

## ORIGINAL ARTICLE

# Prediction of declarative memory profile in panic disorder patients: a machine learning-based approach

Felipe Dalvi-Garcia,<sup>1,2</sup> Laiana Azevedo Quagliato,<sup>1</sup> Donald J. Bearden,<sup>3,4</sup> Antonio Egidio Nardi<sup>1</sup>

<sup>1</sup>Instituto de Psiquiatria, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil. <sup>2</sup>Escola de Medicina e Cirurgia, Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Rio de Janeiro, RJ, Brazil. <sup>3</sup>Children's Healthcare of Atlanta, Atlanta, GA, USA.

<sup>4</sup>Emory University School of Medicine, Atlanta, GA, USA.

**Objective:** To develop a classification framework based on random forest (RF) modeling to outline the declarative memory profile of patients with panic disorder (PD) compared to a healthy control sample.

**Methods:** We developed RF models to classify the declarative memory profile of PD patients in comparison to a healthy control sample using the Rey Auditory Verbal Learning Test (RAVLT). For this study, a total of 299 patients with PD living in the city of Rio de Janeiro (70.9% females, age 39.9 ± 7.3 years old) were recruited through clinician referrals or self/family referrals.

**Results:** Our RF models successfully predicted declarative memory profiles in patients with PD based on RAVLT scores (lowest area under the curve [AUC] of 0.979, for classification; highest root mean squared percentage [RMSPE] of 17.2%, for regression) using relatively bias-free clinical data, such as sex, age, and body mass index (BMI).

**Conclusions:** Our findings also suggested that BMI, used as a proxy for diet and exercises habits, plays an important role in declarative memory. Our framework can be extended and used as a prospective tool to classify and examine associations between clinical features and declarative memory in PD patients.

**Keywords:** Panic disorder; memory; cognitive dysfunction; random forest classification; Rey auditory verbal learning test

## Introduction

Anxiety disorders (AD) are the most common psychiatric disorders with a lifetime prevalence of 33.7% in North America.<sup>1</sup> ADs are characterized by an excessive sense of fear and dread of real or perceived threat based in the past, present, or future.<sup>2</sup> Panic disorder (PD) is a common AD with a lifetime prevalence of about 1.6-2.2%.<sup>3</sup> PD is defined by recurrent unexpected panic attacks and persistent concern about additional attacks or their consequences. Symptoms of PD may include tachycardia, sweating, trembling, difficulty breathing, feeling of choking, chest pain, nausea, dizziness, derealization, fear of losing control, fear of dying, and chills or hot flushes.<sup>2</sup>

PD is associated with short- and long-term verbal memory problems<sup>4-7</sup> that may have a neurochemical etiology. Quagliato et al.<sup>8</sup> recently found that proinflammatory cytokines and elevated kynurenine/tryptophan ratio may lead to short-term auditory memory dysfunction. However, research on memory problems in PD remains limited and controversial.<sup>9</sup>

Declarative memory refers to the ability to store and recall personal events and factual information.<sup>10</sup> As such, it may worsen anxiety through storage and recollection of anxiety-provoking information<sup>11</sup> and plays an important role in learning healthy coping strategies and reframing maladaptive beliefs.<sup>12</sup> Further, declarative memory problems can negatively affect personal, professional, and social functioning.<sup>13,14</sup>

Rapid advancement of computational processing coupled with easier access to clinical and experimental data have allowed digital health technologies to speed up time between diagnosis and treatment of psychiatric disorders.<sup>15</sup> In this context, machine learning (ML) algorithms have been successfully employed in several research contexts and have shown promising results in identification and prediction of complex psychiatric symptomatology, supporting clinical and treatment decision-making. Such applications include, for example, creation of a risk calculator for attention deficit hyperactive disorder,<sup>16</sup> prediction of treatment outcome in patients with depression,<sup>17</sup> and identification of PD among other AD.<sup>18</sup>

Correspondence: Felipe Dalvi-Garcia, Instituto de Psiquiatria, UFRJ, Av. Venceslau Brás 71, CEP 20290-140, Rio de Janeiro, RJ, Brazil.

E-mail: fdalvigarcia@gmail.com

Submitted Jul 20 2023, accepted Sep 24 2023.

**How to cite this article:** Dalvi-Garcia F, Quagliato LA, Bearden DJ, Nardi AE. Prediction of declarative memory profile in panic disorder patients: a machine learning-based approach. Braz J Psychiatry. 2023;45:482-490. <http://doi.org/10.47626/1516-4446-2023-3291>

Random forest (RF) modeling is an ML technique that builds a set of statistical models (multiple decision trees) based on a given training dataset and can be used for non-linear classification and regression analyses.<sup>19</sup> RF performs classification by using majority voting of target output categorical variables with two or more levels. For regression, RF models use averaging to make predictions.<sup>20</sup> Appealing features of RF include the small number of adjustable parameters required, good performance with small datasets, scale invariance, and, generally, low sensitivity to overfitting.<sup>21</sup> In addition, RF regression calculates variable importance measures and partial dependence, which are useful indicators of the magnitude of influence of input variables and how they are associated with output variables.<sup>22</sup>

The purpose of our study was to develop a classification framework based on RF modeling to outline the declarative memory profile of patients with PD compared to a healthy control sample. In addition, we analyzed associations between declarative memory and salient features calculated by our RF model.

## Methods

The overall design of this study is illustrated graphically in Figure 1.

### *Panic disorder group*

This is a cross-sectional study that includes data collected between December 2017 and February 2020 from 304 participants diagnosed with PD. Patients were recruited through clinician referrals or self/family referrals and were excluded if they had any of the following conditions: uncontrolled cardiovascular, endocrinologic, hematologic, hepatic, renal, or neurologic diseases; autoimmune conditions; chronic infections; history of liver abnormalities; evidence of infection within 1 month of screening; history of cancer, pregnancy, or lactation; history of schizophrenia; active psychotic or depressive disorders; substance abuse and/or dependence within the past 6 months; active eating disorder; obsessive-compulsive disorder; or a score of less than 28 on the Mini-Mental State Examination (MMSE).<sup>8</sup>

All patients were living in neighborhoods in the environs of the Instituto de Psiquiatria, Universidade Federal do Rio de Janeiro (UFRJ), and had been receiving antidepressant medication for at least 3 months prior to entry into the study. Some patients also received adjunctive benzodiazepine treatment. PD was diagnosed with a structured clinical interview according to the DSM-IV-TR administered by a trained psychiatrist or psychologist and was independently confirmed by a senior psychiatrist. For all participants, the following variables were assessed and recorded: sex, age, level of education, body mass index (BMI), the number of individuals residing in the patients' home, socioeconomic status, caffeine consumption, alcohol consumption, and cigarette usage.

### *Declarative memory assessment*

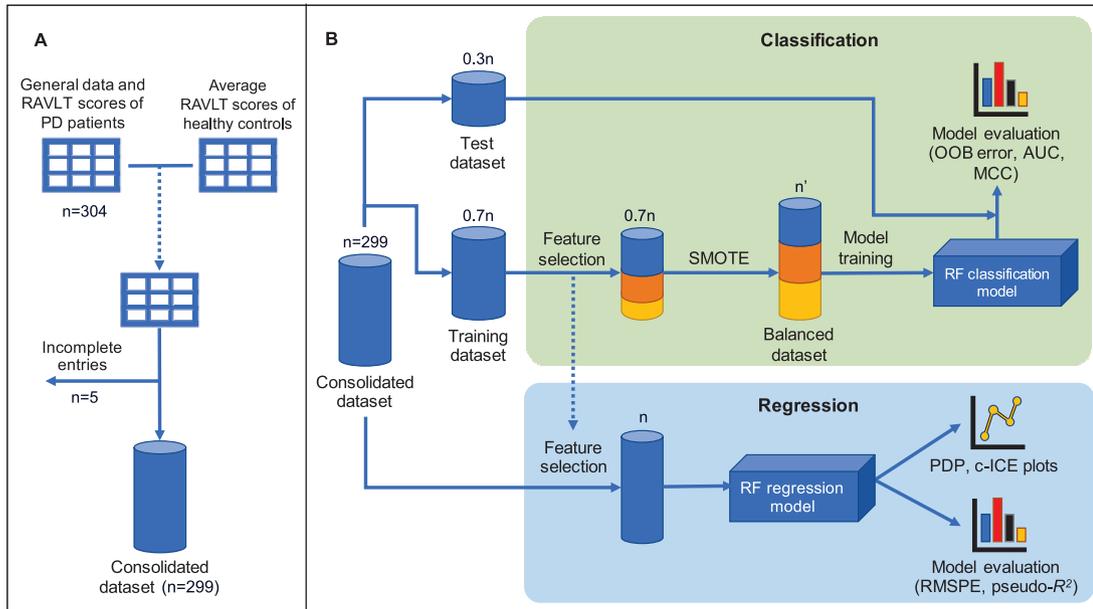
Participants' declarative memory was assessed using the Rey Auditory Verbal Learning Test (RAVLT), a widely used measure of learning, recall, and recognition memory (RM). The RAVLT is composed of a list of 15 nouns (list A) that are read out loud to the participant five times. Following each repetition, the participant is asked to recall as many words as possible and the number of correct answers is summed (scores A1 to A5). After the fifth attempt, an interference list with 15 different nouns (list B) is read to the participant, who is then asked to recall it (score B). Next, the participant is asked to recall the words from list A (score A6) followed by a 20-minute delay, after which the participant is asked to recall list A again (score A7). This is immediately followed by an RM test, during which the participant is asked to identify words from list A when read aloud and intermixed with nontarget words (15 from list A, 15 from list B, and 20 novel words) (Rec score).<sup>23,24</sup>

### *Control group sample data*

Evaluation of RAVLT performance is based on comparison with healthy normative data provided by the test manufacturer that includes demographic and cultural factors that may influence performance.<sup>25,26</sup> We used the normative data made up of a Brazilian sample of  $n=302$  males and females (62.6% male) ranging from 17 to 85 years old ( $50.6 \pm 15.9$  years) with a range of 1 to 20 years of education ( $11.3 \pm 3.7$  years). In this context, a healthy population is defined as individuals who did not exhibit any of the following conditions/factors: a history of psychiatric disturbances or a current state of psychiatric disturbance; diabetes, heart problems or any other related pathological conditions; use of psychoactive drugs or drugs known to have significant side effects that affect memory function; or use of psychoactive drugs within the past 12 months.<sup>24</sup> The group mean ( $\bar{x}_{ij}$ ) and SD ( $\sigma_{ij}$ ) for each  $i$ -th RAVLT score was grouped by age: 17-34 years, 34-49 years, 50-64 years, and 65-85 years.<sup>24</sup> We used these data to calculate 95%CI [ $\bar{x}_{ij} \pm 1.96\sigma_{ij}/\sqrt{n_j}$ ] for each RAVLT score within each  $n_j$ -size age group in the healthy population.

### *Data preparation*

We used the RAVLT scores of PD patients to additionally compute a proactive interference (PI) score (B/A1), a retroactive interference (RI) score (A6/A5), and a forgetting speed (FS) score (A7/A6). We then classified each of the variables (i.e., PI, RI, FS, and RM) as lower, similar or higher based on whether they were below, within, or above the 95%CI for the control group scores. Incomplete entries (i.e., those missing information for any variable) were removed from the dataset ( $n=5$ ), resulting in 299 participants in the consolidated dataset used to develop our RF models.



**Figure 1** Graphical illustration of the study design. A) Consolidation of information collected from panic disorder (PD) patients and normative data provided by the Rey Auditory Verbal Learning Test (RAVLT) manufacturer, collected from healthy controls. B) Modeling process. For classification, the dataset was split into training (70%) and test (30%) datasets; after feature selection and synthetic minority oversampling technique (SMOTE) implementation, the balanced dataset size ( $n'$ ) is different and larger than the original training dataset because all minority classes were oversampled to the size of the majority class. The same features previously selected were used for regression but using the same samples as the consolidated dataset. The steps in the classification and regression boxes were repeated for each target variable. AUC = area under the curve; c-ICE = centered individual conditional expectation; MCC = Matthew correlation coefficient; OOB = out-of-bag; PDP = partial dependence plots; RF = random forest; RMSPE = root mean squared percentage error.

### Random forests modeling

#### Classification

For classification, the consolidated dataset was split 70/30, into training and test datasets respectively. Because correlated predictors may affect estimation of importance scores calculated by RF models, we initially ran an RF recursive feature elimination (RFE) algorithm to perform feature selection. RFE trains RF models iteratively, ranks features, and removes those with the lowest ranking. Also, since RFE reduces the importance of correlated data, the algorithm favors selection of uncorrelated features.<sup>27</sup> Ten-fold cross-validation with five repeats was used to build the control object used in the RFE algorithm, which was run for each target variable.

The class proportion for each target variable in the training dataset is imbalanced, which can result in biased classification in favor of the majority class. Thus, we used a synthetic minority oversampling technique (SMOTE) to artificially balance the ratio between classes at a proportion of 1:1:1, using the number of samples of the majority class as a reference.<sup>28</sup> Since the sample size of the majority class varies depending on the target variable, the final balanced datasets have different sizes and are larger than the original training dataset.

To assess the performance of RF classifiers, we computed the generalization error obtained by out-of-bag (OOB) samples from the training dataset, the

Matthew correlation coefficient (MCC), and the multiclass area under the curve (AUC) from its receiver operating characteristic (ROC) curve, which are calculated based on the test dataset.<sup>29</sup> The OOB samples are observations that are not bootstrapped to train an RF model and are used as an internal validation dataset to estimate the accuracy of that model.<sup>19</sup> In addition, the OOB error is a convenient approach to internally evaluate RF models without performing a cross-validation procedure.<sup>30</sup> The MCC has recently been recommended as one of the best evaluation methods for binary and multi-class classification tasks in addition to its statistical reliability and informative score over the most popular F1-score and accuracy. The MCC ranges from -1 to 1, where 1 indicates perfect classification and -1 indicates perfect misclassification.<sup>31</sup> Finally, the ROC curve represents the compromise between true positive rate (sensitivity) and false positive rate (1-specificity) ranging from 0 to 1. Hence, the AUC is a threshold-independent metric commonly used to evaluate classification performance.<sup>32</sup>

#### Regression

Using RF models to conduct regression can be useful for analyzing partial dependence plots (PDP). In non-linear regression methods like tree-based models, the association between input and output variables is not straightforward, making direct statistical inference more complex.<sup>22</sup> PDP were proposed by Friedman<sup>33</sup> as a method for

visualizing the marginal effect of input variables on predicted output of any ML model. Thus, PDP analysis helps to identify and describe associations between input features and outcome variables. However, the reliability of PDP interpretation is limited by the assumption of independence between variables. Thus, use of uncorrelated variables selected by the RFE algorithm aims to improve the dependability of PDP analysis. Since the RAVLT scores are correlated to their correspondent classes for simplicity, for regression we used the same RFE-selected variables used as input features in the classification task.

One weakness of PDP is that it may mask heterogeneous relationships derived from feature interactions. In such cases, individual conditional expectation (ICE) plots may be useful to correctly interpret PDP.<sup>34</sup> ICE plots show dependence of prediction for each observation separately (i.e., one line per observation that represents how the prediction of that instance changes when a feature changes). Thus, PDP are the average of the lines of an ICE plot built from each observation or subset of observations from the dataset. We used PDP from centered ICE plots (c-ICE) to clarify the behavior of the RF regressions and provide a more accurate interpretation.<sup>34</sup> B/A1, A6/A5, A7/A6, and Rec scores were set as target variables in the RF regression models.

To evaluate the accuracy and the goodness-of-fit of each regression, we computed the root mean squared percentage error (RMSPE) and the pseudo- $R^2$  coefficient. RMSPE is the square root of the average of the squared errors expressed as a percentage. The closer RMSPE is to 0, the more accurate the regression model. Pseudo- $R^2$  can be used to evaluate the extent to which the variance of the data can be explained by the model and is particularly useful when  $R^2$  cannot be used (e.g., to assess fit of nonlinear models).<sup>35</sup>

### Implementation

All analyses and models were implemented using the free R language (version 4.1.3) on the RStudio platform (version 2022.02.0). The number of trees parameter ( $n_{tree}$ ) was set to 500 and optimal split ( $m_{try}$ ) was set to 2. The code and dataset are publicly available at the project's Open Science Framework page and can be accessed at <https://osf.io/ckg37/>.

### Ethics statement

Written informed consent was obtained prior to participation in the study, which was approved by the research ethics committee at UFRJ. This study was performed according to the ethical standards defined in the Declaration of Helsinki.

### Results

Table 1 summarizes descriptive statistics for the RFE-selected features of our PD patients after data preparation. Except for BMI and age, in PI and RM prediction

respectively, all features were significantly different between groups to  $p < 0.01$ .

Performance indicators for each RF model are reported in Table 2. For PI and RI, the RFE algorithm selected age, BMI, socioeconomic status, and level of education as the most important and uncorrelated input variables. In addition to the same aforementioned features, FS and RM were used for training the number of people with whom each patient lived. According to multiclass AUCs and MCCs, our models made the correct classification in 100% of cases. OOB errors were extremely low, even in the worst case (0.57%, for RM prediction). Regarding regression, the overall goodness-of-fit was very high, with a lowest pseudo- $R^2$  of 98.4%, for RM score prediction, and a highest RMSPE of 9.64%, for PI score prediction.

Because some of our predictor variables are vulnerable to social and cultural bias (e.g., educational level, socioeconomic status), we evaluated the ability of our modeling approach to classify tasks using objective variables (i.e., sex, age, and BMI) (Table 2). Even with reduced classification performance in comparison to the previous models that included potential for bias, the highest OOB error in the retrained model was 11.6%, and the lowest AUC and MCC were 0.979 and 0.852, respectively. As for regression, the goodness-of-fit was at least 94.5% and the highest RMSPE was 17.2%, for PI score prediction.

Finally, PDP and ICE/c-ICE plots for PI, RI, FS, and RM are shown in Figure 2. From the original set of variables used in the classification task, only age and BMI are presented because they were the two most important continuous features according to the RFE algorithm.

Assuming uncorrelated features, the initial interpretation of the PDP and c-ICE plots is straightforward: the flatter the line in these plots, the less that feature influences the prediction. Figure 2 shows that the age feature has a clear influence across RAVLT scores. Based on the behavior of the c-ICE curves and their corresponding PDP, a clear declining trend is observed in PI, RI, and RM with respect to the age of patients with PD. Notice that inferences cannot be made from c-ICE plots and PDP outside of the range of the sample (i.e., ages beyond the range 22-52 years and BMI beyond the range 19.1-25.4 kg/m<sup>2</sup>).

### Discussion

The purpose of this study was to develop a tool to help clinicians screen for declarative memory impairments in people with PD. To do this, we created an ML-based model that advances existing classification/regression approaches by identifying complex relationships between clinical features and declarative memory measures based on RAVLT scores. To the best of our knowledge, this is the first study to predict declarative memory profiles in patients with PD using ML modeling.

Our findings suggest that, in general, our models were able to predict PI, RI, FS, and RM status in persons with PD based on age, BMI, socioeconomic status, level of education, and the number of people each patient lives with. The classification models were highly predictive (AUCs and MCCs equal to 1.00) regardless of whether



**Table 2** Performance evaluation of RF classification and regression models

| Input feature/Target variable | RF classification |                  |       | RF regression |                           |
|-------------------------------|-------------------|------------------|-------|---------------|---------------------------|
|                               | OOB error (%)     | AUC <sup>†</sup> | MCC   | RMSPE (%)     | Pseudo-R <sup>2</sup> (%) |
| RFE-selected variables        |                   |                  |       |               |                           |
| PI (B/A1)                     | 0.00              | 1.00             | 1.00  | 9.64          | 98.7                      |
| RI (A6/A5)                    | 0.33              | 1.00             | 1.00  | 1.61          | 99.6                      |
| FS (A7/A6)                    | 0.28              | 1.00             | 1.00  | 3.12          | 98.6                      |
| RM                            | 0.57              | 1.00             | 1.00  | 1.81          | 98.4                      |
| Age, sex, and BMI             |                   |                  |       |               |                           |
| PI (B/A1)                     | 8.41              | 0.996            | 0.913 | 17.2          | 95.9                      |
| RI (A6/A5)                    | 11.6              | 0.979            | 0.852 | 5.28          | 95.1                      |
| FS (A7/A6)                    | 3.67              | 0.994            | 0.855 | 6.23          | 94.5                      |
| RM                            | 0.287             | 1.00             | 1.00  | 1.69          | 98.6                      |

AUC = area under the curve; BMI = body mass index; FS = forgetting speed; MCC = Matthew correlation coefficient; OOB = out-of-bag; PI = proactive interference; RF = random forest; RFE = recursive feature elimination; RI = retroactive interference; RM = recognition memory; RMSPE = root mean squared percentage.

<sup>†</sup> Multiclass AUC, as defined by Hand & Till.<sup>29</sup>

patients with PD had lower, similar, or higher RAVLT scores in comparison to a healthy control sample. Also, the regression models had excellent goodness-of-fit (pseudo-R<sup>2</sup> ranging from 98.4 to 99.6%) and a relatively low error rate (RMSPE ranging from 1.61 to 9.64%) in predicting the RAVLT scores. Interestingly, the models retrained with just the features sex, age, and BMI displayed lower but still good performance in classification (AUCs from 0.979 to 1.00, MCCs from 0.852 to 1.00) and regression (pseudo-R<sup>2</sup> from 94.5 to 98.6%, RMSPE from 1.69 to 17.2%). Although the predictive performance of our model declined due to removal of important RFE-selected features, our unique approach using only three objective features yielded good predictive results, especially when compared to similar approaches in the literature.<sup>17,18,32</sup>

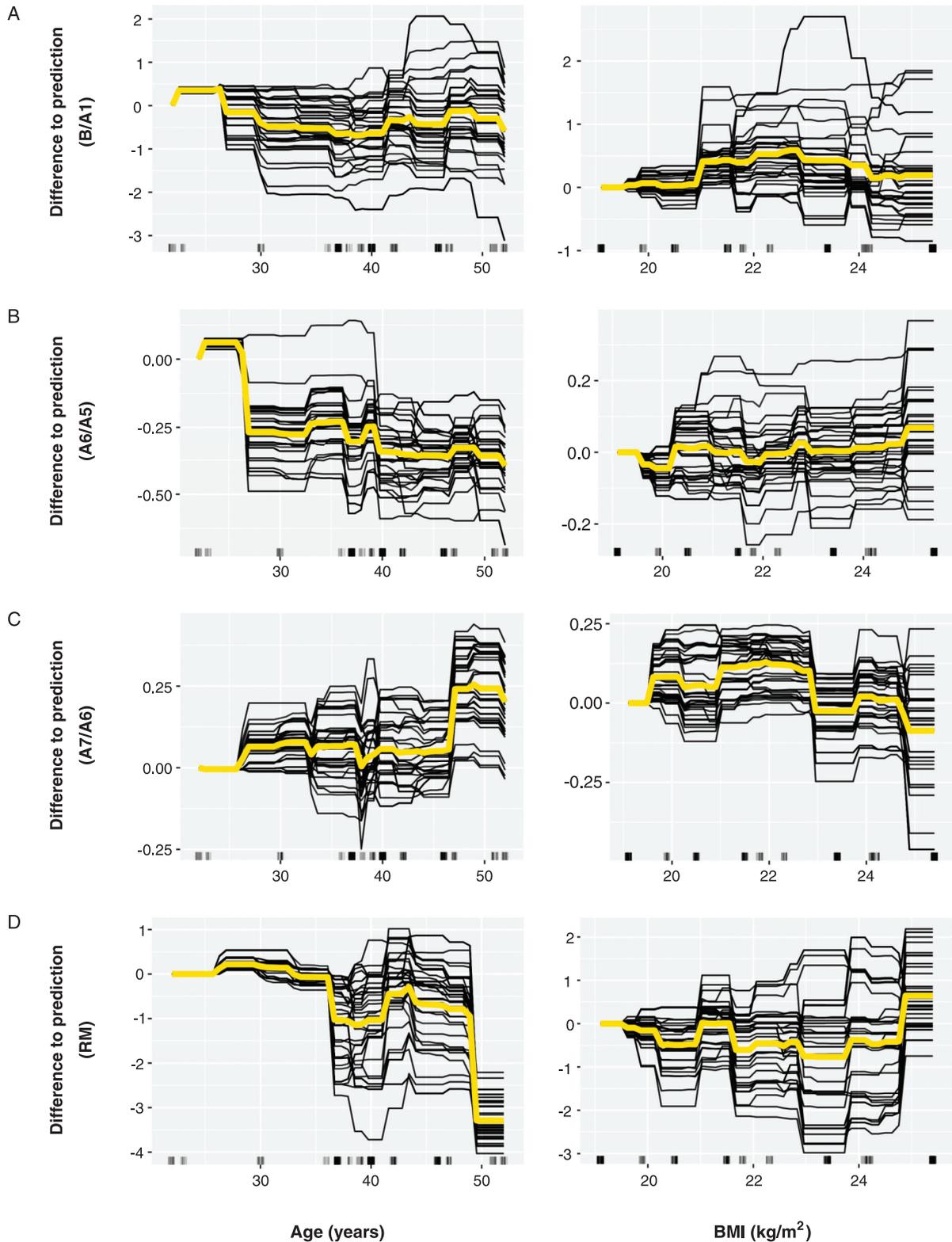
Based on the PDP and c-ICE plots (Figure 2), age had a clear influence on all aspects of declarative memory in our study as measured by the RAVLT, a finding observed in previous research with healthy populations.<sup>24-26,36</sup> Age-related decline was observed in PI, RI, and RM in patients with PD. Interestingly, our results suggest that FS increases from 45 to 50 years of age, but it remains to be seen if this trend continues for older ages.

Despite most study participants having healthy BMI (19.1-25.4 kg/m<sup>2</sup>), our regression model captured the relationship between this feature and declarative memory. The associations between BMI and PI and FS were reliably interpretable in the PDP and c-ICE plots, whereas the associations between BMI and RI and RM were too variable to interpret. In addition, we found trends in our analyses beyond what could be captured in our representative sample that suggest performance declines in PI and FS as BMI increases. Specifically, findings showed that PI is average for patients within the normal BMI range (20-25) and suggested that it declines as BMI exceeds this range. Similarly, FS was average for patients within the normal BMI range, but results suggested that it decreases as BMI increases beyond this range. A large body of research has identified a negative association between weight and cognitive functioning.<sup>37,38</sup> Most studies to date, however, have focused on the effect of

obesity-related neuroinflammation on cognitive functioning. For example, Loprinzi & Frith<sup>38</sup> identified morphological brain changes, insulin resistance, neuroinflammation, hypertriglyceridemia, elevated glucocorticoids, and cerebral metabolites as underlying mechanisms of obesity-related memory impairments. In line with findings from these studies, our results corroborate the association between elevated BMI, which reflects dietary and exercise habits, and declarative memory.

Given the simplicity of our framework, it could be further validated and used by a wide range of health professionals in a variety of settings (e.g., psychiatric care centers, nursing homes, primary health care institutions, etc.). As is true for any psychiatric assessment tool, ours is meant to be used as part of a comprehensive psychiatric/neuropsychological evaluation and should not be used in isolation to diagnose or treat declarative memory impairments in PD. Nevertheless, we believe our tool provides a relatively inexpensive and efficient method of detecting memory dysfunction in patients with PD.

Limitations of our study include the fact that our model is based on data from different cross-sectional studies to represent PD and healthy groups. Although this may not interfere with the modeling aspect, it may lead to biased prediction due to selection bias and confounding factors such as use of antidepressants by PD patients. Second, our PD sample size is relatively small, but it has been shown that RF models in general have good predictive performance even with limited sample size and that the predictive improvement achieved by increasing sample size is limited.<sup>21,39</sup> Also, our models were developed for a highly specific population in a single Brazilian city, which may limit their validity if used in other settings. However, our study design and framework can be used for larger sample sizes and different populations, since the generalizability of ML models is boosted as the training dataset variability increases.<sup>40</sup> Thus, a larger sample size would improve the robustness of classification and regression tasks and provide a more extensive and well-distributed sample for features, especially age and BMI. We encourage further multicenter studies with larger sample sizes representing a greater variety of populations.



**Figure 2** Partial dependence plots (PDP) (yellow lines) and centered individual conditional expectation (c-ICE) plots (black lines) of the effects of the features age and body mass index (BMI) on target variables (A) proactive interference (PI), (B) retroactive interference (RI), (C) forgetting speed (FS), and (D) recognition memory (RM) (from top to bottom). Note: Increase in FS indicates a decline in performance.

To improve the reliability of ML algorithms in classifying cognitive changes or states in patients with PD, we strongly suggest use of a control group in the same study. In addition to the classification task, interpretable ML algorithms such as RF are recommended because they allow for analysis of complex relationships between features and target variables.

In conclusion, we developed a novel, multi-class, highly accurate ML-based tool to help healthcare professionals identify declarative memory dysfunction in patients with PD. Our tool is in line with new applications in the field of computational and precision psychiatry and has shown excellent results based on internal validation. Findings from this study should be externally validated in future research with different and larger populations from various clinical settings in order to improve our model's robustness and increase its generalizability.

### Acknowledgements

We gratefully acknowledge the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for supporting this research.

### Disclosure

The authors report no conflicts of interest.

### References

- 1 Bandelow B, Michaelis S. Epidemiology of anxiety disorders in the 21st century. *Dialogues Clin Neurosci*. 2015;17:327-35.
- 2 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. Arlington: American Psychiatric Publishing; 2013.
- 3 Weissman MM, Bland RC, Canino GJ, Faravelli C, Greenwald S, Hwu HG, et al. The cross-national epidemiology of panic disorder. *Arch Gen Psychiatry*. 1997;54:305-9.
- 4 Boldrini M, Del Pace L, Placidi GPA, Keilp J, Ellis SP, Signori S, et al. Selective cognitive deficits in obsessive-compulsive disorder compared to panic disorder with agoraphobia. *Acta Psychiatr Scand*. 2005;111:150-8.
- 5 Asmundson GJ, Stein MB, Larsen DK, Walker JR. Neurocognitive function in panic disorder and social phobia patients. *Anxiety*. 1994;1:201-7.
- 6 Lucas JA, Telch MJ, Bigler ED. Memory functioning in panic disorder: a neuropsychological perspective. *J Anxiety Disord*. 1991;5:1-20.
- 7 Airaksinen E, Larsson M, Forsell Y. Neuropsychological functions in anxiety disorders in population-based samples: evidence of episodic memory dysfunction. *J Psychiatr Res*. 2005;39:207-14.
- 8 Quagliato LA, Freire RC, Nardi AE. Elevated peripheral kynurenine/tryptophan ratio predicts poor short-term auditory memory in panic disorder patients. *J Psychiatr Res*. 2019;113:159-64.
- 9 O'Sullivan K, Newman EF. Neuropsychological impairments in panic disorder: a systematic review. *J Affect Disord*. 2014;167:268-84.
- 10 Bertola L, Malloy-Diniz LF. Assessing knowledge: psychometric properties of the BAMS semantic memory battery. *Rev Psiquiatr Clin*. 2018;45:33-7.
- 11 Zlomuzica A, Dere D, Machulska A, Adolph D, Dere E, Margraf J. Episodic memories in anxiety disorders: clinical implications. *Front Behav Neurosci*. 2014;8:131.
- 12 Mur M, Portella MJ, Martinez-Aran A, Pifarre J, Vieta E. Influence of clinical and neuropsychological variables on the psychosocial and occupational outcome of remitted bipolar patients. *Psychopathology*. 2009;42:148-56.
- 13 Lengenfelder J, Dahlman KL, Ashman TA, Mohs RC. Psychological assessment of the elderly. In: Goldstein G, Allen DN, DeLuca J, editors. *Handbook of psychological assessment*. 4th ed. Amsterdã: Elsevier; 2019. p. 505-32.
- 14 Cubillas CP. Declarative memory. In: Vonk J, Shackelford T, editors. *Encyclopedia of animal cognition and behavior*. Cham: Springer International Publishing; 2017. p. 1-5.
- 15 Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3:223-30.
- 16 Caye A, Agnew-Blais J, Arseneault L, Gonçalves H, Kieling C, Langley K, et al. A risk calculator to predict adult attention-deficit/hyperactivity disorder: generation and external validation in three birth cohorts and one clinical sample. *Epidemiol Psychiatr Sci*. 2019;29:e37.
- 17 Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3:243-50.
- 18 Na KS, Cho SE, Cho SJ. Machine learning-based discrimination of panic disorder from other anxiety disorders. *J Affect Disord*. 2021;278:1-4.
- 19 Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees. classification and regression trees*. London: Routledge; 2017.
- 20 Dietterich TG. Ensemble methods in machine learning. In: Kitterl J, Roli F, editors. *Lecture notes in computer science*. Berlin: Springer; 2000. p. 1-15.
- 21 Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: a synthetic knowledge synthesis. *Sci Prog*. 2022;105:003685042110297.
- 22 Auret L, Aldrich C. Interpretation of nonlinear relationships between process variables by use of random forests. *Miner Eng*. 2012;35:27-42.
- 23 Moradi E, Hallikainen I, Hänninen T, Tohka J; Alzheimer's Disease Neuroimaging Initiative. Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease. *NeuroImage Clin*. 2016;13:415-27.
- 24 Magalhães SS, Hamdan AC. The Rey Auditory Verbal Learning Test: normative data for the Brazilian population and analysis of the influence of demographic variables. *Psychol Neurosci*. 2010;3:85-91.
- 25 Van der Elst W, Van Boxtel MPJ, Van Breukelen GJP, Jolles J. Rey's verbal learning test: normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *J Int Neuropsychol Soc*. 2005;11:290-302.
- 26 Malloy-Diniz LF, Lasmar VAP, Gazinelli LDSR, Fuentes D, Salgado JV. The Rey Auditory-Verbal Learning Test: applicability for the Brazilian elderly population. *Braz J Psychiatry*. 2007;29:324-9.
- 27 Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Stat Comput*. 2017;27:659-78.
- 28 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-57.
- 29 Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn*. 2001;45:171-86.
- 30 Mitchell MW. Bias of the random forest out-of-bag (OOB) error for certain input parameters. *Open J Stat*. 2011;1:1-7.
- 31 Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS One*. 2012;7:e41882.
- 32 Oh T, Kim D, Lee S, Won C, Kim S, Yang J, et al. Machine learning-based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES. *Sci Rep*. 2022;12:2250.
- 33 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189-232.
- 34 Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat*. 2015;24:44-65.
- 35 Smith G. *Essential statistics, regression, and econometrics*. 2nd ed. Cambridge: Academic Press; 2011.

- 36 Diniz LFM, De Fátima Da Cruz M, De Macedo Torres V, Cosenza RM. O teste de aprendizagem auditivo-verbal de Rey: normas para uma população brasileira. *Rev Bras Neurol.* 2000;3:79-83.
- 37 Dahl AK, Hassing LB. Obesity and cognitive aging. *Epidemiol Rev.* 2013;35:22-32.
- 38 Loprinzi PD, Frith E. Obesity and episodic memory function. *J Physiol Sci.* 2018;68:321-31.
- 39 Luan J, Zhang C, Xu B, Xue Y, Ren Y. The predictive performances of random forest models with limited sample size and different species traits. *Fish Res.* 2020;227:105534.
- 40 Therrien R, Doyle S. Role of training data variability on classifier performance and generalizability. In: Gurcan MN, Tomaszewski JE, editors. *Medical imaging 2018: Digital Pathology.* Houston: SPIE; 2018. p. 5.