Applying machine learning to a virtual serious game for neuropsychological assessment.

1st Javier Marín-Morales
Instituto de Investigación e Innovación en Bioingeniería (i3B)
Universitat Politécnica de Valencia
Valencia, Spain
jamarmo@i3b.upv.es

2nd Lucia A. Carrasco-Ribelles
Instituto de Investigación e Innovación en Bioingeniería (i3B)
Universitat Politécnica de Valencia
Valencia, Spain
lucarri@i3b.upv.es

3rd Mariano Alcañiz
Instituto de Investigación e Innovación en Bioingeniería (i3B)
Universitat Politécnica de Valencia
Valencia, Spain
malcaniz@i3b.upv.es

4th Irene Alice Chicchi Giglioli
Instituto de Investigación e Innovación en Bioingeniería (i3B)
Universitat Politécnica de Valencia
Valencia, Spain
alicechicchi@i3b.upv.es

***Abstract*— Neuropsychological assessment has been traditionally made through paper-and-pencil batteries which usually are time-consuming, decontextualized, and non-ecological. These abilities play a critical role in education since they are very related to learning capacity, academic achievement, social functioning, as well as the inhibition of maladaptive behaviors. Meanwhile, serious games are being used in education and psychology to achieve assessments without these limitations, including neuropsychological assessments. While traditional tests can be analyzed with classical statistics, a large number of variables can be extracted from serious games, the analysis of which can be more complex. Machine learning can handle this large amount of information and find patterns that allow us to recognize behaviors. This study aimed to investigate whether machine learning could be used to improve predictive validity in applying a serious game for neuropsychological assessment. Results were based on 60 subjects, including 42 cognitive activities. The validation process showed best results on attention, memory, planning, and cognitive flexibility, achieving accuracies higher or equal to 0.8 and Cohen's Kappas higher than 0.55, which implies that the Virtual Serious Game could be a valid tool to perform a neuropsychological evaluation along with traditional tests.**

Keywords*— *Executive function, virtual reality, assessment.

I. Introduction

Neuropsychological assessment encloses the evaluation of individual performance in a wide range of domains, such as attention, memory, processing speed, control inhibition, planning, visual perception, control of interferences, and cognitive flexibility, to identify dysfunctions and preserved abilities. These executive functions become relevant when learning abilities [1] and they are related to academic achievement, social functioning and the inhibition of maladaptive behaviours [2]. Neuropsychological assessment includes validated paper-and-pencil batteries, e.g. the Wechsler Adult Intelligence Scale, and/or specific tests, such as the Trail Making Test to assess attention and set switching or the Tower of London for planning abilities [3, 4, 5]. The duration of the assessment can vary and cannot be foreseen.

Along with the examination variability, traditional neuropsychological assessment is showing other limitations: a) the tests are too abstracting and decontextualized, lacking of motivation, attention, alertness and stress; b) and lacking ecological validity, not reflecting behaviours in daily life activities [6]. Videogames are widespread and thanks to their playability can be also utilized for serious purposes mainly in education and psychology, providing three features that can increase the impact in neuropsychological assessment and treatment [7,8]. First, they can be internet-based and therefore they offer the opportunity to reach more people; second, they offer challenges overcoming to win, involving and engaging subjective motivational processes; and third, they provide greater ecological validity and objectivity than traditional tests [7]. Ecological validity encounters two features: a) the similarity between the test and everyday activity demand; b) and the statistical relation between standardized assessment measures and the measures assessing and predicting daily life performance [9].

The monitoring of the performance of the participant in a virtual serious game (VSG) can produce lots of information that need to be processed. In this work, the use of Machine Learning (ML) instead of classical statistics is preferred for the analysis for different reasons: a) ML makes minimal assumptions about data distribution, b) ML can understand and find non-linear relationships, c) ML can work with very numerous variables, and d) ML can extrapolate the learned knowledge from data to estimate or predict the condition of new participants [10]. The objective of this study was precisely to evaluate the predictive capability of a virtual serious game (VSG) for functional neuropsychological evaluation.

II. Materials and Method

A. Participants

A total of 60 subjects (30 women and 30 men; mean age = 35.95; SD = 11.17) participated in the study. The inclusion criteria included: a) age between 18 and 55 years; and b) a cut score higher or equal to 24 in the Mini-Mental State Examination (MMSE) [11]. Before participating in the study, each participant received written information about the study and was required to give written consent for inclusion in the study. The study obtained ethical approval by the Ethical Committee of the Polytechnic University of Valencia.

B. *Neuropsychological Assessment*

First, participants completed a socio-demographic questionnaire on age, gender, and education, and secondly, two neuropsychological batteries were administered to participants: a) an initial screening test for cognitive abilities (MMSE, [11]); b) and a subsequent extensive paper-and-pencil battery - Wechsler Adult Intelligence Scale - WAIS-IV, and Rey-Osterrieth Complex Figure Test [3, 12]. Two indexes of the WAIS-IV have been calculated using the core subtests: The Working Memory Index (Digit Span and Arithmetic) and the Processing Speed Index (Symbol Search and Coding) [3]. Furthermore, visual-spatial abilities and memory function were assessed with the Rey-Osterrieth Complex Figure Test (ROCFT) [12]. Finally, participants completed a total of 5 specific computerized tests (ST): Dot Probe Task (DOT; [13]) to assess selective attention; Go/NoGo Task to assess sustained attention and inhibition control (Fillmore et al, 2006); Stroop Test to assess selective attention, processing speed, and interference control (ST) [14]; Trail Making Task (TMTA-B) to assess visual attention and set switching [15]; and Tower of London-Drexler to assess planning abilities (TOLDX)[5]. The ST was randomly presented and performed on a personal computer.

C. *EXPANSE: the virtual serious game*

A narrative storytelling has been created and placed in a spaceship whose aim was to discover a new land because earth is no longer habitable. Various situations to solve (for example, one of the engines broke) or missions to accomplish were submitted to the participant, including 42 mini-games related to cognitive functions over mentioned.

Fig. 1. Screenshot of the virtual spaceship

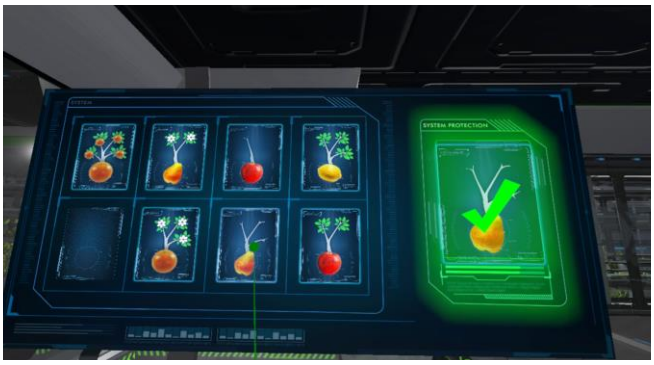

Fig. 2. Screenshot of an attention and visual perception minigame

Most of the mini-games were related to more than one cognitive function. Some examples of these mini-games can be found in Table 1. The behavioural performance data, related to the time spent on each mini-games and the hits achieved have been gathered during the gameplay. The VSG was developed using Unity 5.5.1f1 software, applying C# programing language using the Visual Studio tool.

TABLE I. *EXAMPLES OF THE MINI-GAMES AND THE EXECUTIVE FUNCTIONS THAT WERE RELATED TO THEM.*

Mini-game description	Related executive functions
Switchboard with light switches that were turning on and had to be turned off	Attention and processing speed
A puzzle with the pieces of the spaceship engine had to be solved by reconstructing it	Attention and visual perception
Ingredients fall out and have to be taken on a plate only those that appear on a recipe. The recipe is shown before you start.	Working memory
Some meteors fall at varying speeds and have to be destroyed before they reach the spaceship	Control inhibition and control of interferences

D. *Experimental procedure*

The study took place at a laboratory room and consisted of five sessions of one hour each. In the first session, before the VSG experience, participants were administered the demographic questionnaire, and the standardized neuropsychological tests over mentioned, and a tutorial to familiarize the participant with the virtual reality system. The tutorial consisted of an activity in which the participant had to handle some geometric figures, rotate them, and insert them in the right location. Each VSG session was experienced used a Head Mounted Display (HTC VIVE/Pro).

E. *Data analysis*

1) *Preliminary analysis*

First, a multivariate outlier detection was performed. The outliers within each of the 7 groups of scores (WAIS-IV, ROCFT, DOT, Go/NoGo, ST, TMTA-B, TOLDX) were identified. The Mahalanobis distance between subjects according to the scores in each group was calculated. The probability that it belongs to a Chi-square distribution was calculated, and if it was below 1%, the participant was defined as an outlier. 1 outlier subject belonged to the WAIS-IV scores, 2 to the ROCFT, 3 to the DOT, 1 to the ST, 2 to the TMTA-B, and 1 to the TOLDX.

After this analysis, all the scores were divided into “High” and “Low” score depending on if the values were above or below the median of all participants. A description of the results of this categorization can be found in Table 2. During data exploration 2 missing values were found on DOT Mean Time and DOT Hits Proportion and 9 on Go/NoGo Mean Time, and were removed from the analysis. These were due to a failure on the recording system. 8 datasets were prepared, collecting the variables from games related to the cognitive domains involved in the VSG. These datasets included: attention (289 variables), memory (165), control inhibition (75), processing speed (150), planning (132), perception (60), control of interferences (64) and cognitive flexibility (85).

TABLE II. *MEDIAN VALUE OF EACH SCALE. THE UNBALANCEMENT IS NOT TOO LARGE. MISSING VALUES AND OUTLIERS ARE ALREADY REMOVED.*

Scale	**Categorization**		
	Low	***High***	***High/Total***
DOT Trials	23	34	0.6
DOT Hits proportion	20	35	0.64
DOT Mean Time	27	28	0.51
DOT Total Time	30	27	0.47
Go/NoGo Hits proportion	28	32	0.53
Go/NoGo Mean Time	26	25	0.49
ST Mean Time	30	29	0.49
ST Hits proportion	21	38	0.64
TMTA-B Total Time A	27	31	0.53
TMTA-B Total Time B	29	29	0.50
TOLDX Excess of movements (mean)	28	31	0.53
TOLDX Mean of Execution Times and hits	28	31	0.53
TOLDX Mean of Initial Times and hits	28	31	0.53
TOLDX Total Time	28	31	0.53
TOLDX Total Time	28	31	0.53
TOLDX Total Score	22	37	0.63
WAIS Working Memory	28	31	0.53
WAIS Processing Speed	27	32	0.54
ROCFT Memory	25	33	0.57
ROCFT Copy	28	30	0.52

2) *Machine Learning*

To find the best set of mini-games-related variables able to estimate the subject’s level on each score, ML models were performed. The modelling process was the following, for each score and per each algorithm specified in Table 3: (1) Remove any subject which score was missing or considered as outliers. (2) *Feature selection*. The feature selection was performed using the wrapper method *backward sequential feature selection* [16], which starts from a model with all the variables and in each step removes the variable decreasing the performance measure the most. A maximum of 15 variables was fixed to avoid overfitting. (3) *Hyperparameter tuning*. When the set of variables was obtained, the hyperparameters (Table 3) of the algorithm were tuned. 10 equal-sized values in the range defined for each hyperparameter were tested. In the case of the SVM hyperparameters, an exponential transformation (2^x) was applied to the values obtained. (4) *Modelling*. Having the best set of variables and hyperparameters, the model was built and validated with 3 times Repeated Cross-Validation of 5 folds. The average of the metrics obtained (accuracy, Cohen’s kappa, sensibility and specificity) amongst the Repeated Cross-Validation is calculated. Both the feature selection and the hyperparameter tuning were also validated with their respective 3 times 5 folds Cross-Validation. Both the outlier detection and machine learning were performed on the software R (version 3.6.1).

TABLE III. *MACHINE LEARNING ALGORITHMS TESTED, HYPERPARAMETERS TUNED FOR EACH ONE AND RANGE OF VALUES TESTED. THE TESTED K VALUES FOR KNN ARE DEFINED HERE.*

Algorithm	**Parameter**	**Values**
Naïve Bayes	Laplace	(0,10)
Decision Tree	-	-
GLMNet	Alpha	(0,1)
SVM	C	(-10,10)
	Sigma	(-10,10)
kNN	k	(3, 5, 7, 9)

III. RESULTS

The results of the validation process of the machine learning models are shown in Table 4. All the ML algorithms tested managed to produce the best predictor for at least one of the variables, except for Decision Trees. This may be due to the simplicity of this algorithm. The dataset with the information from the attention games was the most successful in generating successful predictors of executive functions, achieving the best results in 9 of the 19 variables, followed by memory (6), planning (3) and cognitive flexibility (1). Games related to processing speed, inhibition, control of interference and perception produced worse results for all the scores.

IV. DISCUSSION

The aim of this study was to study the predictive capability of a virtual serious game for functional neuropsychological assessment through machine learning models. 60 participants took part in five sessions of around 1 hour of a VSG where they played to 42 mini-games directly related to some executive function. The information extracted from the participant's performance during the mini-games was analysed and used to train several ML models. All the models achieved accuracies higher or equal to 0.8 and Kappas higher than 0.55, which implies that the VSG could be a valid tool to perform a neuropsychological evaluation along with traditional tests. Games inside the VSG with attention and memory requirements may be the ones that contain more information to help discriminate between high and low levels on each variable from the classical neuropsychological assessment, as they are the games that have been most selected by ML models. In general, variables related to the time needed were better modelled by ML than those related to the hit ratio. This difference may be due to the fact that the classic assessment tests have generally been evaluated in terms of time required, and not in terms of hits [13-15].

V. CONCLUSIONS

Considering these positive results, it seems that this VSG could be used to make assessments of executive functions. Not only would an automatic assessment be achieved, but also a much more engaging and motivating environment, which would facilitate the assessment process. By facilitating the neuropsychological assessment, greater coverage and access to neuropsychological counselling could be provided, which could improve academic performance across the board.

TABLE IV. Best results for each score in terms of accuracy, kappa, sensibility and specificity, and with which algorithm and set of variables have been achieved.

Variable	**Dataset**	**Model**	**Features (#)**	**Accuracy**	**Kappa**	**Sensibility**	**Specificity**
DOT Trials	Attention (289)	GLMNet	14	0.81	0.61	0.84	0.79
DOT Hits proportion	Memory (165)	GLMNet	15	0.91	0.81	0.92	0.90
DOT Mean Time	Attention (289)	SVM	15	0.92	0.83	0.90	0.93
DOT Total Time	Memory (165)	kNN	13	0.87	0.73	0.84	0.90
Go/NoGo Hits proportion	Planning (132)	GLMNet	13	0.80	0.61	0.82	0.81
Go/NoGo Mean Time	Attention (289)	SVM	14	0.94	0.88	0.95	0.94
ST Mean Time	Memory (165)	GLMNet	15	0.91	0.82	0.91	0.92
ST Hits proportion	Attention (289)	kNN	13	0.84	0.66	0.90	0.77
TMTA-B Total Time A	Attention (289)	GLMNet	12	0.90	0.79	0.89	0.90
TMTA-B Total Time B	Attention (289)	SVM	15	0.85	0.70	0.85	0.88
TOLDX Excess of movements (mean)	Memory (165)	kNN	15	0.83	0.67	0.82	0.87
TOLDX Mean of Execution Times and hits	Planning (132)	GLMNet	12	0.54	0.68	0.86	0.83
TOLDX Mean of Initial Times and hits	Planning (132)	GLMNet	14	0.84	0.69	0.85	0.84
TOLDX Total Time	Attention (289)	SVM	14	0.85	0.70	0.84	0.85
TOLDX Total Score	Memory (165)	GLMNet	14	0.92	0.82	0.97	0.86
WAIS Working Memory	Attention (289)	kNN	15	0.82	0.64	0.88	0.78
WAIS Processing Speed	Attention (289)	kNN	15	0.80	0.59	0.74	0.87
ROCFT Memory	Flexibility (85)	Naïve Bayes	15	0.87	0.74	0.89	0.85
ROCFT Copy	Memory (165)	GLMNet	15	0.82	0.62	0.86	0.78

ACKNOWLEDGMENT

This work was supported by the Spanish Ministry of Economy, Industry and Competitiveness funded project "Advanced Therapeutically Tools for Mental Health" (DPI2016-77396-R) and by the European Union through the Operational Program of the European Regional Development Fund (ERDF) on the Valencian Community 2010-2020 (IDIFEDER/2018/029).

REFERENCES

[1] Altemeier, L., Jones, J., Abbott, R. D., & Berninger, V. W., Executive functions in becoming writing readers and reading writers: Note taking and report writing in third and fifth graders. Developmental neuropsychology, 29(1), 2006, 161-173.

[2] Hofmann, W., Schmeichel, B. J., & Baddeley, A. D., Executive functions and self-regulation. Trends in cognitive sciences, 16(3), 2012, 174-180.

[3] Wechsler, D, (2008a-b), WAIS-IV administration and scoring manual (Canadian). The Psychological Corporation, San Antonio, TX.

[4] Reitan, RM, (1958), Validity of the Trail Making Test as an Indicator of Organic Brain Damage. Perceptual and Motor Skills, 8, 3, pp. 271 - 276.

[5] Culbertson, W, Zillmer, E, (1999), Tower of London Drexel University, examiner's manual resarch version. Multi-Health Systems, Toronto.

[6] Valladares-Rodríguez, S, Pérez-Rodríguez, R, Anido-Rifón, L, Fernández-Iglesias, M, (2016), Trends on the application of serious games to neuropsychological evaluation: A scoping review. Journal of Biomedical Informatics, 64, 296 - 319.

[7] Fleming, TM, Bavin, L, Stasiak, K, Hermansson-Webb, E, Merry, SN, Cheek, C, Lucassen, M, Lau, HM, Pollmuller, B, Hetrick, S, 2017, Serious Games and Gamification for Mental Health: Current Status and Promising Directions. *Frontiers in Psychiatry*, **7**, pp. 215 - 215.

[8] Rosa, PJ, Sousa, C, Faustino, B, Feiteira, F, Oliveira, J, Lopes, P, Morais, D, (2016), The effect of virtual reality-based serious games in cognitive interventions: a meta-analysis study. Proceedings of the 4th Workshop on ICTs for improving Patients Rehabilitation Research Techniques, pp. 113 -116.

[9] Chaytor, N, & Schmitter-Edgecombe, M, (2003), The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology review*, **13**, *4*, pp. 181-197.

[10] Bzdok, D., Altman, N., & Krzywinski, M. (2018) "Statistics versus machine learning". *Nature Methods*, 15, 233-234.

[11] Folstein, MF, Folstein, SE, Mchugh, PR, (1975), Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, **12**, *3*, pp. 189 - 198.

[12] Shin, MS, Park, SY, Park, SR, Seol, SH, Kwon, JS, (2006), Clinical and empirical applications of the Rey-Osterrieth complex figure test. Nature protocols, 1, 2, pp. 892 – 892.

[13] Miller, MA, Fillmore, MT, (2010), The effect of image complexity on attentional bias towards alcohol-related images in adult drinkers. Addiction, 105, 5, pp. 883 – 890.

[14] Stroop, JR, (1935), Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 6, pp. 643–662.

[15] Reitan, RM, (1958), Validity of the Trail Making Test as an Indicator of Organic Brain Damage. Perceptual and Motor Skills, 8, 3, pp. 271 - 276.

[16] Doak, J. (1992). An evaluation of feature selection methods and their application to computer security. *Techninal Report CSE-92-18*