

A Machine Learning Approach to the Early Diagnosis of Alzheimer's Disease Based on an Ensemble of Classifiers

Sonia Valladares-Rodríguez[1], Luis Anido-Rifón[1(✉)], Manuel J. Fernández-Iglesias[1], and David Facal-Mayo[2]

[1] Department of Telematics Engineering, University of Vigo, Vigo, Spain
lanido@gist.uvigo.es
[2] Unit of Psychogerontology, University of Santiago de Compostela, Santiago de Compostela, Spain

Abstract. The present article discusses the ability of an ensemble of machine learning models to implement classification strategies to discriminate among mild cognitive impairment (MCI), Alzheimer's disease (AD) and Cognitive Unimpaired (CU). For this, a battery of games that assesses the most relevant markers for the early diagnosis of cognitive impairment (i.e., memory, executive functions, attention and gnosias) was implemented, and a pilot study was carried out with 64 individuals (28 CU, 16 MCI and 20 AD). Participants were administered a collection of classical pen-and-paper tests, and then interacted with all games in the battery. A set of four ensembles of classifiers were applied and relevant metrics were computed to assess the classification power of this approach. According to the classification metrics computed, best classification results are obtained using the Random Forest Classifier algorithm, with average F1 score = 0.99, accuracy = 1.00, and Cohen's kappa = 0.98. The experiments performed indicate that artificial intelligence techniques and serious games can be used to automate some aspects of the clinical diagnosis of individuals with cognitive impairment. However, more research is needed to obtain the required normative data for clinical validity.

Keywords: Machine learning · Ensemble methods · Alzheimer's disease · Early detection · Cognitive impairment · Serious games

1 Introduction

There were an estimated 46.8 million people worldwide living with dementia in 2015, and this number was close to 50 million in 2017. This figure will almost double every 20 years, reaching 75 million in 2030 and 131.5 million in 2050 [1]. Particularly, Alzheimer's disease is the most common cause of dementia, as it accounts for 60% to 80% of all cases. It is considered a slowly progressive brain disease that begins well before clinical symptoms emerge. Alzheimer (AD) is one of the most prevalent diseases and one of the most relevant health threats worldwide [2].

Diagnosing AD requires a careful and comprehensive medical evaluation, including an examination of the person's physical and mental status. Focusing in cognitive

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11619, pp. 383–396, 2019.
https://doi.org/10.1007/978-3-030-24289-3_28

markers, state-of-the-art diagnostic procedures for AD are based on neuropsychological paper-and-pencil tests [3, 4]. These tests have some relevant limitations, as they are perceived as intrusive [5] and influenced by the white-coat effect (WCE) [6]; they provide a late diagnosis [7]; they lack ecological validity [8]; and they are strongly dependent on confounding factors (e.g., age, educational level [9], practice effect [3]), and prone to processing errors due to their manual processing.

Due to the need of alternative mechanisms supporting an early diagnosis, we can find in the scientific literature several proposals such as the digitalization of classical tests, the introduction of game-inspired elements (e.g., rewards, challenges, simulated environments) and the implementation of technology-based solutions such as immersive 3D environments or online interactive software applications, among others.

The proposal discussed in this paper relies on the use of gamification techniques and machine learning, and the introduction of digital touch devices. More specifically, this paper discusses the ability of a digital battery composed of 7 games to assess the cognitive status in a non-intrusive and ecological way [10–12]. Cognitive areas tackled are episodic memory, executive functions, attention, semantic memory, working memory, procedural memory and gnosias. These cognitive areas, and specially the three first ones, are early markers of cognitive alterations with a relevant diagnostic value for mild cognitive impairment (MCI) and AD [13, 14].

To support the diagnosis of cognitive impairment, traditional analysis procedures rely on survival analysis [15], polynomial regression analysis [16], multinomial logistic regression analysis [17] and Markov models [13]. These techniques' performance could be improved thorough machine learning (ML), a scientific discipline that focuses on how computers learn from data [18] to build classifiers by automatically learning the inherent structure of a data set. In this study, we focused on supervised learning ML, whose initial goal is the prediction of a known output or target. This innovative technique has a great potential for supporting diagnosis prediction, risk estimation, response prediction and classification learning for diagnosis [19–22]. These techniques, and more specifically ensemble methods such as Random Forest (RF), Extra Trees classifier (ET) or boosting classifiers, are used for the automated classification of medical datasets (e.g., MRI scans, PET images, biomarkers, etc.) as either Alzheimer's disease or healthy controls [23]. Such techniques do not rely on a single cognitive area of interest, which may result in low specificity and sensitivity due to higher inter-subject variability of the aforementioned disease. On the contrary, these techniques analyse different biomarkers or cognitive constructs, resulting in higher discriminative power. They are discussed in this paper as a means to automate some aspects of the clinical diagnosis of individuals with cognitive impairment.

Section 2 introduces the dataset used and the experiment performed, while Sect. 3 presents the results obtained and Sect. 4 discusses these results and their impact in the early diagnosis of MCI and AD. Finally, the conclusions of this research work and possible future directions are proposed.

2 Materials and Methods

2.1 Data and Instruments

Participants

This study was carried out from a cross-sectional sample of 64 senior adults over 55 years old (range = 57–95; average = 77.03; SD = 7.23, cf. Table 1). All participants were recruited in the Pontevedra province of Galicia, Spain, with the support of the associations of relatives of Alzheimer's patients in the area. The basic inclusion criterion was being 55+ years old, and exclusion criteria included an advanced cognitive impairment, severe motor, hearing or visual disability, and technophobia. Participants should attend a socio-cognitive workshop in any of the collaborating associations, and no previous educational level or technological skills were required. Among the 64 participants, 20 subjects were previously diagnosed with AD, 16 participants had MCI, and 28 non-cognitively-impaired subjects served as controls. This latter group is referred as the CU group in the rest of this paper.

Table 1. Participants' characteristics and their distribution according to the Galician population

Variable		Sample n (%)	Galician population[a] n (%)
Gender	Female	42 (65.63)	571277 (55.4)
	Male	22 (34.38)	459914 (44.6)
		Mean (SD)	
Age (55 + years)	n = 64	77.01 (7.23)	n > 62, confidence level = 95%, error = 7.5% according Galician population[a] and stratified proportional sampling [24]
Educational level		1.84 (1.2)	–
Exercise level		3.64 (0.99)	–
Socialize level		3.28 (1.05)	–
Chronic treatment		0.84 (0.36)	–

Note: Sociodemographic characteristics: gender, age, educational level (i.e., 0 = illiterate; 1 = ability to read and write; 2 = primary school; 3 = secondary school; 4 = high school; 5: vocational training; 6 = university), exercise level and socialization level, all of them based on a 5-point Likert scale: 1(never) to 5 (always); and finally, chronic treatment (i.e., 0 = no; 1 = yes).
[a]Galician population of people aged 55 years and over, according to a study by the Galician Institute of Statistics (2017).

The study design was approved by the Galician institutional ethics committee (registration code 2016/477)) and was conducted in accordance with the provisions of the Declaration of Helsinki, as revised in Seoul 2008. Before their involvement in the pilot project, all subjects read and understood patient information sheets and gave their written informed consent to use the anonymized data gathered.

Neuropsychological Assessment

Participants underwent a demographic and health interview to gain some insight on their quality of life and cognition levels [25] prior to the experiment. They also completed an extensive neuropsychological and cognitive assessment, including the MMSE – Mini-metal Examination State test [26], the Spanish version of CVLT - California Verbal Learning Test [27], AD8 - Adapted Dementia Screening Interview, ADL - Barthel scale of activities of daily living, and a questionnaire about memory complaints. Data gathered using these tests were used as golden standard data and also to initially discriminate subjects suffering MCI from healthy controls. According to standard diagnostic criteria, MCI would likely be present in participants that would present normal cognitive functions and their capabilities to carry out everyday activities were not compromised, they would refer memory complaints, they would obtain a low score in the Spanish version of CVLT according to subject's age and educational, and a level equal or greater than 2 in the AD8 test. Participants who suffered AD—facilitated by participating organizations with a prior medical diagnosis—were not asked to take the Spanish version of CVLT to avoid frustration.

Game Assessment

Data acquisition in order to detect MCI or AD was carried out through specifically designed digital games running in touch devices. All games are based on the gamification of existing neurophysiological tests [28–30], and were designed according to a methodology based on participatory design focused on the end user. In other words, the vision of senior adults captured in participatory design workshops and focus groups [31, 32] drove the design process from a technical and aesthetical perspective. Besides, experts in neurology, neuropsychology and geriatrics, and participating patients' associations provided the expertise required to ensure the validity of the new tool for cognitive assessment. The collaboration with health professionals [33] from an early design stage was deemed necessary to guarantee both content and clinical validity.

The battery of games, named Panoramix [10–12], is composed of seven serious games targeting seven different cognitive areas: (1) Episodix is based on the gamification of the California Verbal Test (CLVT), which is widely used to assess episodic memory; (2) Attentix evaluates attention capabilities of player; (3) Executix addresses executive functions and it is based on the gamification of the Tower of Hanoi (TOT) game [34]; (4) Workix, which is based on the Corsi Cubes Test [30], evaluates working memory; (5) Semantix is based on the gamification of the Pyramids and Palm-trees test [35] to evaluate semantic memory; (6) Prospectix addresses prospective memory through the Pursuit Rotor Task test [36]; and finally (7) Gnosix focuses on visual gnosias, that is, the ability to visually recognize different elements and assigning meaning to them (Fig. 1).

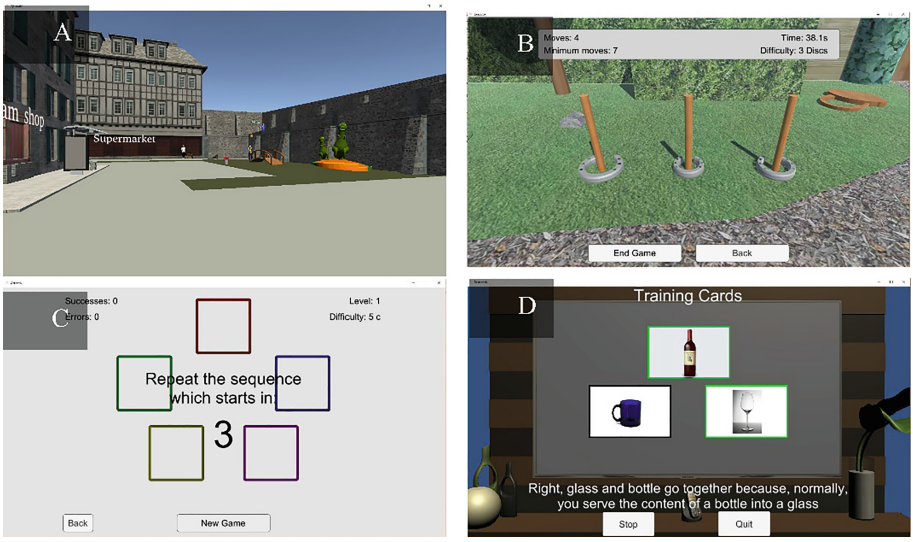

Fig. 1. Some captures of Panoramix: A: Episodix; B: Executix; C: Attentix and D: Semantix

2.2 Organisation of the Experiment

The proposed approach consisted of three phases, namely data collection, pre-processing, and classification.

1. **Data Collection**

The procedure to gather data from participants, based on game assessment, consisted on three consecutive stages:

1. Firstly, subjects would play the first part of Episodix, which targets immediate and short-term memory.
2. Secondly, they would play two additional games in the battery (i.e., Executix, Attentix, Procedurix, Semantix, Workix or Gnosix).
3. Finally, participants would play the second part of Episodix to assess long-term memory and their recognition capabilities.

Note that game playing for data collection [10] emulates the administration of CVLT, only that using serious games instead of pen-and-paper tests, and breaks being used to evaluate other cognitive markers. Moreover, data collection was carried out at the same time and location as the compulsory socio-cognitive workshops to guarantee the ecological validity of neuropsychological evaluation by integrating the latter in participants' daily routines. This was particularly relevant for patients with cognitive impairment due to the high prevalence of the white-coat effect [6].

2. **Pre-processing**

Data analytics was performance under a Python ecosystem and using the Scikit-Learn machine learning library [37]. Before training and testing data with the ML

ensemble classifiers, features were pre-processed by merging the datasets from the seven games, and cleaning, normalizing and preparing the complete dataset. Missing values due to the different difficulty levels achieved by different participants were addressed by means of substitution by cognitive group (i.e., AD, MCI or HC). This popular interpolation method replaces the missing data for a given attribute by the mean of all known values of that attribute, for the different cognitive groups [38, 39].

3. **Classification**

Regarding this phase, we selected popular supervised learning techniques in medical research [19–21]. More specifically, we used the ensemble methods enumerated below:

- Random Forest classifier (i.e., RF) [40]. This collection of algorithms, also known as random decision forest, is a perturb-and-combine technique specifically designed for decision trees where random alterations are introduced in the learning process. It provides better prediction capabilities than other decision tree-based methods, and thee most common criterion employed for classification is information gain; namely, at each split, the decrease in entropy due to this split is maximized.
- Extra Trees classifier (i.e., ET) [41]. This ensemble of classifiers is based on unpruned top-down decision trees and also on the perturb-and-combine strategy. They are typically used to enhance precision in predictions provided by traditional decision trees. The ET algorithm builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure.
- Gradient Boosting classifier (i.e., GB) [42]. A classification technique based on decision trees, which produces a prediction solution in the form of an ensemble of weak prediction models. It optimizes a cost function by iteratively choosing a function—weak hypothesis—that points towards the negative gradient direction.
- Ada Boost (i.e., AB) [43]. This machine learning meta-algorithm, is adaptive in the sense that subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. AB maintains a set of weights over the original training set and adjusts these weights after each classifier is learned by the base learning algorithm. This algorithm is sensitive to noisy data and outliers, and also, it tends not to over-fit.

For each of the previous algorithms, metrics were computed based on the four possible results of the classification of participants according to their cognitive status, that is, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). "Positive" means that the subject was classified in that cognitive category, while "Negative" means that the subject was not classified in that category. "True" means that the classification was correct, and "False" means that it was erroneous. From these results, the following values are defined:

- F1-score, which represents a weighted harmonic average of precision and recall.

$$F1(F1\ score) = 2 * \frac{P * R}{P + R}$$

- Accuracy, a combined value of precision and specificity.

$$Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)}$$

- The Cohen's kappa coefficient (K) [44] was included as a robust measure of the concordance of the observed participants' classification for each algorithm along the experiments performed. In other words, it provides a measure of the validity of the classification when confronted with a random classification. K value ranges are interpreted according to an existing consensus [45]: 0-0.20 Insignificant (i.e., ML-based classification would not be distinguishable from random classification); 0.21–0.40 Median; 0.41–0.60 Moderate; 0.61–0.80 Substantial; and 0.81-1.00 Almost perfect.
- Sensitivity or true positive rate:

$$Sensitivity\ (TPR) = \frac{TP}{(TP + FN)}$$

- Specificity or true negative rate:

$$Specificity\ (TNR) = \frac{TN}{(TN + FP)}$$

For the above ML algorithms, we plotted the Receiver Operating Characteristic (ROC) curve, which represents sensitivity as a function of fall-out, and is widely used in the diagnostic field.

Finally, the normalized confusion matrix, will be also plotted, as useful graphs in diagnosis-related fields [46].

Metrics above are utilized as macro values, as they are the best metrics to show the performance of classifiers on non-uniformly distributed datasets with respect to the categories considered. Furthermore, a 10-fold cross-validation strategy [47] was adopted to train and evaluate the proposed classifiers in order to prevent overfitting due to the use of the same data for the training and testing of classifiers. Data analytics in all cases was carried out in a Jupyter Notebook/Python ecosystem under the machine learning library Scikit-Learn [37].

3 Results

All participants played all the games in the Panoramix battery twice. Obtained datasets were merged and processed with the machine learning algorithms above, that is, Random forest (RF); Extra Trees classifier (ET); Gradient Boosting classifier (GB); and Ada Boost classifier (AB). Then, with regard to predicted validity, the metrics in the previous section were computed to estimate the classification abilities by cognitive group (cf. Table 2):

Table 2. Metrics about predicted validity of Panoramix (dataset = 7-games).

	Metrics		
Ensemble methods	F1-score	Accuracy	Cohen's kappa ↓
Random Forest (RF)	0.99	1.00	0.98
Extra Tree (ET)	0.97	1.00	0.97
Gradient Boosting (GB)	0.97	1.00	0.97
Ada Boost (AB)	0.96	0.99	0.96

- F1 score: average values obtained were 0.99 for RF; 0.97 for ET; 0.97 for GB; and 0.96 for AB.
- Accuracy: the average value obtained for correctly classified subjects was the maximum value (i.e., Accuracy = 1) for RF, ET and BG; and slightly lower, 0.99 for AB.
- Cohen's kappa: the average value obtained for the concordance of the observed participants' classification for each algorithm was 0.98 for RF; 0.97 for ET and GB; and finally, 0.96 for GB.

Experiments were performances with a 10-fold cross validation.

3.1 Specificity or True Negative Rate

In relation to the ability of Panoramix to correctly reject cognitively-unimpaired subjects, that is, the specificity of Panoramix, was also computed for each ML algorithm and cognitive group (cf. Table 3). On the one hand, specificity value for CU controls and also for AD group, is the maximum value, 1.00, for all classifiers. On the other hand, in the case of participants affected by MCI, the specificity score is lower, 0.7 for the four ML algorithms.

Table 3. Specificity or true negative rate of Panoramix (detailed by cognitive group)

Specificity (TNR)	CU	MCI	AD
Random Forest (RF)	1.00	0.7	1.00
Extra Tree (ET)	1.00	0.7	1.00
Gradient Boosting (GB)	1.00	0.7	1.00
Ada Boost (AB)	1.00	0.7	1.00

3.2 Sensitivity or True Positive Rate

In relation to the capabilities of Panoramix to correctly detect subjects with cognitive impairment (i.e., MCI or AD subjects), namely, the true positive rate, the following results were obtained (cf. Table 4). Firstly, in the case to detect participants with AD, the sensitivity score is above 0.92 for the all ML classifiers, achieving the best performance for RF (i.e., 1.00). On the other hand, the best sensitivity performance of Panoramix tool is for RF classifier (i.e., 1.00) for all participants, regardless of their

cognitive group. In the case of subject with MCI, the sensitivity score achieved the unit for all ML ensemble algoritms.

Table 4. Sensitivity or true positive rate of Panoramix (detailed by cognitive group)

Sensitivity (TPR)	CU	MCI	AD
Random Forest (RF)	1.00	1.00	1.00
Extra Tree (ET)	0.98	1.00	0.97
Gradient Boosting (GB)	0.98	1.00	0.95
Ada Boost (AB)	0.98	1.00	0.92

3.3 ROC Curve

The ROC curve was computed to assess the classifier's performance, detailed by cognitive class, as depicted in the Fig. 2. Particularly, it shows a high average AUC for each cognitive class and for all Ml algorithms, above 0.95. Best performance is for subjects with AD, (i.e., AUC = 0.98–0.99) for all participants. In the case of CU or controls subjects the value of AUC = 0.98, again regardless for each algorithm. Finally, in the case of MCI group, the AUC score slightly slow, 0.96 for RF, ET and GB, while 0.96 for AB.

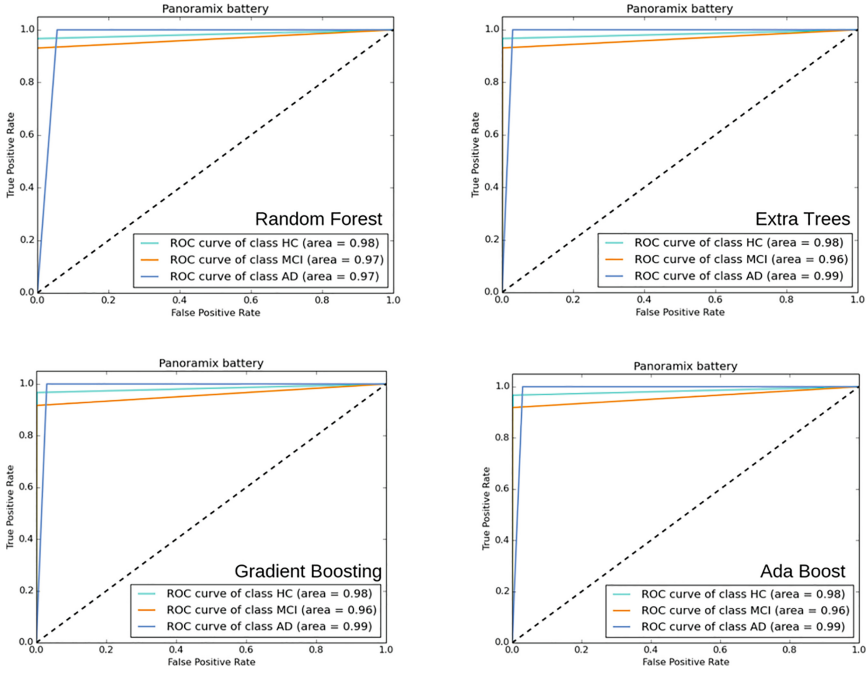

Fig. 2. ROC curve to evaluate ML classifiers' performance detailed by cognitive class

3.4 Confusion Matrix

In the same line as the ROC, the normalized confusion matrix shows scores very close to 1 for all cognitive classes (cf. Fig. 3). Particularly, the ratio between true and predicted labels is 1 for the AD group and for all ML classifiers. In the case of controls or CU group, this rate achieves a value of 0.97. Finally, this score is lower for the participants with MCI: 0.95 for RF; 0.93 for ET; 0.92 for GB; and slightly worse for AB, 0.89.

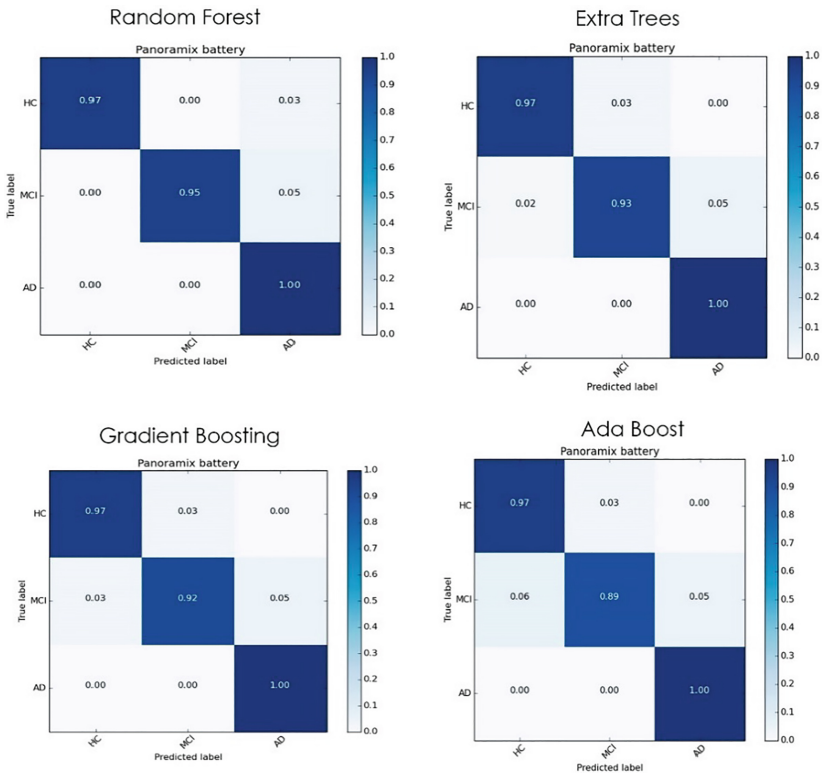

Fig. 3. Matrix confusion to evaluate ML classifiers' performance detailed by cognitive class

To sum up, according to the performance of the metrics above, the best prediction abilities with regard to cognitive status (i.e., CU, MCI or AD) were obtained for the Random Forest classifier (RF).

4 Discussion and Conclusion

The purpose of this study was to evaluate four ensemble machine learning models (i.e., RF, ET, GB, and AB), to construct classification models to discriminate among cognitively healthy individuals (i.e., CU group), participants with MCI and people affected by AD. For this, we used the seven cognitive games in the Panoramix battery [10–12] by means of a pilot experiment with statistical significance for the target population considered. In this study, we applied the classifiers discussed to the complete dataset, to gather more information about the predictive validity of the games and to obtain the best ML configuration to prepare a prototype to be eventually introduced in clinical settings.

Based on the results, the classification performance of four ensembles of classifiers was analysed using a 10-fold stratified cross validation procedure [44]. Best results were obtained for the Random Forest method, in line with previous supervised classification studies targeting the detection of cancer [46] and dementia [47]. In our case, accuracy is the index parameter, with average values Accuracy = 1.00, F1-score = 0.99 and Cohen's kappa = 0.98. These results indicate that Panoramix is a high-quality tool to predict cognitive impairment, both according to the rate of correctly classified subjects (i.e., Accuracy), as well as from the point of view of (the harmonic average of) precision and sensitivity (i.e., F1 score). Moreover, findings are robust when compared to a random prediction, as the values of Cohen's kappa obtained are almost perfect [41] for all classifiers, above 0.96.

Regarding the ability of the game battery to correctly reject cognitively-unimpaired subjects, namely, the specificity of Panoramix, excellent results were obtained with all the algorithms (i.e., true negative rate above 0.97 for all cognitive classes). A sensitivity study was also carried out to assess the ability of Panoramix to correctly detect subjects with cognitive impairment, MCI or AD. In this case, the RF classifier performs with practically full sensitivity for all cognitive groups (i.e., 1.00 regardless cognitive class).

In relation to the limitations of this study, it is worth mentioning that for the results to be soundly considered as highly representative for the Western European population and to guarantee clinical normative validity, a larger trial, including subjects adequately distributed according to age and gender groups, and involving at least two countries to consider cultural or linguistic factors, would be necessary. However, the results discussed in this paper are highly promising and already have statistical significance for Galicia. Moreover, despite of applying mean imputation to handle missing data, which may contribute to underrepresent the variability in the data, the impact in actual data sets is limited because this value was computed by cognitive group and not for all subjects.

As it was demonstrated in this paper, ensemble ML methods support the construction of classification models for the discrimination among CU, MCI and AD. Besides, these classifiers' performance is clearly better in comparison to other popular approaches (e.g., Support Vector Machines), without affecting output metrics.

Thus, the experiments reported indicate that artificial intelligence techniques can be used to automate aspects of clinical diagnosis, using the data obtained from a battery of serious games. Furthermore, discrimination was performed in a nonintrusive, non-frustrating way, which is especially relevant when targeting elder adults that may suffer cognitive impairments.

Acknowledgment. This work has been partially supported by 'Rede Galega de Investigación en Demencias' (IN607C-2017/02) funded by Axencia Galega de Innovación GAIN – Xunta de Galicia and 'Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019) funded by Xunta de Galicia and the European Regional Development Fund – ERDF.

References

1. Prince, M.J.: World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends. Alzheimer's Disease International (2015)
2. WHO: First WHO ministerial conference on global action against dementia: meeting report, WHO Headquarters, Geneva, Switzerland, 16–17 March 2015
3. Lezak, M.D.: Neuropsychological Assessment. Oxford University Press, Oxford (2004)
4. Howieson, D.B., Lezak, M.D.: The neuropsychological evaluation. In: Essentials of Neuropsychiatry and Behavioral Neurosciences, pp 29–46. American Psychiatric Pub (2010)
5. Chaytor, N., Schmitter-Edgecombe, M.: The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. Neuropsychol. Rev. **13**, 181–197 (2003)
6. Mario, B., Massimiliano, M., Chiara, M., et al.: White-coat effect among older patients with suspected cognitive impairment: prevalence and clinical implications. Int. J. Geriatr. Psychiatry **24**, 509–517 (2009)
7. Holtzman, D.M., Morris, J.C., Goate, A.M.: Alzheimer's disease: the challenge of the second century. Sci. Transl. Med. **3**, 77sr1 (2011)
8. Knight, R.G., Titov, N.: Use of virtual reality tasks to assess prospective memory: applicability and evidence. Brain Impair. **10**, 3–13 (2009)
9. Cordell, C.B., Borson, S., Boustani, M., et al.: Alzheimer's association recommendations for operationalizing the detection of cognitive impairment during the medicare annual wellness visit in a primary care setting. Alzheimer's Dement **9**, 141–150 (2013)
10. Valladares-Rodriguez, S., Perez-Rodriguez, R., Facal, D., et al.: Design process and preliminary psychometric study of a video game to detect cognitive impairment in senior adults. PeerJ **5**, e3508 (2017). https://doi.org/10.7717/peerj.3508
11. Valladares-Rodriguez, S., Pérez-Rodriguez, R., Fernandez-Iglesias, J.M., et al.: Learning to detect cognitive impairment through digital games and machine learning techniques. Methods Inf. Med. **57**, 197 (2018)
12. Valladares-Rodriguez, S., Fernández-Iglesias, M.J., Anido-Rifón, L., et al.: Episodix: a serious game to detect cognitive impairment in senior adults. A psychometric study. PeerJ **6**, e5478 (2018). https://doi.org/10.7717/peerj.5478
13. Facal, D., Guàrdia-Olmos, J., Juncos-Rabadán, O.: Diagnostic transitions in mild cognitive impairment by use of simple Markov models. Int. J. Geriatr. Psychiatry **30**, 669–676 (2015). https://doi.org/10.1002/gps.4197
14. Juncos-Rabadán, O., Pereiro, A.X., Facal, D., et al.: Prevalence and correlates of cognitive impairment in adults with subjective memory complaints in primary care centres. Dement. Geriatr. Cogn. Disord. **33**, 226–232 (2012)
15. Ding, D., Zhao, Q., Guo, Q., et al.: Progression and predictors of mild cognitive impairment in Chinese elderly: a prospective follow-up in the Shanghai Aging Study. Alzheimer's Dement. Diagn. Assess Dis. Monit. **4**, 28–36 (2016)
16. Cloutier, S., Chertkow, H., Kergoat, M.-J., et al.: Patterns of cognitive decline prior to dementia in persons with mild cognitive impairment. J Alzheimer's Dis. **47**, 901–913 (2015)

17. Han, J.W., Kim, T.H., Lee, S.B., et al.: Predictive validity and diagnostic stability of mild cognitive impairment subtypes. Alzheimer's Dement. **8**, 553–559 (2012)
18. Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.-T.: Learning from Data. AMLBook, New York (2012)
19. Lehmann, C., Koenig, T., Jelic, V., et al.: Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). J. Neurosci. Methods **161**, 342–350 (2007). https://doi.org/10.1016/j.jneumeth.2006.10.023
20. Tripoliti, E.E., Fotiadis, D.I., Argyropoulou, M., Manis, G.: A six stage approach for the diagnosis of the Alzheimer's disease based on fMRI data. J. Biomed. Inform. **43**, 307–320 (2010)
21. Maroco, J., Silva, D., Rodrigues, A., et al.: Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res. Notes **4**, 299 (2011)
22. Patel, M.J., Andreescu, C., Price, J.C., et al.: Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int. J. Geriatr. Psychiatry **30**, 1056–1067 (2015)
23. Williams, J.A., Weakley, A., Cook, D.J., Schmitter-Edgecombe, M.: Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In: Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
24. Barlett, J.E., Kotrlik, J.W., Higgins, C.C.: Organizational research: determining appropriate sample size in survey research. Inf. Technol. Learn. Perform. J. **19**, 43 (2001)
25. Banerjee, S., Smith, S.C., Lamping, D.L., et al.: Quality of life in dementia: more than just cognition. An analysis of associations with quality of life in dementia. J. Neurol. Neurosurg. Psychiatry **77**, 146–148 (2006)
26. Cockrell, J.R., Folstein, M.F.: Mini-mental state examination (MMSE). Psychopharmacol. Bull. **24**, 689–692 (1987)
27. Delis D.C., Kramer, J.H., Kaplan, E., Ober, B.A.: California Verbal Learning Test: Adult version, Psychologi. San Antonio, Texas (1987)
28. Selnes, O.A.: A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, Ilustrada. Oxford University Press, Oxford (1991)
29. Howard, D., Patterson, K.: The Pyramid and Palm Trees Test: A Test of Semantic Access from Words and Pictures. Thames Val Test Company, Bury St Edmunds (1992)
30. Kaplan, E., Fein, D., Morris, R., Delis, D.: WAIS-R NI manual. Psychological Corporation, San Antonio (1991)
31. Liamputtong, P.: Focus Group Methodology: Principle and Practice. Sage Publications, Thousand Oaks (2011)
32. Diaz-Orueta, U., Facal, D., Nap, H.H., Ranga, M.-M.: What is the key for older people to show interest in playing digital learning games? Initial qualitative findings from the LEAGE project on a multicultural european sample. GAMES Health Res. Dev. Clin. Appl. **1**, 115–123 (2012)
33. Brox, E., Fernandez-Luque, L., Tøllefsen, T., et al.: Healthy gaming–video game design to promote health. Appl. Clin. Inform. **2**, 128–142 (2011)
34. Ahonniska, J., Ahonen, T., Aro, T., Lyytinen, H.: Suggestions for revised scoring of the tower of hanoi test. Assessment **7**, 311–319 (2000)
35. Howard, D., Patterson, K.E.: The Pyramids and Palm Trees Test: A Test of Semantic Access from Words and Pictures. Thames Valley Test Company, Suffolk (1992)
36. Rami, L., Bosch, B., Sanchez-Valle, R., Molinuevo, J.L.: The memory alteration test (M@T) discriminates between subjective memory complaints, mild cognitive impairment and Alzheimer's disease. Arch. Gerontol. Geriatr. **50**, 171–174 (2010)

37. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
38. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. **17**, 519–533 (2003)
39. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**, 263–282 (2010)
40. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
41. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**, 3–42 (2006)
42. Payan, A., Montana, G.: Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv Prepr arXiv:150202506 (2015)
43. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997)
44. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960)
45. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
46. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**, 1145–1159 (1997)
47. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Burlington (2016)