Development and Validation of an iPad-based Serious Game for Emotion Recognition and Attention Tracking towards Early Identification of Autism

1st Chiara Piazzalunga
DEIB
Politecnico di Milano
Milan, Italy
chiara.piazzalunga@polimi.it

2nd Pierpaolo Molino
DEIB
Politecnico di Milano
Milan, Italy
pierpaolo.molino@mail.polimi.it

3rd Chiara Giangregorio
DEIB
Politecnico di Milano
Milan, Italy
chiara.giangregorio@polimi.it

4th Stefania Fontolan
Department of Medicine and Surgery
Università dell'Insubria
Varese, Italy
stefaniafontolan@gmail.com

5th Cristiano Termine
Department of Medicine and Surgery
Università dell'Insubria
Varese, Italy
cristiano.termine@uninsubria.it

6th Simona Ferrante
DEIB
Politecnico di Milano
Milan, Italy
simona.ferrante@polimi.it

Abstract—**The diagnosis of Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD) can be challenging due to limited accessibility and subjective assessments. Autistic individuals often present difficulties in emotional regulation, emotion recognition and imitation, and in maintaining focus. Emotional expressions and attention are thus hallmarks of ASD and ADHD and can be analyzed to identify these conditions. In this study, we developed and validated a serious game that integrates emotion recognition and attention tracking as a novel tool for identification of ASD and ADHD. Leveraging the TrueDepth camera capabilities, our game provides a cost-effective and user-friendly alternative to current face-tracking technologies. We compared the accuracy of emotion recognition using Euclidean distance with calibrated reference expressions and a calibration-free system based on a machine learning model using Random Forest. We also identified children at risk of ADHD using the Bells test and constructed a machine learning model, utilizing Support Vector Machine and Leave-One-Out Cross Validation, trained on attention data and game data to predict this risk. Our game was tested on 20 adults to validate the emotion recognition system, and then on 17 children of the primary school to assess usability and test the constructed models. The emotion recognition system achieved an accuracy of 0.78 for adults and 0.45 for children, while the machine learning model predicted seven emotions in children with an accuracy of 0.50, suggesting the potential for eliminating the need for calibration. The model also obtained good results in predicting valence and arousal values. The attention model showed excellent validation scores (accuracy: 0.94), indicating the possibility of extending it to a larger cohort. The System Usability Score was excellent (85.0), and children found the game enjoyable, making it a promising tool for ASD and ADHD identification.**

Index Terms—**emotion, attention, autism, ADHD, serious game, machine learning**

I. Introduction

Autism corresponds to an atypical neurological functioning that can result in challenges in social interactions, verbal and nonverbal communication, unusual responses to sensory stimuli, preference for routines, and repetitive behaviors. While these are common features of autism, each autistic individual may to exhibit some, but not all of these characteristics to varying degrees. Therefore, autistic people's support needs vary greatly, as do the accomodations necessary to improve their quality of life. In fact, autism has been historically classified by traditional psychiatry as a neurodevelopmental disorder, but recently a growing branch of research considers it a neurodiversity [1]. This perspective recognizes that while autistic individuals may have disabilities and unique support needs, their differences should be accommodated rather than cured or erased. However, research shows that earlier diagnosis and intervention can substantially benefit those who receive them [2]. This is relevant not only for those who have disabilities related to autism, but also for "high functioning" autistic people, who have lighter symptoms and whose difficulties are often dismissed due to the fact that, to an external observer, their functioning seems typical, if only a bit odd. These individuals may struggle to receive a diagnosis, both during childhood, since they do not exhibit signs that parents or caregivers may find worrying, and during adulthood, having by then learned to mask those behaviors that make them feel out of place. Living with unidentified autism can take its toll on a person's mental health, and can lead to anxiety, depression, and substance abuse, but the diagnosis is not easy to obtain [3].

979-8-3503-2745-8/23/$31.00 ©2023 European Union

Some of the main criticalities of the diagnosis process are caused by general faults of healthcare systems: lack of trained specialists, absence of specialized facilities close to home, long waiting lists which force people to seek private practitioners [4]. This makes it particularly difficult for people with a low socioeconomic status. Other problems arise due to bias, such as the one regarding women, who are much less likely to receive a diagnosis [5], or are identified later in life [6], because golden standards have been defined and normed on a male population. Although the recent increase in the incidence of autism can be seen as a sign that access to diagnosis has improved, there is still much to do.

The same issues are relevant to the diagnosis of neurodevelopmental disorders. For example, the most common one, ADHD (Attention Deficit and Hyperactivity Disorder), often goes undiagnosed, especially in adults [7]. ADHD is, in fact, generally considered a childhood disorder, but symptoms persist into adulthood. Adults with ADHD experience impaired educational achievement and higher risks of substance abuse and imprisonment [8].

Autism and ADHD often co-occur: it has been estimated that 1 in 8 children with ADHD is also autistic, while ADHD is the most common comorbidity in autistic children, with rates from 40% to 70% [9]. Improving the diagnosis process of one of these conditions can smooth the path towards the identification of the other one.

To overcome bias and improve accessibility to both diagnoses, technology can come in handy. In fact, many contexts already benefit from the use of e-health monitoring software to identify anomalies in children's neurological development, thanks to their ability to capture patterns in features that could go unnoticed even to the most expert eyes [10][11]. However, finding objective indicators of autism traits or of ADHD is hard, as phenotypes can be extremely varied. Consequently, tools developed for this purpose should be modular and give access to ways to analyze all of the different manifestations of autism and ADHD. Some of the starting points that we have considered in this work are emotions and attention.

Although the fact that autistic people are incapable of feeling emotions is only a common misconception, autism affects both emotion recognition and emotion imitation [12]. Studies have found that neurotypical individuals recognize emotions better if they see them on realistic depictions (such as videos or photos), while autistic people perform better with cartoon depictions [13]. Regarding imitation, in autistic individual emotion imitation is slower and less precise [14]. However, there is nothing more subjective than emotions, and clinicians' bias can greatly influence their assessment.

Attention, on the other hand, is an indicator for ADHD, as the name of the disorder itself says. In fact, people with ADHD tend to have more difficulties in maintaining focus for a prolonged period of time [15].

II. Related works

There have been many attempts to use technology to gather quantitative data to investigate autism and ADHD and to obtain a more timely and accurate diagnosis. However, automated systems for emotion recognition are sometimes very invasive and expensive. Some examples are electromyography [16], neuroimaging [17], or marker-based systems [18]. Testing autistic individuals is particularly delicate, since they could have sensory issues that may be triggered by electrodes and markers.

Attention, on the other hand, has been put in relation with the blink rate and with the rotation and position of the head, and previous studies have computed quantitative measures to evaluate it based on these parameters [19][20]. However, a quantitative measure of these features needs to be collected by an automated system, and as of now such tools are, like the ones that track emotions, invasive or expensive.

Eye tracking has been explored as well, particularly to analyze attention, both for autism and ADHD, although the usual approaches involve either expensive trackers or computationally heavy systems based on computer vision and video analysis [21].

Some researchers have turned to games and mobile applications as potential solutions, as they are readily accepted and appreciated by children [22] [23]. However, these approaches still rely on video analysis and fail to address challenges such as computational complexity and the need for anonymity in data collection, as pointed out by [24] regarding voice analysis.

A solution can be found in the TrueDepth system, a technology embedded in Apple cameras on some iPhone and iPad models introduced after 2017. TrueDepth includes an infrared camera and a projector that projects over 30,000 dots over one's face, feeding then these data to a neural engine that constructs a mathematical model of a face. Apple then provides the ARKit, a framework to develop augmented reality applications that gives access to the 52 parameters extracted from the TrueDepth camera [25]. The accuracy is surely lower than the ones obtained in state-of-the-art systems, but the price and the complexity are significantly reduced. Besides, children are accustomed to smartphones and tablets, so their experience with the tool would probably be more natural. For instance, most autistic children have a natural affinity for technology and a good attitude toward learning on computers [26], proving to be a promising explorable tool. However, a rigorous work of validation should take place to understand whether this technology is sensitive and reliable enough to be employed in this field.

Given these premises, the aim of this work is to present the design and development of a platform of serious games that exploit the TrueDepth camera of an iPad Pro for emotion recognition and attention tracking.

III. Materials and Methods

A. Definition of emotions and of valence and arousal values

Before trying to recognize emotions, we had to define which ones we wanted to take into consideration, to find a trade-off between simplicity and sensitivity.

Emotions are variegated and full of nuances, so it is not easy to arbitrarily isolate a subset of them. However, they can be mapped into a two-dimensional space through their valence and their arousal. Valence, or hedonic tone, is the property which specifies if the emotion is positive or negative [27], while arousal indicates the intensity of that emotion [28].

While choosing the emotions to consider, we thus included the six basic emotions [29], but added others, such as neutrality, desperation, enthusiasm, and tiredness, to cover the whole spectrum of both valence and arousal. The complete set of emotions chosen is reported in Figure 1 [30].

Each emotion was assigned with a value of valence and arousal. To simplify the mapping and to allow for a classification of each emotion in a reduced number of categories, the only values assigned were 1, 0, and -1. They are reported in Table I.

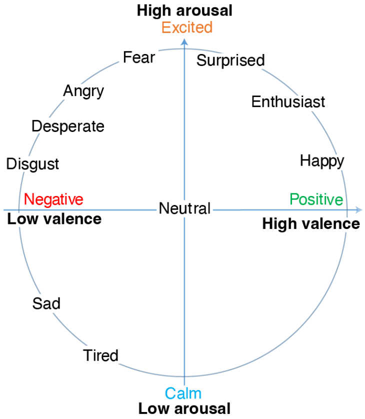


Fig. 1: Chosen emotions for the validation.

TABLE I: Valence and arousal values

Emotion	Valence	Arousal
Neutrality	0	0
Happiness	+1	0
Enthusiasm	+1	+1
Surprise	0	+1
Desperation	-1	+1
Anger	-1	+1
Disgust	-1	0
Fear	0	+1
Sadness	-1	-1
Tiredness	-1	-1

B. Game design

The game was developed in Unity 2021.1.11f1 and built for iOS on an iPad Pro 12.9" (5th generation), equipped with the TrueDepth camera and the ARKit [25], which exposes 52 parameters which describe the facial expression. Each one ranges from 0 to 1, where 0 is the neutral position and 1 is the maximum movement.

Two versions of the serious game were developed: the first one served as a validation tool to assess the accuracy of the emotion recognition system, while the second one was the actual game provided to children. Some parts were common to both versions, including the insertion of personal data and the choice of the avatar (shown in Figure 2), which will be superimposed on the face of the player and mimic their facial movements, and a calibration phase.

Fig. 2: Selectable avatars.

1) Calibration: The app asks the player to mimic ten emotions in succession to perform the calibration. Each emotion is prompted three times.

For each expression, the app saves the values of the parameters that the ARKit exposes. Finally, the three configurations of each expression are summarized into one by the arithmetic mean of their parameters. At the end of this calibration, each subject has a reference expression for each emotion.

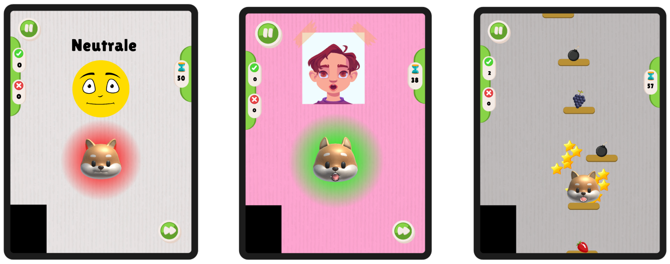
2) Validation game: To assess the accuracy with which our game could identify emotions, a validation app was developed. After the calibration phase described in III-B1, the app asked the user to mimic the same emotions provided during calibration, in the same order four times in a row, for a total of 40 emotions. The classification was made based on the Euclidean distance between the performed facial expression and subject-specific references obtained through the calibration phase. The Validation scene is shown in Figure 3a.

3) Emotion Recognition game: This game, shown in Figure 3b, is based on the ability to recognize and reproduce expressions. Literature shows that autistic people read emotions better in cartoons than in real photos, and they also did it with better accuracy than neurotypical participants. In the game, seven expressions made by a cartoon are first shown and must be reproduced by the user. Then, the player is asked to confirm the emotion they identified by selecting it with the corresponding button, to distinguish between an error made by the tool and one made by the player.

4) Fruit Salad game: This game is an endless game, in which the avatar constantly jumps vertically and moves horizontally depending on the player's head position, going from one platform to the next. The avatar opens its mouth when the player does, and can eat the in-game items, namely fruits, which give points, and bombs, which must be avoided. A screenshot of this game can be seen in Figure 3c.

C. Protocol

Recruitment was performed in accordance with relevant guidelines and regulations and in accordance with the Declaration of Helsinki; the protocol was approved the university's Ethical Committee n. 04/2021.



(a) (b) (c)

Fig. 3: Screenshots showing (a) the Validation scene, (b) the Emotion Recognition game, and (c) the Fruit Salad game.

1) Game: The tablet was kept on a stand positioned at an angle of 14 degrees. The subject was asked to keep a distance ranging from 50 to 60 cm from the screen throughout the duration of the game to guarantee repeatability and consistency between measurements. The game provided visual feedback in case the subject moved too close or too far.

2) System Usability Scale: All subjects graded the System Usability (SUS) [31]. Each statement can be rated on a scale from 1 to 5, with 1 representing the lowest approval with the statement and 5 representing the highest approval. The SUS scale is then used to assess whether the video game is too complex, difficult to navigate, or cumbersome.

3) Custom satisfaction questionnaire: In addition, subjects were asked to answer open-ended questions:

- What did you like the most?
- What did you like the least?
- Would you change anything in the game?

Finally, subjects were asked:

- Was the game fun?
- Was the game boring?

The answers were five-points Likert scales, from “Not at all” (coded as 1) to “A lot” (coded as 5).

4) Bells test: Children were administered the Bells test [32], which is a cognitive test, commonly used to assess attention and executive functions. Children had to find all of the bells in a crowded image. For each of them, accuracy and speed were recorded.

D. Data analysis

The data are stored both locally and online, in Firebase’s Realtime Database. Stored data include:

- Calibration data: 10 emotions composed of 52 features each;
- Characterization information: gender, age, eyeglasses;
- Game data (only for children):
  - Emotion Recognition game: prompted, recognized and selected emotion.
  - Fruit Salad game: game time, score, bombs eaten, bombs avoided, fruits missed.

Data analysis was performed offline with Python 3.9.7.

1) Tool validation: At the end of the validation process a total of 800 records, each including the 52 features recorded by the TrueDepth system and the labels identifying the prompted emotion and the recognized emotion, were obtained. The emotion recognition system was evaluated by means of overall accuracy and of emotion-specific accuracy, to assess the tool’s recognition power and to understand whether errors were due the similarity between certain emotions.

Since conventional face trackers sometimes function poorly when the subject wears glasses, due to the noise produced by glasses’ reflections [33], we wanted to verify if our game was robust to this. We thus performed a t-test to assess the impact of eyeglasses on the accuracy of the emotion recognizer. Another factor that could have a confounding effect is gender, so we computed its impact in the same way.

2) Emotion classifier: The emotion recognition tool we used for validation is dependent on the calibration, which can improve accuracy, but also pose some criticalities and limitations. Firstly, our application is designed for children, particularly autistic ones, which may have difficulties sitting still during this phase. Failure in properly acquiring reference expressions can lead to an unsatisfying functioning of the tool. Secondly, removing the need for the calibration can make the game more enjoyable and its administration faster. We thus implemented a machine learning model to understand whether it was possible to predict the performed emotion even without a calibration phase.

We considered a variety of algorithms:

- Gaussian Naive Bayes (Gaussian NB);
- Random forest (RF);
- Support Vector Machine (SVM);
- CatBoost.

The models were trained on the 52 features extracted by the TrueDepth system during adults’ calibration and validation, and validated through Cross Validation (CV) to assess their quality. The best one was then fitted on adults’ calibration and validation data in which the classes were balanced through SMOTE (Synthetic Minority Oversampling TEchnique), then tested on children’s calibration data. The performance was evaluated three times, when predicting the performed emotion, the valence, and, finally, the arousal.

3) Analysis of attention and Bells test classification: There is no objective way to compute the level of attention during a task. Some studies rely on eye movements, which are often an expression of visual attention, while others measure the excitation level of some areas of the brain [34]. However, these techniques are often invasive or costly. We have instead identified some elements that are usually regarded as markers of attention and that TrueDepth can recognize. These are the blink rate, computed as the average number of blinks over a period of 5 seconds while playing the Emotion Recognition game, the rotation of the head, and the number of times the subject has looked outside the screen. Regarding the dependent variable, children were divided between at-risk and not-at-risk depending on the results in the Bells test: those who had less than -2 in the Z-score of the accuracy in the test were labeled as at-risk. We considered the same models used in Section III-D2 and selected the best one through the same strategy. However, since the dataset only comprises data from children and is thus fairly small, the final model was not evaluated by its performance on a test set, but through a Leave-One-Out Cross Validation, after augmenting the minority class with SMOTE. SMOTE and the CV were steps of a pipeline that made sure to only use real records as the validation samples, as to not contaminate the validation process with augmented data. The best hyperparameters for the SVM were chosen through a Grid Search.

IV. Results

A. Sample description

37 subjects participate in the study, 20 adults (age 26.0 ± 4.32, 8 males and 12 females) to validate the expression tracking model, and 17 children (age 7.88 ± 0.58, 9 males and 8 females) for the testing. 8 adults and 1 child wore glasses.

B. System Usability Scale and satisfaction questionnaire

The SUS score obtained was 85 ± 7.5, which is way higher of the minimum score necessary to consider a system usable, which is 68, and equal to the score that makes a system excellent. In general, children's feedback was positive. The question "Was the game fun?" obtained a score of 4.94 ± 0.23, while the question "Was the game boring?" obtained a score of 1.27 ± 0.43.

C. Tool validation

The general accuracy of the emotion recognition system on adults' expressions is 0.67. Figure 4 shows the confusion matrix, with the results for each emotion. It is apparent that some emotions are recognized more easily than others. For example, Desperation has a fairly low accuracy (0.25), while Surprise and Fear are often confused with one another. For this reason, the initial set of 10 emotions was resized to contain only 7, after merging Surprise and Fear, and removing Desperation and Tiredness. This brings the accuracy from 0.67 to 0.78.

Regarding the robustness of the tool to gender and eyeglasses, a t-test was performed to check whether there were differences in the distribution of the number of recognized emotions in the two populations. The test showed that there was no significant difference in neither of the two cases (p=0.85 for gender, p=0.95 for eyeglasses).

D. Emotion classification

Figure 5 reports the confusion matrix of the children's emotions prompted and recognized in the Emotion Game. In this case, the accuracy is significantly lower, standing at 0.45, which prompted us to search for a machine learning-based method to recognize emotions.

We constructed four different machine learning algorithms to assess which was the best at predicting the correct emotion, the valence value, or the arousal value. The models were

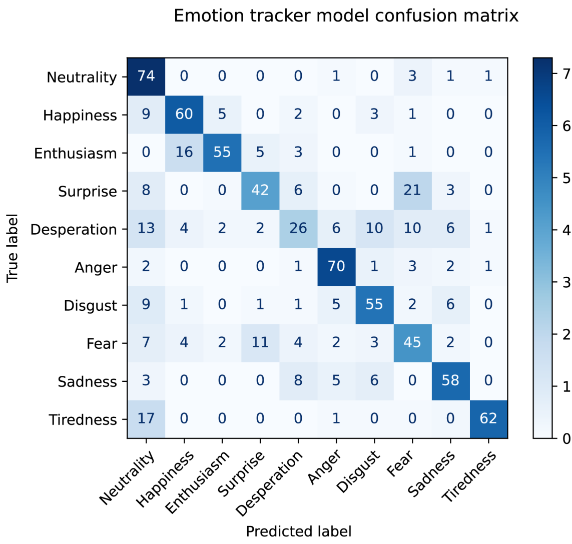


Fig. 4: Confusion matrix of the emotions recognized by the game in the adults' validation phase.

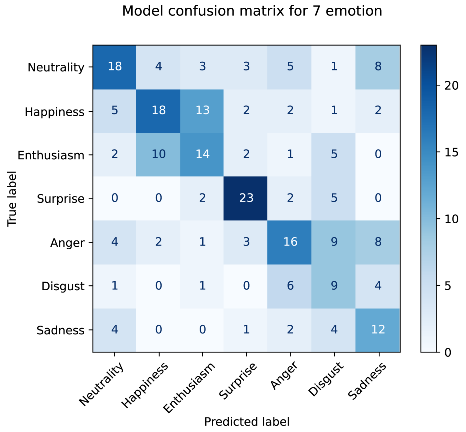


Fig. 5: Confusion matrix of the emotions recognized by the Emotion Recognition game in children.

TABLE II: Accuracy values of classifiers obtained on validation set

Algorithm	Emotion	Valence	Arousal
	(7 classes)	(3 classes)	(3 classes)
Gaussian NB	0.19	0.35	0.42
CatBoost	0.25	0.44	0.49
SVM	0.24	0.45	0.45
RF	0.26	0.45	0.49

validated through a 10-fold Cross Validation and its results are reported in Table II.

The best model was the one using the RF algorithm. It was tested using the children's calibration data and its results are reported in Table III.

TABLE III: Accuracy, precision, recall, and F1 scores of the RF classifier on test data

Target	Accuracy	Precision	Recall	F1
Emotion	0.50	0.50	0.50	0.50
Valence	0.71	0.71	0.71	0.71
Arousal	0.55	0.52	0.56	0.53

E. Attention classification

We constructed a machine learning model to predict the level of attention of the children. Considering the risk threshold set at a Z-score of -2 in the Bells test, 11 children were not at risk, while 6 were at risk. We considered four models (Gaussian NB, CatBoost, SVM and RF), and trained them on attention data (blink rate, head rotation, gazes outside of the screen) and game data (time, score, bombs eaten and avoided, and fruits missed). The accuracy scores on the validation set are reported in Table IV.

TABLE IV: Accuracy values of classifiers obtained on validation set

Algorithm	Accuracy (2 classes)
Gaussian NB	0.25
CatBoost	0.45
SVM	0.70
RF	0.45

The best model was the one using the SVM algorithm. It was then optimized through hyperparameter tuning and class balancing. The Grid Search selected the following hyperparameters:

- C = 0.001
- gamma = 1000
- kernel = radial basis function

After these refinements, the accuracy of the best estimator reached 0.94 ± 0.23.

V. Discussion

Autism is an atypical neurological function that is associated, among other things, to difficulties in recognizing and imitating emotions and to attention problems, often due to its co-occurrence with ADHD. Autism and ADHD often go undiagnosed, leading to a worse quality of life and a higher risk of depression, substance abuse and suicide. Autism and ADHD diagnoses present several problems, like accessibility to the diagnostic process and clinicians' biases. Technological tools should thus be employed to support the early screening of autistic traits, providing the possibility to begin the diagnostic process as soon as possible and access appropriate support. The tool should be gamified and fun, and cover a variety of abilities. We thus developed a serious game focused on emotion recognition and on attention tracking, exploiting the iPad's TrueDepth system.

The main objectives of this study were the validation of the emotion recognition system and the definition of a model of attention, towards the creation of a tool that will not be intended for diagnosis but rather will to offer alerts to teachers or clinicians, signaling potential cases of undetected autism and ADHD.

The game was at first tested on 20 adults (age 26.0±4.32), who only performed the calibration and the validation phase. The emotion recognition system, based on the comparison of the performed emotion and the calibrated expression previously saved, had an accuracy of 0.67, which is fairly high if we consider the fact that it had to discriminate among 10 classes. The results seem to be in accordance with the six main emotions that are traditionally identified (Happiness, Anger, Disgust, Fear, Surprise, Sadness), although our system had difficulties in differentiating between Surprise and Fear. This may be due to the fact that each subject had a different way of interpreting those expressions, and sometimes the features that one associated with Fear were the same that another associated to Surprise. However, it seems reasonable to unify them, as they have similar manifestations. Desperation and Tiredness were eliminated, on the account of their scarce accuracy and on the fact that they are not present in the usual six emotions. Accuracy improved to 0.78 when the initial set was resized to contain only 7 emotions. T-tests showed no effect of gender (p=0.85) nor of eyeglasses (p=0.95), which means that our system is robust to these factors.

17 children (age 7.88±0.58) tested the game. They performed the calibration and played the Emotion Recognition game and the Fruit Salad game. In general, their opinion on the game was good, as the SUS reached a score of 85±7.5. They also thought that the game was fun (4.94±0.23 out of 5) and not boring (1.27±0.43 out of 5). However, for these subjects, calibration proved to be more problematic: in general, they had difficulties understanding when they had to stand still and, even when they did, they were easily distracted or tired. The accuracy of emotion recognition system is significantly lower, standing at 0.45. Moreover, some of them expressed boredom towards the calibration phase. Because of this, we tried a machine learning approach to verify if it was possible to include a ML recognizer in future versions of the app, in which the calibration would not be necessary anymore. We used a Random Forest classifier fitted on the TrueDepth data of the adults. The model obtained overall good results when recognizing the emotions performed by children, with an accuracy of 0.50, which is higher than the one reached using the Euclidean distance. This means that the model can generalize quite well, even though it is fairly simple and trained on a small dataset, and suggests the possibility to strengthen the model and eliminate the calibration step entirely. Moreover, head rotation was not included in the features on which the model was trained, but for several

subjects, particularly children, it had a great relevance. For example, the Sadness emotion was often imitated through the lowering of the head. Thus, it would be useful to include it in a future version of the application.

Trying to identify the specific emotion can be useful when analyzing a subject's ability to recognize or imitate it, but there are other applications in which knowing the valence or arousal would be enough. For example, the valence can give an idea of the general mood in which the player is, while the arousal may give an indication of the emotional involvement. For this reason, the same model used for emotions was fitted on the same data two more times, predicting the valence and the arousal values, which both were either +1, 0, or -1, reaching respectively accuracies of 0.71 and 0.55. The valence model showed good results, while the arousal one had lower performances, probably because there were significantly less samples with a -1 arousal value, since only Sadness has this value, as can be seen in I. The oversampling solved in part this problem, which, however, could not be eliminated completely. This transpires from the fact that the class that has an arousal value of 1 has a F1-score of 0.61, the 0-arousal class has a 0.55 F1-score, while the -1-arousal class has a 0.36 F1-score.

Regarding attention, a model leveraging the SVM algorithm and evaluated through Leave-One-Out Cross Validation (LOOCV) was constructed. It was fitted on attention data gathered during gameplay, such as blink rate, head rotation, and gazes outside of the screen, and game data, such as score and points collected. The target variable of the model was the risk category of the Bells test. The accuracy reached by the model is 0.94±0.23. This is a very good result, although we must remember that this model has not been evaluated on test data, due to the scarcity of samples, but nevertheless this suggests the possibility to use this setup in future works. At the same time, LOOCV is adequate for our application, as it demonstrates that the model can learn effectively and obtains good results when tested on a subject it has never encountered.

Ethical Impact Statement

A. Issues related to human subjects

The study was carried out after the approval of the university's Ethical Committee. The participants signed an informed consent. Since some of them were minors, they signed a simplified version of the informed consent form, and their parents or legal guardians signed the official informed consent. On recruitment, each participant was assigned an alphanumeric code. Their data was identified by that code only, while the correspondence between the code and the participant identity was known to a subset of researchers, who had it to be able to delete the participant's data if they wished to do so. Moreover, the expressions are saved only through parameters and the app does not record images nor videos.

B. Issues related to potential negative societal impact

The application can give information about emotions expressed and attention during gameplay. It can classify people at risk of having ADHD. A potential misuse could be the discrimination born from the result of the classification: a healthy person could be classified as at-risk and be discriminated for it, or an at-risk person could be classified as healthy and not receive the treatment they may need. However, the classification is not validated nor 100% certain and should not be treated as such. Moreover, difficulties with emotion recognition and imitation is not the only autistic trait in existence, so an autistic person with no difficulty regarding emotions could feel misrepresented. We would like to extend the application to include a variety of other aspects. We did not consult with autistic people or individuals with ADHD yet, because this game was developed in the framework of a larger project that did not have people with these characteristics among its participants. However, future interactions with stakeholders is certainly a priority for us.

C. Issues related to limits of generalizability

Each person's interpretation of emotions is different. In fact, the first validation technique (Euclidean distance) struggle to identify the emotions with a lot of variability (e.g., tiredness). However, the machine learning approach has mitigated this problem, even though the criticality will remain until the dataset on which the models are trained becomes much bigger and varied.

Acknowledgment

This work is funded by European H2020 project ESSENCE (SC1-PHE-CORONAVIRUS-2020-2B; G.A. 101016112).

References

[1] Deborah J Morris-Rosendahl and Marc-Antoine Crocq. Neurodevelopmental disorders the history and future of a diagnostic concept. *Dialogues Clin. Neurosci.*, 22(1):65–72, March 2020.

[2] Susan E Bryson, Lonnie Zwaigenbaum, and Wendy Roberts. The early detection of autism in clinical practice. *Paediatr. Child Health*, 9(4): 219–221, April 2004.

[3] Yvette Hus and Osnat Segal. Challenges surrounding the diagnosis of autism in children. *Neuropsychiatric Disease and Treatment*, Volume 17:3509–3529, December 2021.

[4] Natasha Malik-Soni, Andrew Shaker, Helen Luck, Anne E Mullin, Ryan E Wiley, M E Suzanne Lewis, Joaquin Fuentes, and Thomas W Frazier. Tackling healthcare access barriers for individuals with autism from diagnosis to adulthood. *Pediatr. Res.*, 91(5):1028–1035, April 2022.

[5] Meng-Chuan Lai, Michael V. Lombardo, Greg Pasco, Amber N. V. Ruigrok, Sally J. Wheelwright, Susan A. Sadek, Bhismadev Chakrabarti, and Simon Baron-Cohen and. A behavioral comparison of male and female adults with high functioning autism spectrum conditions. *PLoS ONE*, 6(6):e20835, June 2011.

[6] Rachel Loomes, Laura Hull, and William Polmear Locke Mandy. What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474, June 2017.

[7] Ylva Ginsberg, Javier Quintero, Ernie Anand, Marta Casillas, and Himanshu P Upadhyaya. Underdiagnosis of attention-deficit/hyperactivity disorder in adult patients. *Prim. Care Companion CNS Disord.*, June 2014.

[8] Joseph Austerman. Adhd and behavioral disorders: Assessment, management, and an update from dsm-5. *Cleveland Clinic Journal of Medicine*, 82(11 suppl 1):S2–S7, 2015. ISSN 0891-1150.

[9] Camille Hours, Christophe Recasens, and Jean-Marc Baleyte. ASD and ADHD comorbidity: What are we talking about? *Front. Psychiatry*, 13, February 2022.

[10] Khaleghi Ali, Fatemeh Heydari, Hadi Haedar, Takhttavani Maede, and Alireza Soltani Nezhad. An approach to diagnose cognitive deficits: gamifying adhd children diagnosis questionnaire. 04 2018.

[11] Kokol Peter, Helena Blazun Vosner, Jernej Zavrnik, Joeri Vermeulen, Samaa Shohieb, and Frank Peinemann. Serious game-based intervention for children with developmental disabilities. *Curr. Pediatr. Rev.*, 16(1): 26–32, 2020.

[12] Hannah Meyer-Lindenberg, Carolin Moessnang, Bethany Oakley, Jumana Ahmad, Luke Mason, Emily J. H. Jones, Hannah L. Hayward, Jennifer Cooke, Daisy Crawley, Rosemary Holt, Julian Tillmann, Tony Charman, Simon Baron-Cohen, and Tobias Banaschewski. Facial expression recognition is linked to clinical and neurofunctional differences in autism. *Molecular Autism*, 13, 11 2022. ISSN 2040-2392.

[13] Gray Atherton and Liam Cross. Reading the mind in cartoon eyes: Comparing human versus cartoon emotion recognition in those with high and low levels of autistic traits. *Psychological Reports*, 125, 03 2021.

[14] Hanna Drimalla, Irina Baskow, Behnoush Behnia, Stefan Roepke, and Isabel Dziobek. Imitation and recognition of facial emotions in autism: a computer vision approach. *Molecular Autism*, 12(1), April 2021.

[15] Lara Tucha, Anselm B. M. Fuermaier, Janneke Koerts, Rieka Buggenthin, Steffen Aschenbrenner, Matthias Weisbrod, Johannes Thome, Klaus W. Lange, and Oliver Tucha. Sustained attention in adult ADHD: time-on-task effects of various measures of attention. *Journal of Neural Transmission*, 124(S1):39–53, July 2015.

[16] Rongjie Li, Yao Wu, Qun Wu, Nilanjan Dey, Rubén González Crespo, and Fuqian Shi. Emotion stimuli-based surface electromyography signal classification employing markov transition field and deep neural networks. *Measurement*, 189:110470, 2022. ISSN 0263-2241.

[17] Madeline B. Harms, Alex Martin, and Gregory L. Wallace. Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review*, 20(3): 290–322, September 2010. doi: 10.1007/s11065-010-9138-6. URL https://doi.org/10.1007/s11065-010-9138-6.

[18] Ulrik Söderström, Songyu Li, Harry Claxton, Daisy Holmes, Tom Ranji, Carlos Santos, Carina Westling, and Harry Witchel. Toward emotional recognition during hci using marker-based automated video tracking. *ECCE 2019: Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 49–52, 09 2019.

[19] Antonio Maffei and Alessandro Angrilli. Spontaneous blink rate as an index of attention and emotion during film clips viewing. *Physiology and Behavior*, 204:256–263, 5 2019. ISSN 1873507X.

[20] Masaaki Goto, Tetsuo Tanaka, and Kazunori Matsumoto. Estimating attention level from blinks and head movement. 77:52–59, 2021. ISSN 2398-7340.

[21] Jessica S. Oliveira, Felipe O. Franco, Mirian C. Revers, Andréia F. Silva, Joana Portolese, Helena Brentani, Ariane Machado-Lima, and Fátima L. S. Nunes. Computer-aided autism diagnosis based on visual attention models using eye tracking. *Scientific Reports*, 11(1), May 2021. doi: 10.1038/s41598-021-89023-8. URL https://doi.org/10.1038/s41598-021-89023-8.

[22] Nicholas Deveau, Peter Washington, Emilie Leblanc, Arman Husic, Kaitlyn Dunlap, Yordan Penev, Aaron Kline, Onur Cezmi Mutlu, and Dennis P. Wall. Machine learning models using mobile game play accurately classify children with autism. *Intelligence-Based Medicine*, 6:100057, 2022. ISSN 2666-5212. doi: https://doi.org/10.1016/j.ibmed.2022.100057. URL https://www.sciencedirect.com/science/article/pii/S2666521222000102.

[23] Haik Kalantarian, Peter Washington, Jessey Schwartz, Jena Daniels, Nick Haber, and Dennis P. Wall. Guess what? *Journal of Healthcare Informatics Research*, 3(1):43–66, October 2018. doi: 10.1007/s41666-018-0034-9. URL https://doi.org/10.1007/s41666-018-0034-9.

[24] Jung Hyuk Lee, Geon Woo Lee, Guiyoung Bong, Hee Jeong Yoo, and Hong Kook Kim. Deep-learning-based detection of infants with autism spectrum disorder using auto-encoder feature representation. *Sensors*, 20(23), 2020. ISSN 1424-8220. doi: 10.3390/s20236762. URL https://www.mdpi.com/1424-8220/20/23/6762.

[25] Apple developer arkit documentation. https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation. Accessed: 2023-04-12.

[26] Chu-Sui Lin, Shu-Hui Chang, Wen-Ying Liou, and Yu-Show Tsai. The development of a multimedia online language assessment tool for young children with autism. *Research in Developmental Disabilities*, 34(10): 3553–3565, 2013. ISSN 0891-4222.

[27] Juliette Vazard. Feeling the unknown: Emotions of uncertainty and their valence. *Erkenntnis*, July 2022.

[28] James A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, 1994.

[29] P. Ekman and W.V. Friesen. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Number v. 10 in Spectrum book. Malor Books, 2003.

[30] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1): 42–50, January 2021.

[31] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[32] Louise Gauthier, Francois Dehaut, Yves Joanette, et al. The bells test: a quantitative and qualitative test for visual neglect. *International journal of clinical neuropsychology*, 11(2):49–54, 1989.

[33] Su Gwon, Chul Cho, Hyeon Lee, Won Lee, and Kang Park. Gaze tracking system for user wearing glasses. *Sensors (Basel)*, 14(2):2110–2134, January 2014.

[34] Giulia Towey, Rosa Fabio, and Tindara Caprì. *Measurement of Attention*, pages 41–83. 05 2019.