Towards Clinical Phenotyping at Scale with Serious Games in Mixed Reality

Mariem Hafsia [1], Romain Trachel[2], and Guillaume Dumas[1,3,4]

[1]Department of Psychiatry, University of Montreal
[2]Eidos Sherbrooke, Sherbrook C
[3]Mila - Quebec AI Institute
[4]CHU Sainte-Justine Research Center

Abstract

Context: Mental healthcare systems are facing an ever-growing demand for appropriate assessment and intervention. Unfortunately, services are often centralized, overloaded, and inaccessible, resulting in greater institutional and social inequities. Therefore, there is an urgent need to establish easy-to-implement methods for early diagnosis and personalized follow-up. In recent years, serious games have started to offer such a clinical tool at scale. **Problem:** There are critical challenges to the development of secure and inclusive serious games for clinical research. First, the quality of the data and features analyzed must be well defined early in the research process in order to draw meaningful conclusions. Second, algorithms must be aligned with the purpose of the research while not perpetuating bias. Finally, the technologies used must be widely accessible and sufficiently engaging for users. **Focus of the paper:** To tackle these challenges, we designed a participatory project that combines three innovative technologies: Mixed Reality, Serious Gaming, and Machine Learning. We analyze preliminary data with a focus on the identification of the players and the measurement of classical biases such as sex and environment of data collection. **Method:** We co-developed with patients and their families, as well as clinicians, a serious game in mixed reality specifically designed for evaluation and therapeutic intervention in autism. Preliminary data were collected from neurotypical individuals with a mixed reality headset. Relevant behavioral features were extracted and used to train several classification algorithms for player identification. **Results:** We were able to classify players above chance with slightly higher accuracy of neural networks. Interestingly, the accuracy was significantly higher when players were separated by sex. Furthermore, the uncontrolled condition showed better levels of accuracy than the controlled condition. This could mean that the data are richer when the player interacts freely with the game. Our proof of concept cannot exclude the possibility that this last result is linked to the experimental setup. Future development will clarify this point with a larger sample size and the use of deep learning algorithms. **Implications:** We show that serious games in mixed reality can be a valuable tool to collect clinical data. Our preliminary results highlight important biases to consider for future studies, especially for the sex and context of data collection. Next, we will evaluate the usability, accessibility, and tolerability of the device and the game in autistic children. In addition, we will evaluate the psychometric properties of the serious game, especially for patient stratification. This project aims to develop a platform for the diagnosis and therapy of autism, which can eventually be easily extended to other conditions and settings such as the evaluation of depression or stroke rehabilitation. Such a tool can offer novel possibilities for the study, evaluation, and treatment of mental conditions at scale, and thus ease the burden on healthcare systems.

Key Words: Interactive Psychometrics, Mental Health, Deep Phenotyping, Machine Learning, Autism

1 Introduction

With the recent increase in mental health awareness comes a pressing need for evaluation and intervention. However, health services struggle to meet this demand, mainly due to centralized systems and a lack of trained professionals. In mental health, the wide variability of symptoms between and within individuals makes standard psychometric tools unreliable. Diagnosis and treatment often require multiple appointments with several specialized professionals. In addition to overburdening health facilities, this exhausting and expensive process can critically affect access to the care system and further reinforce social and institutional inequalities. For a condition that requires lifelong follow-up, such as autism, these implications can be even more critical.

Autism Spectrum Disorders (ASDs) involve a large variety of symptoms, including difficulties in social interaction and communication, as well as repetitive behaviors and restricted interests [10]. Although the severity of symptoms varies between individuals, some people may require assistance throughout their lives. Early and intensive interventions have been effective in helping autistic children improve their communication and social skills[6] but can, in time, be very costly. For an autistic person, uncovered insurance expenses care are estimated at $4,000 per year.[12] The lifetime cost associated with autism has been estimated between $2.4 and $3.2 million per person[8, 9]. These numbers underline the urgency to develop less onerous and more broadly accessible alternatives.

In recent years, advances in technology have offered new ways to tackle these challenges using serious games. In addition to being entertaining, they can offer a more affordable and accessible alternative to conventional methods [4]. This approach provides controlled and reproducible environments for the study, intervention, and assessment of mental health, while also adapting to specific needs and preferences. In recent years, head-mounted displays have gained popularity in clinical settings for providing immersive and engaging experiences. For Autism, Mixed Reality (MR), which is a hybrid between virtual reality and augmented reality, has the advantage of allowing social interaction with the subject. Several studies have shown its potential to help people with cognitive disabilities learn new skills and become more autonomous[1].

These devices can record a great quantity of data generated by engaging with a video game (e.g. score, movements, reaction times). Machine Learning analyses, which are sophisticated methods based on mathematical and probabilistic formulas, can then leverage those rich data to extract meaningful patterns. For mental health research, deep phenotyping provides a promising approach to characterize and identify patients based on multimodal observations and brings us one step closer to precision medicine [13]. However, machine learning can introduce and perpetuate biases, for instance regarding sex in healthcare [3], and this becomes critical to carefully control for those effects.

The Human Dynamic Clamp (HDC) is a human-computer interface consisting of a virtual partner whose behavioral model is empirically supported by findings from social neuroscience and experimental psychology. This neuroinspired avatar can interact with humans and responds in real-time. A 2D version of the HDC paradigm has been validated in neurotypicals and has demonstrated its ability, in interaction, to transfer new behaviors to participants [5]. Early research with children and adolescents in this therapeutic setting has shown the ability of the system to unravel sensory-motor and social-cognitive skills of autism [2]. To make this tool more attractive to pediatric populations in a clinical context, a gamified 3D version of the paradigm was developed for MR.

Here, we evaluate the resulting video game with multiple algorithms using preliminary data collected from neurotypical adults. The results illustrate how the combination of serious games, MR, and machine learning can help with clinical phenotyping.

2 Methods

2.1 Mixed Reality Setup

We used the HoloLens 2 MR helmet (Microsoft Corp., Redmond, WA). It is currently the most comfortable and safe MR device on the market, certified for industrial and medical use. The entire device is untethered, so users can move freely without wires or external packs. It is a self-contained computer with Wi-Fi connectivity and multiple sensors that scan the environment and measure body gestures. At device initialization, a 3D spatial mapping phase allows the environment to become interactive, allowing players to view and interact with virtual objects projected as holograms in the room. The default interaction mechanism (air tap) built into HoloLens can be difficult to master, especially for people with motor disabilities. Therefore, we adapted our game so that the player can interact with the environment using the index finger directly to pop balloons.

2.2 Serious Game

Pop'Balloons is our first prototype of a serious game for autistic children (Supplementary Figure S1). As its name suggests, this serious game requires the player to move around the room to pop virtual balloons. Pop'Balloons was first meant to assess the player's motor skills. Therefore, it includes a "motricity" mode with four levels of progressive difficulty, as described in the Supplementary Material section. When the game is launched, the main menu allows the user to create or modify player profiles, access the game's parameters, or launch a new play session. When the last option is selected, the game starts with a countdown. A screen indicates to the player the number of balloons to burst. When the countdown is over, the first virtual balloon appears in the room. The player must then move and pop it. The next one then appears outside the player's field of vision. Arrows and spatialized sound effects guide the player in their task. There is no time limit to complete a level. However, to access the next one, all balloons must be popped. At the end of each level, a star-shaped balloon appears. The player has to pop it to get bonus points. If it does not burst after a few seconds, the balloon disappears by itself. Finally, the serious game ends when all four levels are completed.

2.3 Data collection

Preliminary data were collected from n = 10 neurotypical participants (6 males). Most of them completed the 4 levels of the game. On average, the participants played the game 3.80 times +/- 1.08 and scored 113.76 +/- 24.56. The first 6 participants played the game in a controlled condition, while the others played the game in an uncontrolled condition. The main difference between these two conditions is the environment in which the game was played. In the controlled condition, the game was played in a small empty meeting room (5m x 10m), while in the uncontrolled condition it was in an open space (hall of the student social club) with different playable areas across attempts. Importantly, participants were instructed under the exact same conditions, so that the differences in the collected data maximally reflect the variability between individuals.

2.4 Preprocessing

The trajectories of the player between each balloon were split into chunks of data for each level of the game. In total, 148 chunks of data were recorded by the MR headset, and 6 features were extracted from the motion trajectories as follows, for each chunk:

- The median of the deviation from the ideal trajectory over time. This trajectory is computed by linear interpolation between each consecutive balloon. Then, trajectory deviations are computed with the Euclidean distance between the player's trajectory and the ideal one.
- The minimum angular velocity with respect to the x-axis (pitch) in rad/s.
- The maximum angular velocity with respect to the y-axis (yaw) in rad/s.
- The median of the proportion of time spent motionless, that is, the proportion of time samples where the head of the player keeps motionless w.r.t. the whole balloon presentation. More specifically, we computed the norm of the movement at a given time, then accumulated the time spent below a criterion of 0.005 times the maximum of the movement norm across the dataset.
- The user's score, calculated by adding the points earned from popping the balloons. The maximum score given by a balloon is 35 points when it explodes in 7 seconds. After that, the score value of a balloon starts to decrease by 5 points each second. A ten-point bonus is attributed if the final bonus balloon is popped in less than 7 seconds.
- The median velocity computed in an episode using the norm speed at each time step.

2.5 Machine Learning

Several classification algorithms were trained on player identification tasks using the Scikit-Learn library [11]. A one-vs-rest (OVR) strategy was applied for algorithms that are not inherently multiclass such as SVMs. The effect of players' sex and experimental conditions was studied by splitting the dataset into 5 groups: a player group with every participant, a male group with 6 players, a female group with 4 players, a controlled group with 6 players, and an uncontrolled group with 4 participants. The classification accuracy was computed by cross-validation on 50 random subsets with 20% test data in each group. The significance of our results was calculated with a null hypothesis (H0) by shuffling participants in the player group, shuffling gender in the male/female groups, and shuffling experimental conditions in the controlled/uncontrolled groups.

3 Results

The results show that the MLP classifier was the most accurate (mean = 34.13%, std = 8.18%, H0 = 9.27%, std = 5.47%) among all models when all players were considered (Figure 1). Interestingly, accuracy was slightly improved for most classifiers when groups were separated by gender (Supplementary Figure S2), with the best accuracy achieved using the linear SVM (mean = 47.60%, std = 10.83%; H0 = 29.90%, std = 9.77%) for male players and the MLP classifier (mean = 59.09%, std = 15.45%; H0 = 24.55%, std = 12.76) for female players. Furthermore, the uncontrolled condition showed the best levels of accuracy than the controlled condition for most classifiers (Supplementary Figure S3) and was the best for the SVM linear models (mean = 71.00%, std = 15.00%; H0 = 27.20%, std = 14.15%). This could indicate that the data is of better quality when the player interacts with the game freely and in a natural environment. Given the small dataset analyzed, we cannot rule out that these results may be related to the experimental setup. These findings direct our hypothesis towards deep learning classification algorithms, although they alert to biases in the classifiers. Supplementary Tables describe all the results in detail.

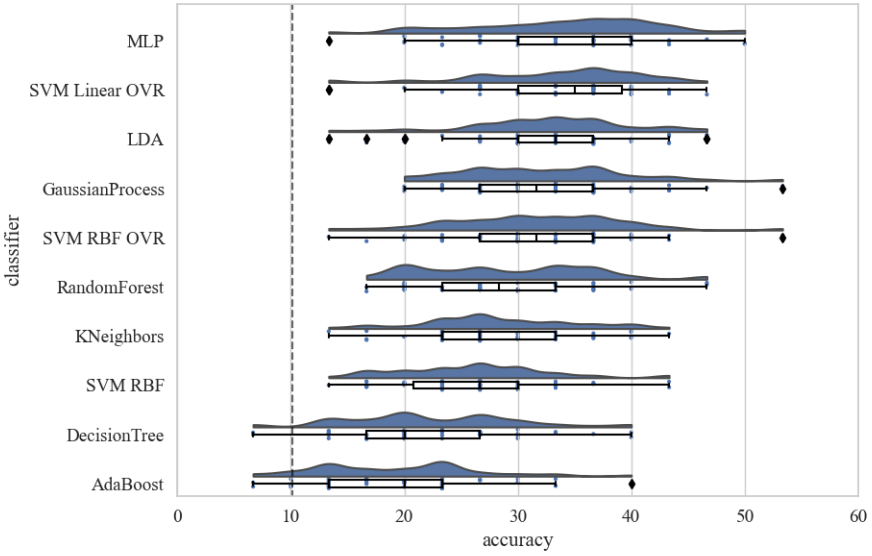


Figure 1: Classification results across all the algorithms, taking all players into accounts. The black dashed vertical line indicates chance level (i.e., H0). Accuracy is the percentage of right guesses.

4 Discussion

Our study demonstrates that serious games in MR can acquire valuable clinical data. We were able to identify the participants with high accuracy using several classification algorithms. The preliminary results reveal considerable bias related to sex and environmental context. When the groups were divided by sex, the females performed better in almost all models. Interestingly, the accuracy levels were much higher when participants were in an uncontrolled context, implying that a natural environment promotes more natural behavior. Environmental bias may also explain this trend. Although more data are needed to draw an appropriate conclusion, the results highlight the need for a more thorough understanding of the impact of these variables on classification.

To assess the usability of the device and the game for autistic children, we will collect significant amounts of data so we can generate accessibility guidelines for an inclusive design of MR applications, especially serious games. Furthermore, we will assess the psychometric characteristics of the serious game, particularly to evaluate the precision of clinical phenotyping in ecological tasks. Ultimately, this will help in the stratification of patients based on clinically interpretable characteristics across different domains (e.g., sensorimotor, cognitive, social). The logic of those subgroups of patients will then follow a more dimensional perspective on mental conditions (e.g., (RDoC) [7]) and depart from a monolithic taxonomy of mental conditions. Hopefully, addressing these theoretical questions and practical challenges could help patients and their families, as well as clinicians and researchers.

Acknowledgments

We thank the ECP students who participated in the early stages of this project: Maxime Fetiveau, Younes Laaboudi, Jules Massin, Hugo Perrin, Olivier Polidori, Seung Eun Yi, Matthieu Divet , Julien Malle , Sarah Saidani , Paul Viossat , Zairan Wang , Xavier Tinel , Adil Dinia , Marianne Clary , Clémence Kopff , Grégoire Martin, Jose Menoci Neto, Niraj Srinivas, Malek Adel, Charlotte Arnaud, Eva Chatry, Mohamed Choraichi, Clémence Giffin, Nadir Larhdir, Simon Moliner, and Zhuoying Wu. We also thank the Institut Pasteur, especially Antonio Borderia, for the initial support of the project, and Actimage, especially Thomas Klein, Jérémy Boistiere, and Alexis Giraudet, for the development of Pop'Balloons. G.D. is funded by the Institute for Data Valorization (Grant CF00137433), Montréal, and the Fonds de Recherche du Québec (Grant 285289). This research was supported by funding from Fondation Orange, MITACS, and Eidos Interactive Corp.

References

[1] B. Aruanno, F. Garzotto, E. Torelli, and F. Vona. Hololearn: Wearable mixed reality for people with neurodevelopmental disorders (ndd). In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 40–51, 2018.

[2] F. Baillin, A. Lefebvre, A. Pedoux, Y. Beauxis, D. A. Engemann, A. Maruani, F. Amsellem, J. S. Kelso, T. Bourgeron, R. Delorme, et al. Interactive psychometrics for autism with the human dynamic clamp: Interpersonal synchrony from sensorimotor to sociocognitive domains. *Frontiers in psychiatry*, 11:510366, 2020.

[3] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1):1–11, 2020.

[4] R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer. *Serious games*. Springer, 2016.

[5] G. Dumas, G. C. de Guzman, E. Tognoli, and J. S. Kelso. The human dynamic clamp as a paradigm for social interaction. *Proceedings of the National Academy of Sciences*, 111(35): E3726–E3734, 2014.

[6] S. Eldevik, R. P. Hastings, J. C. Hughes, E. Jahr, S. Eikeseth, and S. Cross. Meta-analysis of early intensive behavioral intervention for children with autism. *Journal of Clinical Child & Adolescent Psychology*, 38(3):439–450, 2009.

[7] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders, 2010.

[8] A. Karpur, V. Vasudevan, A. Lello, T. W. Frazier, and A. Shih. Food insecurity in the households of children with autism spectrum disorders and intellectual disabilities in the us: Analysis of the national survey of children's health data 2016-2018. *medRxiv*, 2021.

[9] T. A. Lavelle, M. C. Weinstein, J. P. Newhouse, K. Munir, K. A. Kuhlthau, and L. A. Prosser. Economic burden of childhood autism spectrum disorders. *Pediatrics*, 133(3):e520–e529, 2014.

[10] C. Lord, T. S. Brugha, T. Charman, J. Cusack, G. Dumas, T. Frazier, E. J. Jones, R. M. Jones, A. Pickles, M. W. State, et al. Autism spectrum disorder. *Nature reviews Disease primers*, 6(1): 1–23, 2020.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] N. Rogge and J. Janssen. The economic costs of autism spectrum disorder: A literature review. *Journal of Autism and Developmental Disorders*, 49(7):2873–2900, 2019.

[13] C. Weng, N. H. Shah, and G. Hripcsak. Deep phenotyping: embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of biomedical informatics*, 105:103433, 2020.

Appendix

Supplementary Tables

	Players	Games	Levels	Scores	Sex	Condition
0	1	5	4	140.75	Male	Controlled
1	2	6	3.5	130.83	Male	Controlled
2	3	4	4	84.37	female	Controlled
3	4	5	4	131.75	Male	Controlled
4	5	3	4	69.17	Female	Controlled
5	6	3	4	110.42	Male	Controlled
6	7	3	4	138.75	Female	Uncontrolled
7	8	3	4	107.08	Female	Uncontrolled
8	9	3	3.67	95.69	Male	Uncontrolled
9	10	3	4	128.75	Male	Uncontrolled

Supplementary Table S1: Summary of demographic and behavioral data

Classifier	mean	std	H0 mean	H0 std
MLP	34.13	8.18	9.27	5.47
SVM Linear OVR	33.47	7.33	10.27	4.66
LDA	32.87	7.06	9.53	4.32
GaussianProcess	32.13	7.17	9.80	5.47
SVM RBF OVR	31.53	7.43	9.40	5.06
RandomForest	29.00	7.52	11.13	5.52
KNeighbors	28.40	6.71	10.67	5.70
SVM RBF	26.07	6.56	10.33	5.04
DecisionTree	22.47	7.54	10.47	5.08
AdaBoost	19.40	7.11	10.33	5.47

Supplementary Table S2: Classification accuracy with all players.

Classifier	mean	std	h0 mean	h0 std
SVM Linear OVR	47.60	10.83	29.90	9.77
MLP	47.10	10.91	31.00	10.15
LDA	46.90	11.91	30.70	9.49
SVM RBF OVR	41.40	10.58	29.40	9.62
GaussianProcess	39.70	8.97	29.00	10.72
RandomForest	36.30	9.10	27.70	9.50
SVM RBF	35.60	8.81	24.50	10.16
DecisionTree	30.20	12.37	25.50	9.55
KNeighbors	29.90	9.14	26.10	10.21
AdaBoost	24.10	9.88	16.30	7.47

Supplementary Table S3: Classification accuracy with male players.

Classifier	mean	std	h0 mean	h0 std
MLP	59.09	15.45	24.55	12.76
GaussianProcess	58.36	13.98	22.18	13.00
KNeighbors	54.91	13.48	22.18	13.00
RandomForest	54.73	12.66	22.18	11.81
SVM Linear OVR	50.91	12.98	26.55	13.46
SVM RBF	50.36	13.39	19.82	11.31
DecisionTree	49.64	14.34	19.64	14.13
LDA	49.45	15.12	27.64	13.36
SVM RBF OVR	47.27	11.92	24.00	12.30
AdaBoost	37.64	17.35	15.64	9.45

Supplementary Table S4: Classification accuracy with female players.

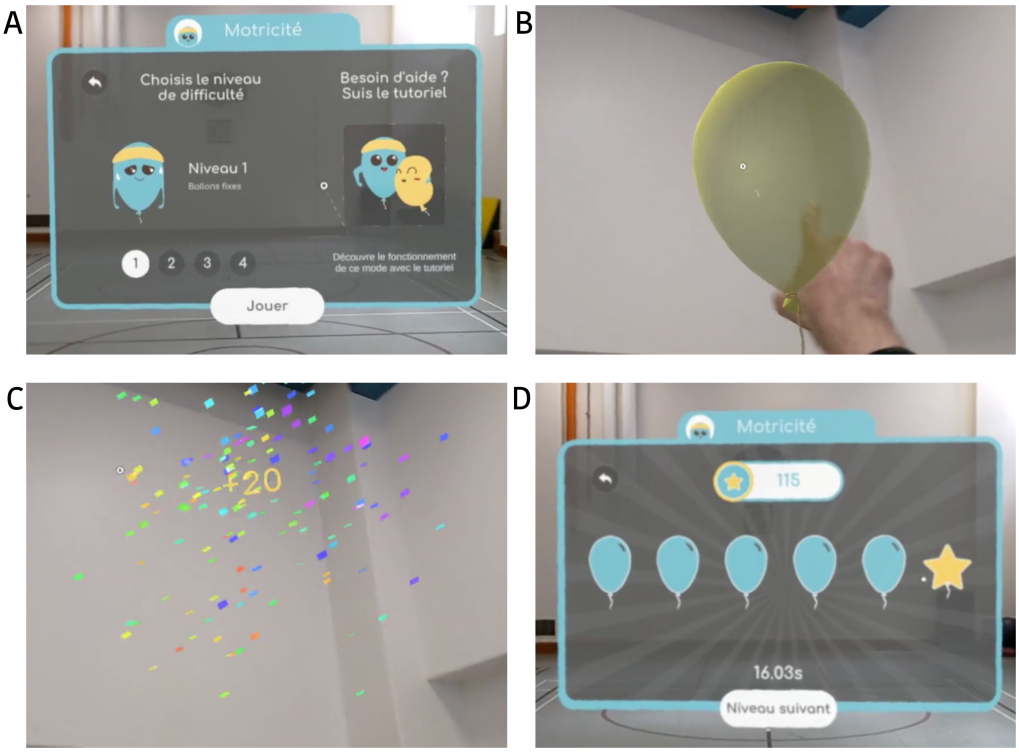
Classifier	mean	std	h0 mean	h0 std
GaussianProcess	45.14	8.90	28.95	7.55
SVM Linear OVR	44.00	9.31	29.52	9.71
MLP	43.43	8.34	30.57	8.58
SVM RBF OVR	43.24	8.51	28.48	9.21
RandomForest	41.14	8.97	24.48	7.25
LDA	40.86	9.38	29.24	8.03
DecisionTree	35.05	9.37	21.33	6.33
SVM RBF	34.86	11.08	24.10	8.95
KNeighbors	34.76	8.58	25.43	9.36
AdaBoost	32.67	10.30	16.76	6.88

Supplementary Table S5: Classification accuracy with players in the controlled condition.

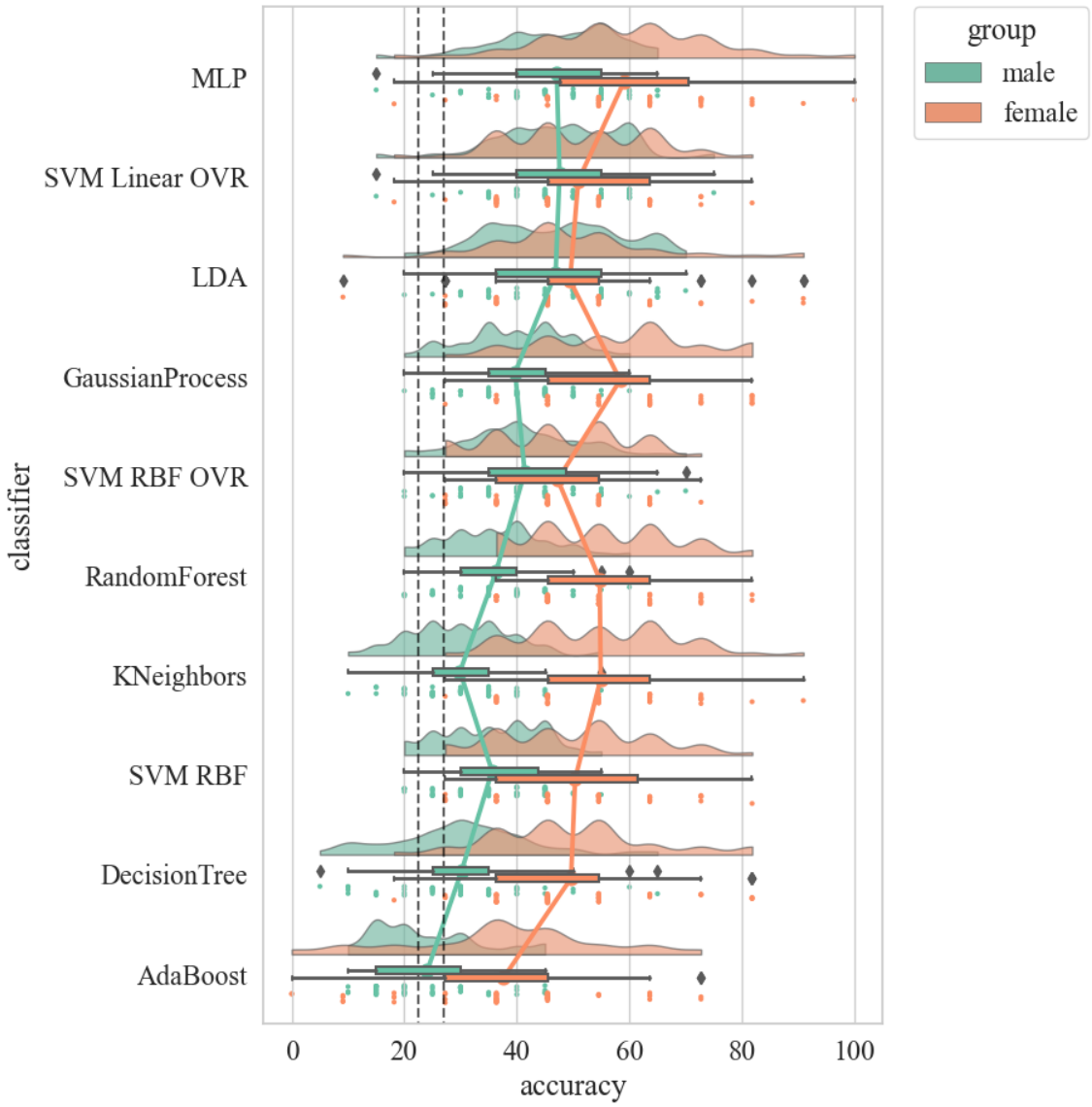
classifier	mean	std	h0 mean	h0 std
SVM Linear OVR	71.00	15.00	27.20	14.15
MLP	67.20	14.84	26.00	15.23
SVM RBF OVR	60.00	15.23	23.80	14.27
RandomForest	58.20	16.21	23.40	14.09
LDA	56.40	13.97	26.80	14.20
SVM RBF	55.40	16.27	19.40	14.62
GaussianProcess	54.20	16.74	23.80	14.13
DecisionTree	47.80	16.53	20.20	11.91
KNeighbors	45.40	14.99	20.60	14.20
AdaBoost	32.20	13.61	14.80	10.44

Supplementary Table S6: Classification accuracy with players in the uncontrolled condition.

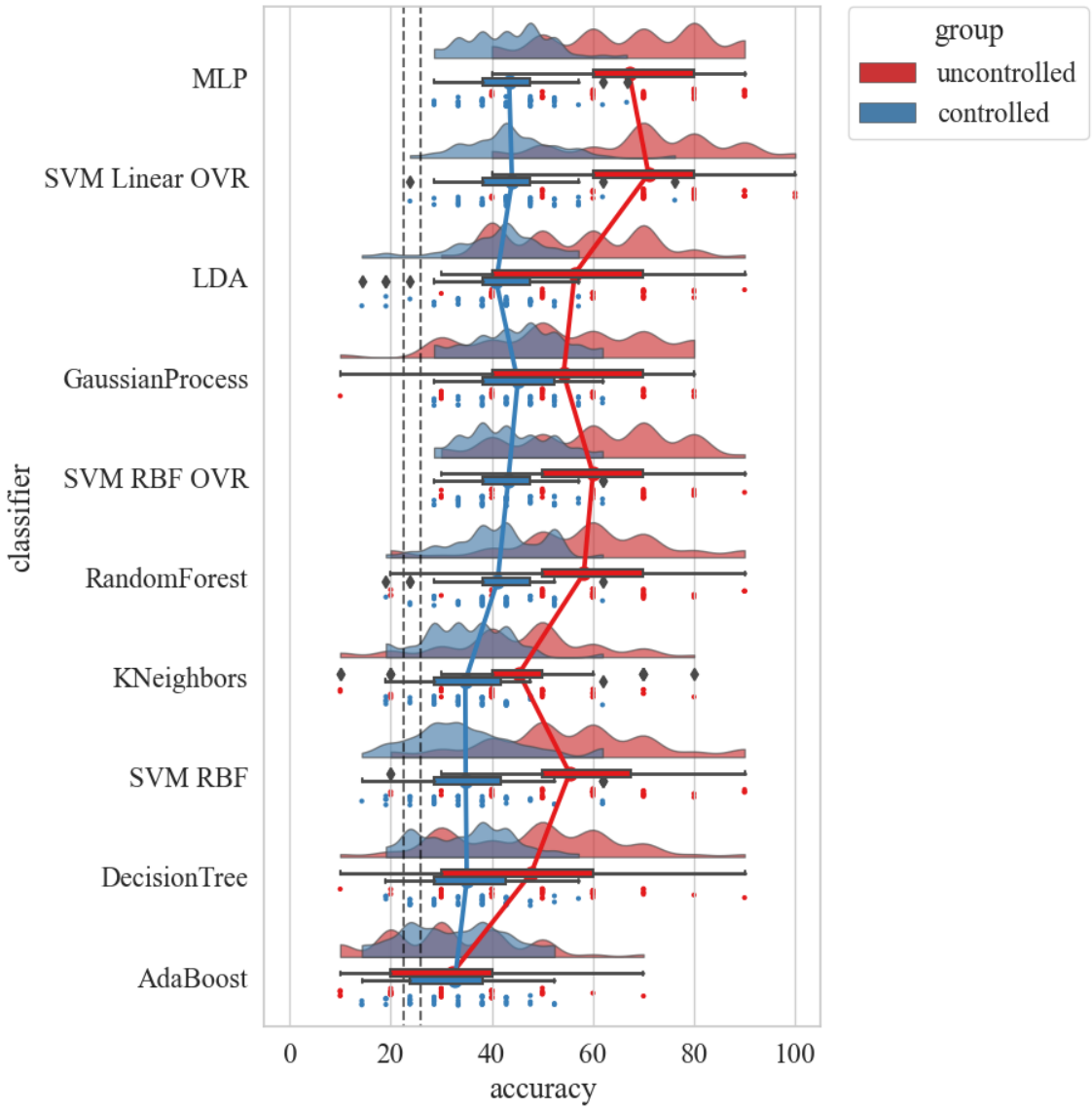
Supplementary Figures



Supplementary Figure S1: Images captured from the Pop'Balloons video games. Panel A shows a screen that allows you to choose the game level, access the tutorials, and start a new play session. The virtual ball before and after being popped is respectively shown in panels B and C. Panel D shows the results board at the end of a game, displaying the final score and which balloons were popped or missed.



Supplementary Figure S2: Classification results for female vs male players. The black dashed vertical line indicates chance level on average across classifiers for each group (i.e., H0). Accuracy is the percentage of right guesses.



Supplementary Figure S3: Classification results for players in the controlled vs uncontrolled groups. The black dashed vertical line indicates chance level on average across classifiers for each group (i.e., H0). Accuracy is the percentage of right guesses.

Supplementary Material

A Description of Levels

There are four levels in the Pop'Balloons video game:

- **Level 1:** Static balloons. They do not move.
- **Level 2:** Floating balloons. They slowly move left and right.
- **Level 3:** Falling balloons. They fall until they reach the ground. If they touch the ground, they explode and you do not earn any points.
- **Level 4 :** Fast falling balloons. They fall faster than the previous level.

Glossary

ASD Autism Spectrum Disorder. 2

HDC Human Dynamic Clamp. 2

KNeighbors K-Nearest Neighbors. 6, 7

LDA Linear Discriminant Analysis. 6, 7

MLP Multi-layer Perceptron. 4, 6, 7

MR Mixed Reality. 2–4

OVR One-vs-Rest. 3, 6, 7

RBF Radial Basis Function. 6, 7

RDoC Research Domain Criteria. 4

SVM Support Vector Machine. 3, 4, 6, 7