industry applications: Exploring efficacy of GEP, ANN, and RSM models. *AIP Adv.* **2024**, *14*, 085216. [CrossRef]
26. Kaur, G.; Kaur, J.; Sharma, A.; Jain, A.; Kumar, R.; Alsubih, M.; Islam, S.; Ammarullah, M.I. Techno-economic investigation and empowering rural resilience through bioengineering: A case study on self-sustainable village energy models. *Int. J. Low-Carbon Technol.* **2024**, *19*, 1275–1287.
27. Farooq, M.S.; Tehseen, R.; Sabir, M.; Atal, Z. Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *Sci. Rep.* **2023**, *13*, 9605. [CrossRef] [PubMed]
28. Awad, M.; Khanna, R.; Awad, M.; Khanna, R. Support vector machines for classification. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: Berkeley, CA, USA, 2015; pp. 39–66.
29. Panda, N.R. A review on logistic regression in medical research. *Natl. J. Community Med.* **2022**, *13*, 265–270. [CrossRef]
30. Raj, S.; Masood, S. Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Comput. Sci.* **2020**, *167*, 994–1004. [CrossRef]
31. Yang, F.J. An Implementation of Naive Bayes Classifier. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 12–14 December 2018; pp. 301–306. [CrossRef]
32. Anava, O.; Levy, K. k*-nearest neighbors: From global to local. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
33. Ayeni, J. Convolutional neural network (CNN): The architecture and applications. *Appl. J. Phys. Sci.* **2022**, *4*, 42–50. [CrossRef]
34. Wang, M.; Yang, N. OTA-NN: Observational therapy-assistance neural network for enhancing autism intervention quality. In Proceedings of the 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 8–11 January 2022; pp. 1–7.
35. Wang, M.; Yang, N. OBTAIN: Observational Therapy-Assistance Neural Network for Training State Recognition. *IEEE Access* **2023**, *11*, 31951–31961. [CrossRef]
36. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: A comprehensive review. *Comput. Soc. Netw.* **2019**, *6*, 1–23. [CrossRef]
37. Zahan, S.; Gilani, Z.; Hassan, G.M.; Mian, A. Human Gesture and Gait Analysis for Autism Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3327–3336.
38. Papaefstathiou, E. A Thorough Presentation of Autism Diagnostic Observation Schedule (ADOS-2). In *Interventions for Improving Adaptive Behaviors in Children With Autism Spectrum Disorders*; IGI Global: Hershey, PA, USA, 2022; pp. 21–38.
39. Prakash, V.G.; Kohli, M.; Kohli, S.; Prathosh, A.; Wadhera, T.; Das, D.; Panigrahi, D.; Kommu, J.V.S. Computer vision-based assessment of autistic children: Analyzing interactions, emotions, human pose, and life skills. *IEEE Access* **2023**, *11*, 47907–47929. [CrossRef]
40. Kojovic, N.; Natraj, S.; Mohanty, S.P.; Maillart, T.; Schaer, M. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Sci. Rep.* **2021**, *11*, 15069. [CrossRef]
41. Song, C.; Wang, S.; Chen, M.; Li, H.; Jia, F.; Zhao, Y. A multimodal discrimination method for the response to name behavior of autistic children based on human pose tracking and head pose estimation. *Displays* **2023**, *76*, 102360. [CrossRef]
42. Stenum, J.; Cherry-Allen, K.M.; Pyles, C.O.; Reetzke, R.D.; Vignos, M.F.; Roemmich, R.T. Applications of pose estimation in human health and performance across the lifespan. *Sensors* **2021**, *21*, 7315. [CrossRef] [PubMed]
43. Vallée, L.N.; Lohr, C.; Kanellos, I.; Asseu, O. Human Skeleton Detection, Modeling and Gesture Imitation Learning for a Social Purpose. *Engineering* **2020**, *12*, 90–98. [CrossRef]
44. Conti, D.; Trubia, G.; Buono, S.; Di Nuovo, S.; Di Nuovo, A. Evaluation of a robot-assisted therapy for children with autism and intellectual disability. In Proceedings of the Annual Conference Towards Autonomous Robotic Systems, Bristol, UK, 25–27 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 405–415.
45. Peterson, T.; Dodson, J.; Sherwin, R.; Strale, F., Jr.; Strale, F., Jr. Evaluating the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP) Scores Using Principal Components Analysis. *Cureus* **2024**, *16*, e66602. [CrossRef]
46. Bringmann, K.; Fischer, N.; van der Hoog, I.; Kipouridis, E.; Kociumaka, T.; Rotenberg, E. Dynamic Dynamic Time Warping. In Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). SIAM, Alexandria, VA, USA, 7–10 January 2024; pp. 208–242.
47. Wang, Z.; Ning, J.; Gao, M. Complex Network Model of Global Financial Time Series Based on Different Distance Functions. *Mathematics* **2024**, *12*, 2210. [CrossRef]
48. Kraprayoon, J.; Pham, A.; Tsai, T.J. Improving the Robustness of DTW to Global Time Warping Conditions in Audio Synchronization. *Appl. Sci.* **2024**, *14*, 1459. [CrossRef]
49. Wang, H.; Li, Y.; Jin, Y.; Zhao, S.; Han, C.; Song, L. Remaining Useful Life Prediction Method Enhanced by Data Augmentation and Similarity Fusion. *Vibration* **2024**, *7*, 560–581. [CrossRef]

50. Molina, M.; Tardón, L.J.; Barbancho, A.M.; De-Torres, I.; Barbancho, I. Enhanced average for event-related potential analysis using dynamic time warping. *Biomed. Signal Process. Control.* **2024**, *87*, 105531. [CrossRef]
51. Castellano Ontiveros, R.; Elgendi, M.; Menon, C. A machine learning-based approach for constructing remote photoplethysmogram signals from video cameras. *Commun. Med.* **2024**, *4*, 109. [CrossRef]
52. Liu, Y.; Guo, H.; Zhang, L.; Liang, D.; Zhu, Q.; Liu, X.; Lv, Z.; Dou, X.; Gou, Y. Research on correlation analysis method of time series features based on dynamic time warping algorithm. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
53. Stenger, M.; Leppich, R.; Foster, I.; Kounev, S.; Bauer, A. Evaluation is key: A survey on evaluation measures for synthetic time series. *J. Big Data* **2024**, *11*, 66. [CrossRef]
54. Martins, A.A.; Vaz, D.C.; Silva, T.A.; Cardoso, M.; Carvalho, A. Clustering of Wind Speed Time Series as a Tool for Wind Farm Diagnosis. *Math. Comput. Appl.* **2024**, *29*, 35. [CrossRef]
55. Liu, X.; Zhang, S.; Wang, X.; Wu, R.; Yang, J.; Zhang, H.; Wu, J.; Li, Z. Clustering Method Comparison for Rural Occupant's Behavior Based on Building Time-Series Energy Data. *Buildings* **2024**, *14*, 2491. [CrossRef]
56. Novák, V.; Mirshahi, S. On the similarity and dependence of time series. *Mathematics* **2021**, *9*, 550. [CrossRef]
57. Berthold, M.R.; Höppner, F. On clustering time series using euclidean distance and pearson correlation. *arXiv* **2016**, arXiv:1601.02213.
58. Cuemath. Euclidean Distance Formula. Available online: https://www.cuemath.com/euclidean-distance-formula/ (accessed on 12 October 2023).
59. Zhang, W.; Wang, J.; Zhang, L. Cosine Similarity: A Comprehensive Review. *J. Stat. Res.* **2020**, *54*, 175–185.
60. Nakamura, T.; Taki, K.; Nomiya, H.; Seki, K.; Uehara, K. A shape-based similarity measure for time series data with ensemble learning. *Pattern Anal. Appl.* **2013**, *16*, 535–548. [CrossRef]
61. To, S.H. Correlation Coefficient: Simple Definition, Formula, Easy Steps. Available online: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/ (accessed on 12 October 2023).
62. Müller, M. *Information Retrieval for Music and Motion*; Springer: New York, NY, USA, 2007; Volume 2.
63. AudioLabs. Dynamic Time Warping (DTW). Available online: https://www.audiolabs-erlangen.de/resources/MIR/FMP/C3/C3S2_DTWbasic.html/ (accessed on 12 October 2023).
64. Dong, C.; Du, G. An enhanced real-time human pose estimation method based on modified YOLOv8 framework. *Sci. Rep.* **2024**, *14*, 8012. [CrossRef]
65. Nguyen, T.D.; Kresovic, M. A survey of top-down approaches for human pose estimation. *arXiv* **2022**, arXiv:2202.02656.
66. Bisht, S.; Joshi, S.; Rana, U. Comprehensive Review of R-CNN and its Variant Architectures. *Int. Res. J. Adv. Eng. Hub (IRJAEH)* **2024**, *2*, 959–966.
67. Chung, J.L.; Ong, L.Y.; Leow, M.C. Comparative analysis of skeleton-based human pose estimation. *Future Internet* **2022**, *14*, 380. [CrossRef]
68. Kim, J.W.; Choi, J.Y.; Ha, E.J.; Choi, J.H. Human pose estimation using mediapipe pose and optimization method based on a humanoid model. *Appl. Sci.* **2023**, *13*, 2700. [CrossRef]
69. Google. Hand Landmarks Detection. Available online: https://developers.google.com/mediapipe/solutions/vision/hand_landmarker/ (accessed on 12 October 2023).
70. Google. Pose Landmark Detection. Available online: https://developers.google.com/mediapipe/solutions/vision/pose_landmarker/ (accessed on 12 October 2023).
71. Perry, A.; Condillac, R.A.; Freeman, N.L.; Dunn-Geier, J.; Belair, J. Multi-site study of the Childhood Autism Rating Scale (CARS) in five clinical groups of young children. *J. Autism Dev. Disord.* **2005**, *35*, 625–634. [CrossRef] [PubMed]
72. Google for Developers. Yin Guobing's Facial Landmark Detector. Available online: https://github.com/yinguobing/facial-landmark-detection-hrnet (accessed on 12 October 2023).
73. Wu, Y.; Ji, Q. Facial landmark detection: A literature survey. *Int. J. Comput. Vis.* **2019**, *127*, 115–142. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.