Computers in Biology and Medicine 180 (2024) 108955

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Designing an effective semantic fluency test for early MCI diagnosis with machine learning

Alba Gómez-Valadés *, Rafael Martínez, Mariano Rincón

Universidad Nacional de Educación a Distancia, Madrid, 28040, Comunidad Autónoma de Madrid, Spain[1]

ARTICLE INFO

Keywords:
Semantic fluency test
MCI
Machine learning
Switching
Clustering
Early diagnosis

ABSTRACT

Semantic fluency tests are one of the key tests used in batteries for the early detection of Mild Cognitive Impairment (MCI) as the impairment in speech and semantic memory are among the first symptoms, attracting the attention of a large number of studies. Several new semantic categories and variables capable of providing complementary information of clinical interest have been proposed to increase their effectiveness. However, this also extends the time required to complete all tests and get the overall diagnosis. Therefore, there is a need to reduce the number of tests in the batteries and thus the time spent on them while maintaining or increasing their effectiveness. This study used machine learning methods to determine the smallest and most efficient combination of semantic categories and variables to achieve this goal. We utilized a database containing 423 assessments from 141 subjects, with each subject having undergone three assessments spaced approximately one year apart. Subjects were categorized into three diagnostic groups: Healthy (if diagnosed as healthy in all three assessments), stable MCI (consistently diagnosed as MCI), and heterogeneous MCI (when exhibiting alternations between healthy and MCI diagnoses across assessments). We obtained that the most efficient combination to distinguish between these categories of semantic fluency tests included the *animals* and *clothes* semantic categories with the variables *corrects*, *switching*, *clustering*, and *total clusters*. This combination is ideal for scenarios that require a balance between time efficiency and diagnosis capability, such as population-based screenings.

1. Introduction

Mild cognitive impairment (MCI) is one of the main early indicators of several neurodegenerative diseases, including Alzheimer's Disease (AD) [1]. The interest in AD has been growing as the population ages [2–4], as it is the main neurodegenerative disease among the elderly population [5]. Therefore, early detection of MCI is indispensable for appropriate treatment to reduce the disease progression and to improve patients' quality of life [3–6]. In this context, it has become essential to develop methods for the early diagnosis of MCI that are effective and, at the same time, quick and easy to administer and assess [3,4,7,8], which is the ideal combination to be used in screening campaigns [3]. Neuropsychological tests are considered ideal for this task due to their good balance of diagnostic capability, cost-effectiveness, speed, and non-invasiveness compared with the alternatives [6,7].

Semantic fluency tests, one of the most widely used tests in neuropsychological batteries, are among them [9,10]. Performance on those tests is sensitive to variations in executive function and semantic memory [11] since language problems appear in the earliest stages of MCI [4,11]. Therefore, those tests have been the focus of a large number of studies, analyzing their performance [2,12–14], searching for new variables capable of providing complementary information of clinical interest such as *switching*, *clustering* or non verbal-elements [3,4,9–11,15–19], automating the evaluation of such tests [20,21], or combining them with other types of tests [7].

In a typical semantic fluency test, the subject must list the maximum number of words belonging to a given semantic category within a time limit, usually 60 s [22]. The traditional scoring method counts only the number of valid terms in the produced sequence, excluding any *perseveration* [7,11]. However, there is strong evidence that these tests may provide more clinically valid information than is currently considered [3,9–11,14–16,18,21,23]. This evidence has been sought primarily in switching and clustering. These scores are based on the subjects' tendency to spontaneously generate successive items within the same semantic subcategory during the verbal production and, once that subcategory is exhausted, switch to another one during verbal production [9,15,18,24,25].

* Corresponding author.
E-mail addresses: albagvb@dia.uned.es (A. Gómez-Valadés), rmtomas@dia.uned.es (R. Martínez), mrincon@dia.uned.es (M. Rincón).
[1] https://www.uned.es/universidad/facultades/informatica.html.

https://doi.org/10.1016/j.compbiomed.2024.108955
Received 11 April 2024; Received in revised form 4 July 2024; Accepted 26 July 2024
Available online 16 August 2024

However, administering these tests remains time-consuming for both the subjects, who may become fatigued after a series of tests [26,27], and the healthcare professionals, as the evaluation is still manual [11,27]. Moreover, manual evaluation is prone to errors and depends to some extent on the subjectivity of the evaluator [3,5,11,27,28], particularly when MCI symptoms are so mild that they could be mistaken for consequences of normal aging [11,29]. Automating data collection, evaluating the efficiency of data and metrics to identify redundant and low-relevance items, and analyzing new variables of potential clinical interest and their implementation feasibility emerge as a necessity to enhance the efficiency of neuropsychological test batteries used in screenings [3,19,28,30].

On the other hand, machine learning (ML) models have been used in the medical research context since their beginnings to assist the work of physicians [6,30]. Within the neuropsychological and dementia areas, they allow replicable, almost immediate data collection and evaluation compared to manual scoring methods [3,4,11,28]. These automatic methods are more objective than manual scoring since they apply the same criteria to all cases [3,11]. Moreover, they allow the simultaneous assessment of large amounts of data, easing the screening and providing a fast first approximation of the cognitive state of subjects [5,6,8,21]. However, their use in normal clinical practice, outside the research context, is currently scarce. The reasons are due to the difficulty of implementation, mostly if they are associated with complex equipment or are dependent on external repositories [3,11,19].

These ML methods can also be used to analyze both traditional and novel tests and variables to determine their diagnostic utility [7,30,31], helping to propose more efficient batteries and variables and as a first analysis of subjects‘ cognitive state [32]. Within the projects to improve the diagnostic efficiency of semantic fluency tests, it has been proposed to automate the extraction of word sequences from oral production [4,28,33], to obtain and evaluate both new variables and combinations of different diagnostic techniques [3,7,17,19,20], and to perform results analysis, and diagnosis support [3,19,21].

In this study, our aim is to identify the most effective semantic fluency test combination, one that can accurately differentiate between individuals with MCI and healthy subjects, while also being the most concise version of those test combinations. This ensures quicker administration, which will benefit both specialists and patients and allow it to be used in population screenings. To achieve this, we automated the collection of all scores and evaluated several combinations of semantic categories and variables using ML models to identify the most efficient and discriminative combination for MCI detection. To mirror real-world scenarios, where subjects may not fit neatly into defined healthy or MCI profiles, our study included individuals with light MCI or in the process of transitioning to MCI.

The rest of the article follows as detailed. Section 2 provides a summary of previous work in this area, which allows us to put this research in context. Section 3 describes the process of automatic scoring from the transcribed sequence of words and the test-variable combinations that were analyzed with ML models. Section 4 shows the results obtained from the analyses described in the previous section. Section 5 discusses the results obtained in the different stages of this work, and Section 6 summarizes the main achievements.

2. Related work

Semantic Fluency

Semantic fluency tests have traditionally been evaluated by the count of the total number of correct words [11]. This simple and straightforward method is done manually, which generates an additional time cost for the evaluators and the possibility of introducing both manual errors and inconsistencies. The latter occurs due to the presence of ambiguous terms where the evaluator must make subjective decisions about their inclusion or exclusion (*e.g.*, “dinosaur” in the *animals*’ semantic fluency test) [5,19]. The inclusion of complementary scores implies more time for healthcare professionals to evaluate each test, and a higher chance of introducing manual errors and inconsistencies [11]. Test automation would apply the same criteria to all patients, objectifying the scores and helping in the reproducibility of results [3,5], as well as obtaining the scores almost immediately [3]. This would enable healthcare professionals to get the most out of the tests.

In this context, automatic scoring can be conducted directly from oral production using speech recognition or from transcribed word sequences. Speech recognition techniques obtain the associated scores from the recorded verbal production. This allowed them to obtain silences and elements outside word production that may be indicative of underlying brain damage [4]. However, it is not currently feasible to use these variables in routine clinical practice due to the necessary technical requirements [3,19,28] and problems in handling noise and extraneous elements of the verbal production itself [28,33]. On the approaches based on the transcribed sequence of words, different methods have been proposed. One of them is the Latent Semantic Analysis (LSA) [34,35], a technique based on the co-occurrence of words in a large text corpus since it assumes that words with similar meanings appear in analogous texts. To calculate the proximity between two words, machine learning methods are trained with a corpus generated from texts of previous clinical studies. This method was employed in the study of Ledoux et al. [17] to obtain the scores of *clustering* and *switching* automatically. Their results were comparable to the ones obtained by Troyer et al. [16], avoiding the previous manual work of defining the semantic categories. Another approach is the Explicit Semantic Analysis (ESA) [36], a method based on the quantification of word relationships inside a “conceptual space” defined by the analysis of Wikipedia entries. This method was used in the study of Woods et al. [19], and was further explored in later studies, combining it with other techniques such as neural word embedding [37], or with variants of word2vec models [11]. This method has the advantages of not needing a previous corpus, and novel words are treated consistently [11].

Methods based on both LSA and ESA allow automating scores, especially *clustering* and *switching*, by saving the step of creating the semantic subcategories [3,11,25]. However, both methods require a sufficiently large and curated corpus in the same language as the tests, which is not always possible [3,11,18]. Moreover, these methods do not automatically indicate what semantic subcategories are involved, and they may not be best suited to capture both the semantic associations made by humans and the sociocultural context of the sample [37]. For example, “shark” and “tiger” appeared as strongly related by these methods, as both were superpredators. However, in the semantic subcategories defined by Troyer [25], they appeared in different clusters. And vice versa, words included in the same manually defined semantic subcategory might present a poor relationship with those proposed by automatic methods, a more significant problem in the ESA [19]. Due to these problems, we considered it more appropriate to define the semantic subcategories manually, and then obtain the scores of the associated variables automatically.

In addition to the *animals* task and the *switching* and *clustering* variables, other semantic categories and variables have been analyzed. Although variables related to changes between subsemantic categories in the *animals* test have attracted great attention, other semantic categories have also been evaluated. In terms of variables, the most common ones have been *repetitions* or *perseverations* [7,19,23], *intrusions* [7], and *phonetic continuity* [10,14], typically yielding worse results than *corrects*, *clustering* or *switching* [7,10,19], although this is not always the case [14]. Regarding tests, some of the most used ones have been *fruits*, *vegetables*, or *supermarket items*. The results obtained with these semantic categories are usually equal or inferior to those obtained with the *animals* test. In particular, the semantic categories of *fruits* and *vegetables* scored worse than *animals* [7,23], while the

supermarket-items test was highly discriminative, and could be used as a reliable complementary semantic fluency test to *animals* [25].

Machine Learning

On the other hand, ML and deep learning techniques have been increasingly employed for the automatic detection of both MCI and AD, where their use in conjunction with neuroimaging tests is experiencing a boom. This can be seen in the large number of systematic reviews in this area conducted in recent times [38–40]. However, in contrast to the neuropsychological tests, it is not feasible to use neuroimaging tests in mass population screening due to their high cost, specialized equipment, and time needed [41].

Focusing on neuropsychological tests, we can highlight the following studies. Adelson et al. [42] compared the performance of an ML-based technique fed with sociodemographic and neuropsychological test data with the MMSE and another three models, obtaining better results for the ML-based technique. Bucholc et al. [41] combined both unsupervised and supervised learning using a large battery of tests, where the labels needed by the supervised methods were generated using the results obtained by the unsupervised learning, improving the results of the system. Franciotti et al. [43], employed three ML methods (Random Forest, Gradient Boosting, and eXtreme Gradient Boosting) on a multimodal dataset, obtaining the best results when the different ML models were combined.

Particularized on semantic fluency tests, ML techniques have proven to be especially useful for establishing an automatic a priori classification of subjects [3,7,20,37,44,45], obtaining and evaluating the variable scores automatically [3,11,19,37]. However, most of these studies focus only on a few models, either because of their power, like Random Forest [18,44], Support Vector Machine [7,13] or XGBoost in more recent times [42,43], or because of their interpretability, like models based on relationships between concepts, such as knowledge graphs or probabilistic networks [20,21,45]. Several of these research also used external databases, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [40,42,43], the National Alzheimer's Coordinating Center Uniform Data Set (NACC-UDS) [41] or Wikipedia [11,37]. The usage of publicly available external databases allows research to be reproducible. However, there is no guarantee that these databases have all the tests or scores needed for a particular research, nor to be in the needed language [46]. Language is critical in tests such as the semantic fluency task, as it is language-based. Therefore, tests optimized for English may not be suitable for other languages [47]. This problem is pronounced in certain situations, such as facing a monolingual cohort [3,48].

Finally, ML has also been used to identify the most efficient combination of tests in neuropsychological test batteries to define more concise batteries that take less time to complete and evaluate [49]. This has emerged as a necessity to avoid long waiting times during mass screening [50]. In this line Loewenstein et al. [51] analyzed the performance of English-speaking and Spanish-speaking subjects on different tests, finding that the most relevant tests varied between the two cohorts. Weakley et al. [30] found that the ML methods were able to correctly classify subjects as healthy or MCI and reduce the number of variables needed. Battista et al. [49] used ML methods to analyze the predictive ability between healthy subjects, MCI, and severe impairment of twelve state-of-the-art neuropsychological tests, obtaining four of them as best predictors. Wang et al. [52] optimized a neuropsychological test battery through ML methods, managing to cut it to six tests, reducing the completion time by half while maintaining a good relationship between precision and recall. Garcia-Gutierrez et al. [53] analyzed several neuropsychological tests using evolutionary algorithms, reducing the number of tests and scores. However, these studies focused on already-defined variables, without considering possible emerging variables.

Inspired by these previous works, this study aims to obtain the most efficient (in terms of optimization between the time required to be performed and evaluated and diagnostic ability) combination of test variables of the semantic fluency test able to discern between healthy and MCI subjects. A cohort of Spanish monolingual subjects will be used, as the final goal is to design a reduced battery that could be used on real population screenings in the region. Because the variables' scores are obtained automatically, only the number of tests will be considered as a time-limiting factor. This is done under the perspective of a population screening scenario, where the focus lies on identifying the smallest and therefore, faster to complete, and most efficient combination of tests and variables. Several ML systems (a total of six, described in Section 3.4 will be used together to avoid the results being conditioned by the chosen ML algorithm.

3. Methods

3.1. Overview of the proposal

Fig. 1 shows the scheme of the different steps followed in this proposal. First, the database was digitized. Next, the variables were defined, and their scores were obtained automatically from the digitized database. A relevance analysis was carried out to define an initial split between highly relevant variables and low-relevant variables to MCI diagnosis. These findings were used to determine the test-variable combinations, which were subsequently analyzed and compared using different ML methods to identify those with the best performance and the lower number of tests. The ML method with the best general results across the different combinations was also identified.

3.2. Dataset description

3.2.1. Database

We used an anonymized database from a large ongoing longitudinal study about the prevalence of MCI in the Autonomous Community of Madrid (Spain) [2,54–56]. Subjects with previous diagnoses of neurodegenerative disease, disabling chronic disease, psychiatric disorders such as major depression, established neurological abnormality, severe sensory impairment, diabetes, stroke, and loss of consciousness were already discarded. The cognitive and emotional status of the subjects was assessed using the Spanish version of the Mini-Mental State Examination [57] and the Geriatric Depression Scale [58]. The study gathered data from a total of 141 Spanish monolingual subjects, whose ages ranged between 58 and 93 years and with an educational level between 0 and 22 years of study. Each subject underwent three assessments, spaced approximately one year apart, resulting in diagnoses of Healthy and MCI. This process yielded a total of 426 assessments, which we considered as independent in this study to make the most of the small sample. Using these three assessments per subject, we distinguished three groups, which allowed us to approach the diagnoses of stable cases differently from those in the early stages:

- Healthy subjects (Healthy): diagnosis of Healthy throughout the three assessments.
- Stable MCI subjects (MCI_s): diagnosis of MCI throughout the three assessments.
- Heterogeneous MCI subjects (MCI_h): alternating diagnosis between "Healthy" and "MCI" throughout the three assessments. It is considered that they will evolve into MCI in the future.

Despite the interest in differentiating between the subjects with stabilized MCI and those with suspected MCI, it was considered more critical in this analysis to adequately discern between Healthy subjects and those with some degree of MCI symptomatology. Thus, more weight was given to those analyses that evaluate the performance of models discerning between Healthy and both types of MCI separately. Table 1 summarizes the sociodemographic characteristics of our database.

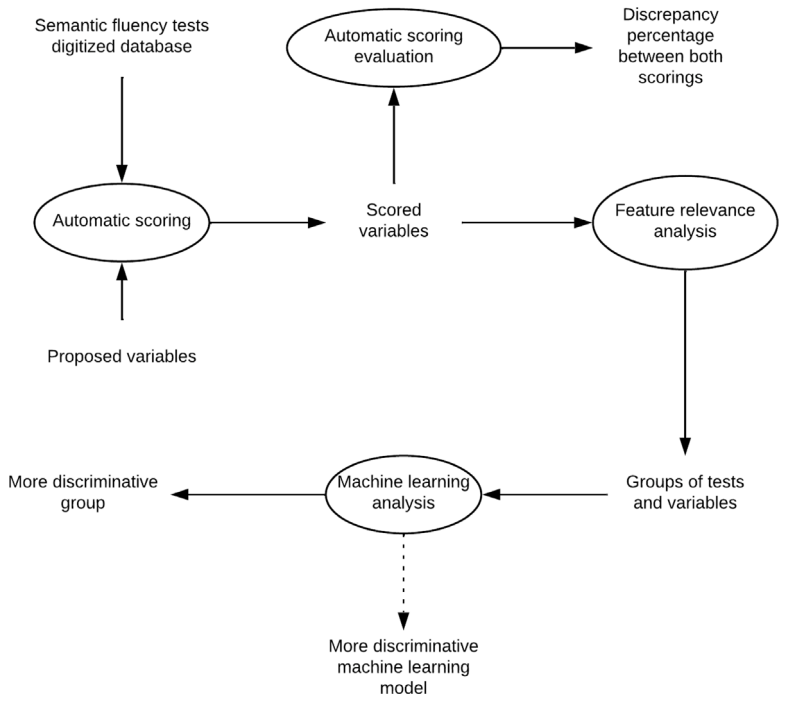

Fig. 1. Scheme of the methodology followed from the initial step of digitalizing the databases to the obtention of the most efficient combination.

Table 1
Summary of the sociodemographic variables of the database, including the MMSE performance.

	Nº of assessments	Men/Women	Age Mean(std)	Schooling Mean(std)	MMSE Mean (std)
Healthy	114	36/108	70.42 (5.55)	11.27 (5.00)	33.08 (1.97)
MCI_h	162	39/123	71.67 (6.96)	9.61 (4.79)	32.22 (2.26)
MCI_s	117	30/87	72.39 (5.95)	6.67 (6.64)	29.27 (3.11)

3.2.2. Semantic fluency test

In this study, four semantic fluency tests were carried out. Subjects were asked to list, in Spanish, the maximum number of terms without repetition in 60 s within a given semantic category (*animals*, *clothes*, *plants*, and *vehicles*). To process the assessment reports, which were on paper, they were first digitized. The terms were manually transcribed in order of occurrence, without omitting *perseverations* or *intrusions*, as no OCR method could deal with handwritten texts with enough accuracy. For poorly readable words, the consensus of the research group was considered.

3.2.3. Feature variables

A total of eight feature variables were defined: seven hypothesized significant variables (*switching*, *clustering*, *total clusters*, *perseverations*, *perseverations with inflections*, *intrusions*, and *phonetic continuity*) along with the traditional scoring method (*corrects*). The criterion defined for obtaining each variable is detailed in the following sections.

Corrects

It is the count of valid and unique terms within the given semantic category. This is the traditional scoring method of semantic fluency tests. However, some terms presented discrepancies in whether they should be considered as *corrects*, *perseverations*, or *intrusions* in all semantic categories, and their final classification is normally left at the discretion of the evaluator.

For example, in the case of *animals*, the Spanish term "pájaro" ("bird", usually only of small size), is considered a subcategory within *animals* because it encompasses several more specific terms. Therefore, it could be considered as *intrusions* (since it refers to several animals at once, instead of being a concrete animal or species), *perseverations* (in case there are animals within that subcategory, such as "chicken" or "goose") or *corrects* depending on the evaluator. Another example is the word "pez" ("fish"). However, this did not happen with other words that could encompass a more specific subset of animals, where all the terms mentioned are taken as *corrects*, even if both subcategories and particular items inside that subcategory appear within the same sequence. Examples could be "dog" with "Yorkshire" and "labrador", "monkey" with "chimpanzee" and "gorilla", or "eagle" with "imperial eagle". This generates inconsistencies between assessments and even within the same assessment. In our database, the usual procedure in most tests was to consider those subcategories as *corrects* regardless of the presence of more concrete words during semantic production. We used this criterion for the automatic scoring process.

In the other three semantic fluency tests, we only found discrepancies in the assessment criterion in terms belonging to certain specific subcategories. These are the cases of terms that could refer to accessories in *clothes*, some fruits and vegetables that could reference both the plant and the consumable product in *plants*, and animals used as transport in *vehicles*. Depending on the test, items corresponding to these subcategories were deemed *corrects* or *intrusions* according to the majority decision in the manual classification.

Intrusions

Intrusions are terms that fall outside the given semantic category, used in some studies as complementary raw scores of *corrects* [7]. When there were doubts about whether a word should be marked as part of the semantic category, we used the criterion established in the data collection.

Perseverations

Perseverations are repetitions of correct words during the verbal production. Although not usually included in clinical assessments, several research of this variable have reported a relationship with neurological

disorders including MCI associated with AD [7,23]. We considered two types of *perseverations* in this study: pure *perseverations* and *perseverations with inflections*. Pure *perseverations* are words that appear at least twice within the same verbal production. *Perseverations with inflections* are those in which the duplicated word has a lexical inflection. The most common inflections were gender (*animals*) and number (all semantic categories).

Switching, Clustering and Total Clusters

Switching, *clustering*, and *total clusters* scores are based on the tendency of subjects to spontaneously generate successive items within the same semantic subcategory, and then switch to another during verbal production [9,10,15,16,25]. To obtain those scores, it was first necessary to define the semantic subcategories of each semantic category. In the literature reviewed, only the subcategories corresponding to *animals* were publicly available [9,15,16]. Therefore, two strategies were employed, one for the *animals* category and another for the *clothes*, *plants*, and *vehicles* categories. For *animals*, we reviewed the semantic subcategories defined by other research groups [9,15,25], adapting them to our context. Those subcategories with low representation in our cohort were discarded, introducing those relevant to our context [7,9,15,25]. This way we ensured that these divisions are faithful to the sociocultural context of our database [9,15].

For the semantic categories of *clothes*, *plants*, and *vehicles*, semantic subcategories were manually created from scratch following the methods described in Troyer [25] as no previous references were publicly available. These subcategories were defined as continuous sequences of semantically related words generated naturally during verbal production. The tables in the APPENDIX show the semantic subcategories defined for the four semantic categories. From these subcategories, switches, and clusters can be obtained. A cluster was defined as a successive grouping of words of the same semantic subcategory, while a switch was a change between clusters. The size of a cluster was calculated by counting the number of words inside the cluster, starting from the second word [25]. If a cluster of a semantic subcategory was inside a larger cluster, only the larger cluster was considered. For example, in the sequence "elephant, leopard, lion, giraffe, there are two clusters: "African animals" and "felines". However, as the cluster "felines" is embedded in the "African animals" cluster, only the latter is considered in the evaluation. If consecutive words in a list belonged to two or more common subcategories, these words were included in both clusters and a switch was scored. For example, in the sequence "dog, cat, tiger", two clusters of size one are identified: "dog-cat" and "cat-tiger" [24].

Clustering, *switching*, and *total clusters* scores were obtained from these criteria. *Clustering* was the average score of all the clusters of a given sequence. *Switching* corresponded to the number of total jumps between semantic subcategories produced within the same verbal production [25]. Finally, the new score of *total clusters* was defined as the number of different clusters presented in a verbal production. Following the indications of the evaluators, we did not include *intrusions* in these counts.

Phonetic continuity

Phonetic continuity, understood as the production of clusters of word sequences beginning with the same letters [15,24,59], has been proposed as a complementary strategy to word generation-based on semantic subcategories in semantic fluency tests. However, a low utility of this variable in differentiating between Healthy and MCI subjects has been reported [7,10]. It is important to note that these studies have been conducted in the English language, where there is no homogeneous relationship between spelling and phoneme, so the natural production of terms initiated with the same letters is disrupted due to the discrepancy between sound and spelling [24]. However, in the Spanish language, each phoneme is associated with a single letter in most cases, so this variable may be more relevant in Spanish than in English. *Phonetic continuity* was calculated as the average of consecutive word sequences beginning with at least the first two letters.

Sociodemographic variables

Although they are not target variables for the analysis of this study, the sociodemographic variables of *age*, *schooling*, and *sex* were included to refine both analysis and predictions made by the ML models. The variable *age* corresponds to the subject's years at the time of testing, *schooling* corresponds to the number of years of education, and *sex* corresponds to the subject's sex, coded in binary form.

3.3. Analysis of the automatic scoring process

To show the improvement obtained by the automatic variable scoring over the manual one, the scores obtained by both methods were compared. We based our analysis on the variable corrects because it was the only variable consistently recorded in all evaluations. All discrepancies were reviewed manually since it was necessary to consult the original reports to identify their origin correctly. These discrepancies were classified into the following categories:

- Undetected *perseverations*: *Perseverations* included as *corrects* in the manual correction.
- Undetected *intrusions*: Words clearly outside the given semantic category included as *corrects* in the manual assessment.
- Subjectivity: Terms evaluated with criteria that vary between evaluations. Only those cases where the criterion was different from those applied in this study were considered (*e.g.*, classifying "fish" or "bird" as *intrusions* or *perseverations*).
- *Perseverations* or *intrusions* not subtracted from *corrects*: Percentage of terms marked as *perseverations* or *intrusions* in the report but included in *corrects* during the manual count. It has been established separately since it was not possible to know if it was a counting error, or if it was finally decided to include these terms deliberately in *corrects*, thus framing it as a problem of subjectivity.
- Counting error: Errors in the manual count. This is considered when there is a discrepancy in the number of correct words, but the word sequence lacks other elements that could explain the discrepancy by any of the other categories described.
- Other reasons: Discrepancies caused by reasons unrelated to the test, such as loss of words in the scan or illegible words. Tests with such discrepancies were discarded when calculating the total percentage of affected tests.

3.4. Identifying the optimal semantic fluency test

The efficiency of a test is measured as the ratio between the accuracy of the test and its cost. To obtain the highest accuracy, we used SOTA ML algorithms on different input configurations, which were defined to minimize the test cost without a significant loss in accuracy. The semantic fluency battery used in this study included four subtests, one subtest per semantic category (*animals*, *clothes*, *plants*, and *vehicles*). This way, we identified two stages when defining the input combinations: the number of semantic categories used in the test battery (test groupings) and the variables used to describe each test (variable groupings). The number of tests directly affects the cost of the test both temporally and economically, not the number of variables, as their scores are obtained automatically. However, the number of variables could influence the accuracy of the ML model.

The following groupings of tests and variables were established to evaluate their relevance in discriminating between Healthy, MCI_h, and MCI_s:

- Test groupings: Each of the four semantic categories serves as an individual test within the overall semantic fluency assessment. As the main objectives are to reduce the number of tests to be performed by a person and to reduce the acquisition time without affecting the diagnosis, combinations with one, two, three, and four tests were analyzed (in total, 15 different combinations). Each test is identified with the initial letter of the semantic category analyzed (A, *animals*; C, *clothes*; P, *plants*; and V, *vehicles*). Only within these combinations, related to the time necessary to perform the tests, does it make sense to evaluate different combinations of variables, which will be related to a higher performance of the ML algorithms.
- Variable groupings: Three groups of variables were defined, each group being identified with a number: (1) all proposed variables, (2) only relevant variables, and (3) only the variable corrects, which would be equivalent to the traditional evaluation method. To determine which variables to include in Group 2, a relevance analysis of their discriminance between Healthy/MCI was performed. This division into variable groups was necessary due to the high number of variables used (a total of 8), which generates a higher number of combinations.

These groupings were combined to define a total of 45 combinations of tests and variables. Each combination was identified by a series of letters and a number. The letters correspond to the included semantic fluency tests and the number to the variable grouping used in that combination. For example, the combination AP-2 includes the semantic fluency tests of *animals* and *plants* and the relevant variables. Sociodemographic variables of *sex*, *age*, and *schooling* were always included.

To ensure that the selected combination is the most efficient, understanding this as the one that obtains the best results with the lowest time cost in general terms and the result is not conditioned by the chosen ML method, the combinations were analyzed by six ML methods with different foundations. The selected ML methods were ADABoost (ADAB), Bagging (BAG), Random Forest (RF), Logistic Regression (RLog), Support Vector Machine (SVM) and XGBoost (XGB). To get the most out of our sample, 10-fold cross-validation was used to evaluate the performance of the predictive models. As a measure to ensure the robustness of the results and to avoid possible biases, the analysis was repeated ten times with different randomization in each repetition.

We used the *F1-score*, the harmonic mean between *recall* and *precision*, as the main metric to evaluate and compare the results. There are several motivations for this choice. One is that neither *precision* nor *recall* alone can describe the general efficiency of a classifier since a good performance in one of those metrics does not necessarily mean a good performance in the other [5,60]. *F1* ignores *true negatives*, so it is suitable for situations where the main interest is to correctly identify the positive class, such as in medical situations [61]. *F1* is a metric capable of reflecting a possible bias of ML methods concerning the positive class in unbalanced datasets [61,62]. This makes *F1* the most appropriate one when comparing different ML models.

In this study, the MCI_h group was established for those subjects with alternating assessments. Mixing those cases with the pure MCI subjects would double the number of cases in that profile regarding the Healthy profile, greatly unbalancing the sample. However, a multiclass classification would give equal importance to all profiles, and there is greater interest in obtaining the combination that best differentiates Healthy subjects and subjects with incipient or suspected MCI than between different degrees of MCI. Taking that into account, we decided to perform rankings between two profiles, and then integrate those results by assigning different weights to the results of each comparison (Healthy-MCI_s, Healthy-MCI_h, and MCI_h-MCI_s) as it can be seen in Eq. (1):

$$\begin{aligned} F1_w(x_i, y_j) = (&\alpha \cdot F1_{Healthy-MCIs}(x_i, y_j) \\ &+ \beta \cdot F1_{Healthy-MCIs}(x_i, y_j) \\ &+ \beta \cdot F1_{Healthy-MCIh}(x_i, y_j) \\ &+ \gamma \cdot F1_{MCIh-MCIs}(x_i, y_j))/3 \end{aligned} \quad (1)$$

where x_i is a concrete configuration of tests and variables, y_j is an ML model, $F1_w$ is the weighted *F1*, $F1_{Healthy\text{-}MCIs}$ is the *F1* resulting from the comparison between Healthy and MCI_s, $F1_{Healthy\text{-}MCIh}$ is the *F1* from the comparison between Healthy and MCI_h, $F1_{MCIh\text{-}MCIs}$ is the *F1* from the comparison between MCI_s and MCI_h, α is the weight associated to $F1_{Healthy\text{-}MCIs}$, β is the weight associated to $F1_{Healthy\text{-}MCIh}$, and γ is the weight associated to $F1_{MCIh\text{-}MCIs}$.

Since the purpose of this study is to obtain the configuration that best separates the Healthy subjects from those with MCI or possible MCI, greater weights were assigned to the results that involved discerning between Healthy and MCI_h or MCI_s than between MCI_h and MCI_s. We used the values $\alpha = \beta = 0.4$ and $\gamma = 0.2$.

3.5. Implementation details

All the implemented code for this study, including the code relating to the automatic extraction of the variables score, was written in Python 3. The ML models were trained using the Scikit-learn 3.4 libraries [63]. For the *perseverations with inflection* score, we jointly used Python's natural language processing package NTLK (Natural Language Toolkit, available at https://www.nltk.org/) to obtain word roots. However, the lemmatizer had a certain degree of error in identifying suffixes and therefore did not return the actual word root. Hence, to determine the similarity between two letter sequences obtained by the lemmatizer we used a gestalt pattern matching variant for Python, the difflib SequenceMatcher module (available at https://docs.python.org/es/3/library/difflib.html). This method allowed us to measure the similarity between resulting roots, giving a value in the range [0,1] where the more similar two words are the higher the score. Specific exceptions were established for false positives caused by similar spellings referring to unrelated items, such as different animals with the same root when the last letter is removed. Examples are "caballo" ("horse") and "caballa" ("mackerel") or "mosquito" ("mosquito") and "mosquita" ("gnat"). Exceptions were also established for certain false negative cases, such as those caused by terms that refer to the same element, but whose difference in the length of the sequence and/or the letters that compose it generates a score that is too low. Examples are "pez" ("fish") and "fishes" ("peces") which are singular and plural, and "rosa" ("rose") and "rosal" ("rosebush") or "bici" ("bike") and "bicicleta" ("bicycle"), where both cases are pure synonyms.

4. Results

4.1. Comparison between manual and automatic scoring

Table 2 shows the percentage of errors made during manual collection of the **corrects** variable. The last column indicates the total percentage of tests affected with at least one of these discrepancies leaving out those cases that fall under "Other reasons". Note that the percentage of "Affected tests" was not the arithmetic sum of all the columns because the same test could have several discrepancies simultaneously, especially in long word sequences.

An exhaustive review of the results obtained by the automatic system was conducted to ensure error-free operation. Consequently, the results obtained by this system have been considered as the gold standard. Table 3 shows the average number of errors per sequence of the manual method concerning our system for the *corrects* variable. The average number of words in each semantic category was included for reference.

Table 2
Percentage of affected tests with at least one discrepancy. Of these discrepancies, the percentage corresponding to perseverations and intrusions not detected or not subtracted, subjectivities, counting errors, and digitization errors is detailed.

	Affected tests	Undetected *perseverations*	Undetected *intrusions*	Subjectivity	*Perseverations* or *intrusions* not subtracted	Other reasons	Counting error
Animals	20.1%	36.2%	2.0%	19.5%	11.4%	1.3%	29.5%
Clothes	17.2%	29.9%	1.4%	23.6%	21.5%	1.4%	22.2%
Plants	17.9%	24.4%	14.1%	30.4%	11.9%	4.4%	14.8%
Vehicles	17.5%	20.0%	5.7%	30.7%	18.6%	1.4%	23.6%

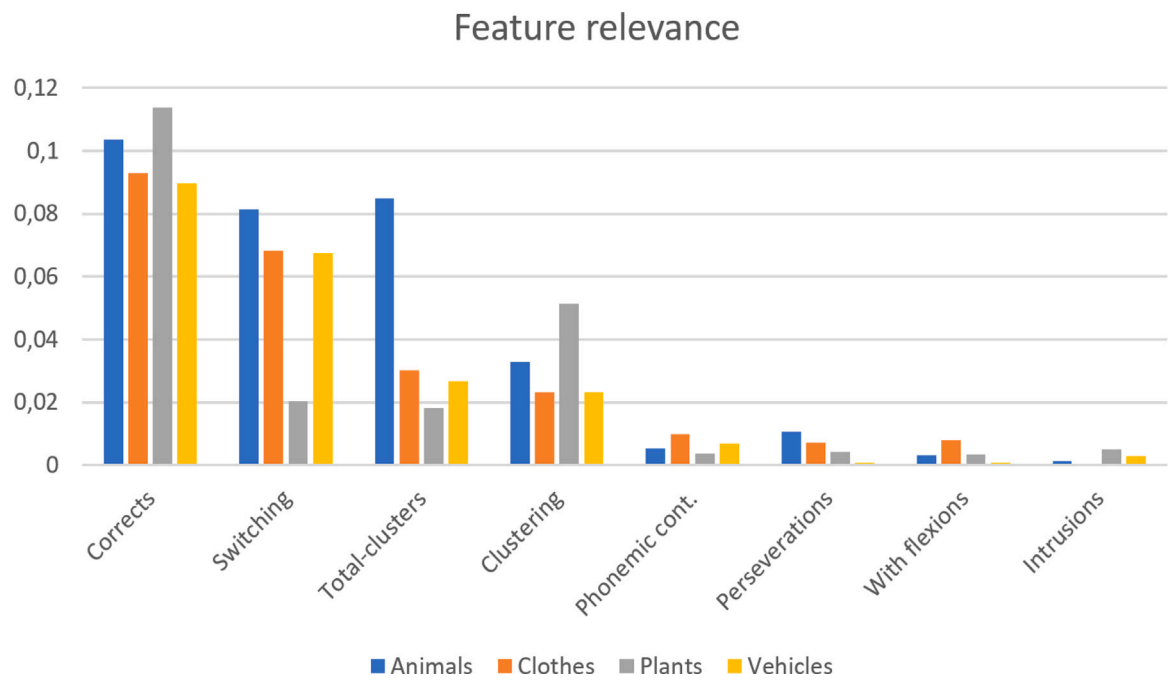

Fig. 2. Relevance of the variables using the feature selection based on the random forest used to discriminate between the diagnoses of the subjects, grouped by variable.

Table 3
Mean and standard deviation of terms erroneously classified as corrects in the manual count, together with the average number of words generated in each semantic category.

	Animals	*Clothes*	*Plants*	*Vehicles*
Mean Error	1,197	1.356	1.387	1.231
Std Error	0.619	0.954	1.292	0.650
Word Average	17	17	12	12

4.2. Analysis of variable relevance

Three methods were used: the k highest scores using both the ANOVA and the mutual information for a discrete target, and the feature importance provided by computing the mean and standard deviation based on the impurity within trees in a random forest. All calculations were performed using Scikit-Learn. On all methods, the variables of *perseverations*, *perseverations with inflections*, *intrusions*, and *phonetic continuity* showed very low relevance in all semantic categories. The *corrects* variable is positioned in all tests and methods as the most relevant variable for discriminating between subjects, followed by *switching*, *total clusters*, and *clustering* variables. From these results, the variables included in the group of "highly discriminating variables" were *corrects*, *switching*, *total clusters*, and *clustering*. Fig. 2 shows the results of the feature selection analysis for the feature selection based on the impurity within trees in a random forest.

4.3. Identifying the optimal semantic fluency test using ML

Using the combinations defined in the 3.4 section, and with sociodemographic variables included in all cases, Table 4 shows the ten highest $F1_w$ combination scores, along with the breakdown by ML models. The table shows the AC-2 combination (*animals* and *clothes* tests with relevant variables) as the most discriminant one with $F1_w$ (AC-2, -) = 0.669, followed by AC-1 ($F1_w$ (AC-1, -) = 0.668). The most efficient ML method for the AC-2 combination was RF, closely followed by BAG and SVM ($F1_{RF}$ = 0.694; $F1_{BAG}$ = 0.689; $F1_{SVM}$ = 0.687). Focusing on the performance of ML models, SVM emerged as the top performer overall, followed by RF and BAG. The scores of the three other methods consistently fell below one or more of the other three. The best ML results for each configuration are highlighted in Table 4.

Table 5 shows the classification of the 10 best combinations. All combinations obtained their best results for the Healthy-MCI_s classification, followed by the Healthy-MCI_h classification. It is worth noticing the presence of the *clothes* test in all groups and the absence of variable group 3 (only *corrects*).

5. Discussion

The automatic collection of variable scores avoided most recurrent errors associated with manual evaluation. It also avoided subjectivity by applying the same criteria to all assessments [3,11]. The analysis of the relevance of the variables highlighted *corrects* as the most discriminative, followed by *switching*, *clustering*, and *total clusters*. On the contrary, the variables of *perseverations*, *perseverations with inflections*, and *intrusions* were classified as low relevant, as in the studies of Clark et al. [7], and in contrast with Pakhomov et al. [23] who did found *perseverations* to be discriminative in the *animals* task. *Phonetic continuity* also presented reduced relevance, indicating that it was not a strategy widely used by the subjects as in the studies of Clark et al. [7,24].

Table 4
Top 10 $F1_w$ for all combinations, as well as the mean value of each ML system for each combination.

Combination	$F1_w$	$F1_{(x,\ ADAB)}$	$F1_{(x,\ BAG)}$	$F1_{(x,\ RF)}$	$F1_{(x,\ RLog)}$	$F1_{(x,\ SVM)}$	$F1_{(x,\ XGB)}$
AC-2	66.9%	62.7%	68.9%	**69.4%**	64.7%	68.7%	67.1%
	(0.13%)	(0.18%)	(0.11%)	(0.07%)	(0.08%)	(0.24%)	(0.13%)
AC-1	66.8%	61.7%	69.7%	68.6%	63.4%	**70.8%**	66.6%
	(0.14%)	(0.26%)	(0.13%)	(0.09%)	(0.10%)	(0.12%)	(0.18%)
C-2	65.9%	61.6%	68.0%	68.4%	61.8%	**68.9%**	66.7%
	(0.11%)	(0.25%)	(0.08%)	(0.06%)	(0.08%)	(0.18%)	(0.12%)
CPV-2	65.8%	60.7%	66.3%	66.8%	64.7%	**69.7%**	66.6%
	(0.11%)	(0.31%)	(0.08%)	(0.07%)	(0.07%)	(0.11%)	(0.12%)
ACP-2	65.6%	61.9%	**67.8%**	67.6%	64.1%	66.0%	66.3%
	(0.13%)	(0.22%)	(0.13%)	(0.12%)	(0.10%)	(0.12%)	(0.11%)
C-1	65.4%	62.1%	**68.6%**	67.0%	60.8%	68.1%	65.8%
	(0.14%)	(0.35%)	(0.08%)	(0.14%)	(0.06%)	(0.12%)	(0.10%)
CP-2	65.4%	61.7%	68.0%	**68.2%**	62.6%	65.0%	66.8%
	(0.10%)	(0.22%)	(0.10%)	(0.07%)	(0.07%)	(0.10%)	(0.11%)
ACPV-2	65.4%	59.9%	66.0%	66.1%	65.3%	**68.4%**	66.4%
	(0.09%)	(0.12%)	(0.07%)	(0.09%)	(0.08%)	(0.08%)	(0.09%)
ACV-2	65.2%	59.3%	66.8%	66.9%	64.1%	**68.6%**	65.7%
	(0.11%)	(0.15%)	(0.16%)	(0.10%)	(0.07%)	(0.10%)	(0.13%)
ACPV-1	65.2%	61.3%	65.8%	64.8%	64.4%	**67.7%**	67.2%
	(0.12%)	(0.24%)	(0.10%)	(0.06%)	(0.10%)	(0.18%)	(0.08%)

Table 5
Ten best $F1_w$ combinations, where $F1_w$ is the weighted mean of the ten repetitions, $F1_{H\text{-}MCIs}$ is the mean of the *F1* of all ML methods for the comparison between Healthy and MCI_s, $F1_{H\text{-}MCIh}$ is the mean of the *F1* of all ML methods for the comparison between Healthy and MCI_h, and $F1_{MCIh\text{-}MCIs}$ is the mean of the *F1* of all ML methods for the comparison between MCI_h and MCI_s. The variances have been added in parentheses.

Combination	$F1_w$	$F1_{Healthy\text{-}MCIs}$	$F1_{Healthy\text{-}MCIh}$	$F1_{MCIs\text{-}MCIh}$
AC-2	66.9% (0.13%)	72.6% (0.10%)	0.668% (0.14%)	0.557% (0.18%)
AC-1	66.8% (0.14%)	73.4% (0.10%)	0.658% (0.16%)	0.557% (0.19%)
C-2	65.9% (0.11%)	71.7% (0.10%)	0.667% (0.10%)	0.527% (0.18%)
CPV-2	65.8% (0.11%)	72.6% (0.10%)	0.643% (0.09%)	0.552% (0.19%)
ACP-2	65.6% (0.13%)	72.2% (0.12%)	0.636% (0.13%)	0.564% (0.16%)
C-1	65.4% (0.14%)	71.8% (0.14%)	0.658% (0.11%)	0.519% (0.14%)
CP-2	65.4% (0.10%)	72.7% (0.10%)	0.637% (0.07%)	0.541% (0.21%)
ACPV-2	65.4% (0.09%)	71.3% (0.07%)	0.641% (0.07%)	0.559% (0.15%)
ACV-2	65.2% (0.11%)	69.8% (0.11%)	0.657% (0.10%)	0.551% (0.16%)
ACPV-1	65.2% (0.12%)	71.8% (0.11%)	0.637% (0.10%)	0.549% (0.22%)

In the analysis to identify the most efficient combination of semantic categories and variables differentiating Healthy from MCI_s and MCI_h, the combination AC-2 (*animals* and *clothes* with the most relevant variables) emerged as the most discriminative ($F1_w$ = 0.669). This is consistent with the fact that combining different tests obtains better results [43]. The RF was the most efficient ML method for that combination with $F1_{RF}$ = 0.694. On the other hand, the combination that would require the least number of tests while keeping an adequate performance, would be C-2 (*clothes* with the most relevant variables) with $F1_w$ = 0.659, the third best combination behind only AC-2 ($F1_w$ = 0.669) and AC-1 ($F1_w$ = 0.668). As Table 4 shows, the traditional evaluation method (*corrects* variable only) does not appear in the top ten results. This indicates that the *corrects* variable alone does not capture all the clinical information of interest presented in these tests. It is striking how all combinations within Group 1 (all variables) always appear below their equivalent Group 2 (relevant variables) combination. For example, the combination AC-1 appears immediately after AC-2; C-1, and ACPV-1 are several positions below C-2 and ACPV-2, and the rest of the combinations are Group 2. This might indicate that low relevant variables introduce noise into the model rather than providing more information.

It is worth noting the presence of *clothes* in the first ten combinations. This might indicate that the semantic category of *clothes* is the most informative semantic fluency test to evaluate the subject's cognitive state. This contrasts with other studies where the *animals* category is usually highlighted as the most discriminative [7,23,59], although only one of the studies reviewed directly compared *animals* and *clothes* [59]. These results can be explained by the fact that they seek to determine how discriminative between Healthy, MCI, and AD were the variables individually, not their joint efficiency. Of the ML methods, the best results corresponded to RF, BAG, and SVM, with very small differences between them in general. The RF ($F1_{RF}$ = 0.694) scored the best result for AC-2, while the best result for AC-1 corresponded to SVM ($F1_{SVM}$ = 0.708). The model that obtained the worst overall results was ADAB (*F1ADAB* = 0.627 for AC-2), therefore being the least suitable for the evaluation of these tests. RF and SVM tend to score the best results for this analysis [42,52].

As expected, all the ML models obtained better results in the comparison between Healthy and MCI_s, the most extreme categories. Similarly, the comparison between MCI_h and MCI_s got the lowest values since both were the closest ones. In general, the cohort used in this study had very mild symptoms in both MCI_h and MCI_s. The good status of both profiles is apparent due to the low rate of *perseverations*, *perseverations with inflections*, and *intrusions*, to the point that those were noise variables for the ML models. The generally low-performance of ML models, particularly in distinguishing between MCI_h and MCI_s, suggests a significant overlap between these groups, underscoring the necessity for a larger dataset to train ML models with greater confidence. This considerable overlap arises from the inclusion of subjects with very weak MCI symptomatology, even within the MCI_s profile, as the aim was to identify these groups in the early stages.

6. Conclusion

In this paper, we have presented a computational tool that automates the collection of variables and analyzes different combinations of semantic categories and associated variables to determine the most efficient configuration. The best one was AC-2 (*animals* and *clothes* with the variables *corrects*, *switching*, *clusters*, and *total clusters*), with RF as the most suitable ML model for that combination. This result was obtained using a database composed by 423 Spanish monolingual instances organized into three groups (healthy, Stable MCI, and

Heterogeneous MCI). Moreover, the automatic collection of variables enabled the elimination of certain common errors associated with the manual score collection, such as counting inaccuracies or subjectivity related to the classification of specific terms. The proposed system could be used as a suitable method during mass population screening, increasing their efficiency and speed of administration and evaluation to reach a larger population. To simulate the cohort of a population screening looking for subjects with early symptomatology, a database composed of healthy, heterogeneous healthy-MCI subjects and stable but very incipient MCI subjects was chosen. This resulted in a high degree of overlap between groups, which explains the performance of the different ML methods.

CRediT authorship contribution statement

Alba Gómez-Valadés: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Rafael Martínez:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Mariano Rincón:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

None declared.

Acknowledgments

We would like to thank Dr. María del Carmen Díaz-Mardomingo and Dr Sara García-Herranz for giving permission to use their databases in our project.

This work was supported by a grant "Ayuda de la UNED para contrato predoctoral para la formación de personal investigador", Spain to Alba Gómez-Valadés as part of the research project presented in this paper.

The authors also gratefully acknowledge the research project CPP 2021-009109 of the Spanish public–private R&D program, Spain.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108955.

References

[1] R.C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, L. Fratiglioni, Mild cognitive impairment: a concept in evolution, Journal of internal medicine 275 (3) (2014) 214–228, http://dx.doi.org/10.1111/joim.12190.

[2] S. García-Herranz, M.C. Díaz-Mardomingo, C. Venero, H. Peraita, Accuracy of verbal fluency tests in the discrimination of mild cognitive impairment and probable Alzheimer's disease in older Spanish monolingual individuals, Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn. (2019) 1–15, http://dx.doi.org/10.1080/13825585.2019.1698710.

[3] A. König, N. Linz, J. Tröger, M. Wolters, J. Alexandersson, P. Robert, Fully automatic speech-based analysis of the semantic verbal fluency task, Dementia Geriatr. Cogn. Disord. 45 (3–4) (2018) 198–209, http://dx.doi.org/10.1159/000487852.

[4] K. López-de Ipiña, U. Martinez-de Lizarduy, P.M. Calvo, B. Beitia, J. García-Melero, E. Fernández, M. Ecay-Torres, M. Faundez-Zanuy, P. Sanz, On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment, Neural Comput. Appl. 32 (2018) 15761–15769, http://dx.doi.org/10.1007/s00521-018-3494-1.

[5] A. So, D. Hooshyar, K. Park, H. Lim, Early diagnosis of dementia from clinical data by machine learning techniques, Appl. Sci. 7 (2017) 651, http://dx.doi.org/10.3390/app7070651.

[6] P. Gurevich, H. Stuke, A. Kastrup, H. Stuke, H. Hildebrandt, Neuropsychological testing and machine learning distinguish Alzheimer's Disease from other causes for cognitive impairment, Front. Aging Neurosci. 9 (2017) http://dx.doi.org/10.3389/fnagi.2017.00114.

[7] D.G. Clark, P.M. McLaughlin, E. Woo, K. Hwang, S. Hurtz, L. Ramirez, J. Eastman, R.-M. Dukes, P. Kapur, T.P. DeRamus, L.G. Apostolova, Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment, Alzheimer's Dementia Diagn. Assess. Dis. Monit. 2 (2016) 113–122, http://dx.doi.org/10.1016/j.dadm.2016.02.001.

[8] P. Goli, E.M. Rad, K. Ghandehari, M. Azarnoosh, Early assessment of mild Alzheimer's Disease using elman neural network, LDA and SVM methods, Mach. Learn. Res. 2 (4) (2017) 148, Number: 4 Publisher: Science Publishing Group.

[9] L. Bertola, M.L. Cunha Lima, M.A. Romano-Silva, E.N. de Moraes, B.S. Diniz, L.F. Malloy-Diniz, Impaired generation of new subcategories and switching in a semantic verbal fluency test in older adults with mild cognitive impairment, Front. Aging Neurosci. 6 (2014) http://dx.doi.org/10.3389/fnagi.2014.00141.

[10] D. Clark, P. Kapur, D. Geldmacher, J. Brockington, L. Harrell, T. DeRamus, P. Blanton, K. Lokken, A. Nicholas, D. Marson, Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease, Cortex 55 (2014) 202–218, http://dx.doi.org/10.1016/j.cortex.2013.12.013.

[11] N. Kim, J.-H. Kim, M.K. Wolters, S.E. MacPherson, J.C. Park, Automatic Scoring of Semantic Fluency, Front. Psychol. 10 (2019) http://dx.doi.org/10.3389/fpsyg.2019.01020.

[12] M.J. Chasles, A. Tremblay, F. Escudier, A. Lajeunesse, S. Benoit, R. Langlois, S. Joubert, I. Rouleau, An examination of semantic impairment in amnestic MCI and AD: What can we learn from verbal fluency? Arch. Clin. Neuropsychol. 35 (1) (2020) 22–30, http://dx.doi.org/10.1093/arclin/acz018.

[13] A. Delgado-Álvarez, J.A. Matias-Guiu, C. Delgado-Alonso, L. Hernández-Lorenzo, A. Cortés-Martínez, L. Vidorreta, P. Montero-Escribano, V. Pytel, J. Matias-Guiu, Cognitive processes underlying verbal fluency in multiple sclerosis, Front. Neurol. 11 (2021) http://dx.doi.org/10.3389/fneur.2020.629183.

[14] K.D. Mueller, R.L. Koscik, A. LaRue, L.R. Clark, B. Hermann, S.C. Johnson, M.A. Sager, Verbal fluency and early memory decline: Results from the wisconsin registry for Alzheimer's prevention, Arch. Clin. Neuropsychol. 30 (5) (2015) 448–457, http://dx.doi.org/10.1093/arclin/acv030.

[15] S. Ramanan, J. Narayanan, T.P. D'Souza, K.S. Malik, E. Ratnavalli, Total output and switching in ategory fluency successfully iscriminates Alzheimer's disease from Mild Cognitive Impairment, but not from frontotemporal dementia, Dementia Neuropsychol. 9 (3) (2015) 251–257, http://dx.doi.org/10.1590/1980-57642015dn93000007.

[16] A.K. Troyer, M. Moscovitch, G. Winocur, L. Leach, M. Freedman, Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease, J. Int. Neuropsychol. Soc. 4 (2) (1998) 137–143, http://dx.doi.org/10.1017/s1355617798001374.

[17] K. Ledoux, T. Vannorsdall, E. Pickett, L. Bosley, B. Gordon, D. Schretlen, Capturing additional information about the organization of entries in the lexicon from verbal fluency productions, J. Clin. Exp. Neuropsychol. 36 (2014) http://dx.doi.org/10.1080/13803395.2013.878689.

[18] F. Paula, R. Wilkens, M. Idiart, A. Villavicencio, Similarity measures for the detection of clinical conditions with verbal fluency tasks, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 231–235, http://dx.doi.org/10.18653/v1/N18-2037.

[19] D. Woods, J. Wyma, T. Herron, E. Yund, Computerized analysis of verbal fluency: Normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury, PLoS One 11 (2016) e0166439, http://dx.doi.org/10.1371/journal.pone.0166439.

[20] L. Bertola, N.B. Mota, M. Copelli, T. Rivero, B.S. Diniz, M.A. Romano-Silva, S. Ribeiro, L.F. Malloy-Diniz, Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls, Front. Aging Neurosci. 6 (2014) http://dx.doi.org/10.3389/fnagi.2014.00185.

[21] J.M. Guerrero, R. Martínez-Tomás, M. Rincón, H. Peraita, Diagnosis of cognitive impairment compatible with early diagnosis of Alzheimer's Disease. a Bayesian network model based on the analysis of oral definitions of semantic categories, Methods Inf. Med. 55 (1) (2016) 42–49, http://dx.doi.org/10.3414/ME14-01-0071.

[22] R.M. Ruff, R.H. Light, S.B. Parker, H.S. Levin, The psychological construct of word fluency, Brain Lang. 57 (3) (1997) 394–405, http://dx.doi.org/10.1006/brln.1997.1755.

[23] S.V. Pakhomov, L.E. Eberly, D.S. Knopman, Recurrent perseverations on semantic verbal fluency tasks as an early marker of cognitive impairment, J. Clin. Exp. Neuropsychol. 40 (8) (2018) 832–840, http://dx.doi.org/10.1080/13803395.2018.1438372.

[24] D. Clark, V. Wadley, P. Kapur, T. DeRamus, B. Singletary, A. Nicholas, P. Blanton, K. Lokken, H. Deshpande, D. Marson, G. Deutsch, Lexical factors and cerebral regions influencing verbal fluency performance in MCI, Neuropsychologia 54 (2014) 98–111, http://dx.doi.org/10.1016/j.neuropsychologia.2013.12.010.

[25] A.K. Troyer, Normative data for clustering and switching on verbal fluency tasks, J. Clin. Exp. Neuropsychol. 22 (3) (2000) 370–378, http://dx.doi.org/10.1076/1380-3395(200006)22:3;1-V;FT370, Publisher: Routledge.

[26] W. Camara, J. Nathan, A. Puente, Psychological test usage: Implications in professional psychology, Prof. Psychol. Res. Pract. 31 (2000) 141–154, http://dx.doi.org/10.1037/0735-7028.31.2.141.

[27] M. Moetesum, I. Siddiqi, U. Masroor, C. Djeddi, Automated scoring of Bender Gestalt Test using image analysis techniques, in: 2015 13th International Conference on Document Analysis and Recognition, ICDAR, 2015, pp. 666–670, http://dx.doi.org/10.1109/ICDAR.2015.7333845.

[28] S.V. Pakhomov, S.E. Marino, S. Banks, C. Bernick, Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency, Speech Commun. 75 (2015) 14–26, http://dx.doi.org/10.1016/j.specom.2015.09.010.

[29] J. Goñi, G. Arrondo, J. Sepulcre, I. Martincorena, N. Vélez de Mendizábal, B. Corominas-Murtra, B. Bejarano, S. Ardanza-Trevijano, H. Peraita, D.P. Wall, P. Villoslada, The semantic organization of the animal category: evidence from semantic verbal fluency and network theory, Cogn. Process. 12 (2) (2011) 183–196, http://dx.doi.org/10.1007/s10339-010-0372-x.

[30] A. Weakley, J.A. Williams, M. Schmitter-Edgecombe, D.J. Cook, Neuropsychological test selection for cognitive impairment classification: A machine learning approach, J. Clin. Exp. Neuropsychol. 37 (9) (2015) 899–916, http://dx.doi.org/10.1080/13803395.2015.1067290.

[31] C. Tunvirachaisakul, T. Supasitthumrong, S. Tangwongchai, S. Hemrunroj, P. Chuchuen, I. Tawankanjanachot, Y. Likitchareon, K. Phanthumchinda, S. Sriswasdi, M. Maes, Characteristics of mild cognitive impairment using the thai version of the consortium to establish a registry for Alzheimer's Disease tests: A multivariate and machine learning study, Dementia Geriatr. Cogn. Disord. 45 (1–2) (2018) 38–48, http://dx.doi.org/10.1159/000487232.

[32] J. Hastings, W. Ceusters, M. Jensen, K. Mulligan, B. Smith, Representing mental functioning: Ontologies for mental health and disease, in: Towards an Ontology of Mental Functioning (ICBO Workshop), Proceeedings of the Third International Conference on Biomedical Ontology, 2012, p. 5.

[33] M.K. Wolters, N. Kim, J.-H. Kim, S.E. MacPherson, J.C. Park, Prosodic and linguistic analysis of semantic fluency data: A window into speech production and cognition, in: Interspeech 2016, ISCA, 2016, pp. 2085–2089, http://dx.doi.org/10.21437/Interspeech.2016-420.

[34] T.K. Landauer, D.S. McNamara, S. Dennis, K. Walter, Handbook of Latent Semantic Analysis, Routledge & CRC Press, 2005, ISBN 9781138004191.

[35] T.K. Landauer, S.T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychol. Rev. 104 (2) (1997) 211–240, http://dx.doi.org/10.1037/0033-295X.104.2.211, Place: US Publisher: American Psychological Association.

[36] E. Gabrilovich, S. Markovitch, Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, 6, 2006, Journal Abbreviation: AAAI Publication Title: AAAI.

[37] N. Linz, J. Troger, J. Alexandersson, A. Konig, Using neural word embeddings in the analysis of the clinical semantic verbal fluency task, in: 12th International Conference on Computational Semantics, IWCS, 2017, p. 7.

[38] M. Ansart, S. Epelbaum, G. Bassignana, A. Bône, S. Bottani, T. Cattai, R. Couronné, J. Faouzi, I. Koval, M. Louis, E. Thibeau-Sutre, J. Wen, A. Wild, N. Burgos, D. Dormont, O. Colliot, S. Durrleman, Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review, Med. Imag. Anal. 67 (2021) 101848, http://dx.doi.org/10.1016/j.media.2020.101848.

[39] A.D. Arya, S.S. Verma, P. Chakarabarti, T. Chakrabarti, A.A. Elngar, A.-M. Kamali, M. Nami, A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease, Brain Inf. 10 (1) (2023) 17, http://dx.doi.org/10.1186/s40708-023-00195-7.

[40] S. Grueso, R. Viejo-Sobera, Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review, Alzheimer's Res. Ther. 13 (1) (2021) 162, http://dx.doi.org/10.1186/s13195-021-00900-w.

[41] M. Bucholc, S. Titarenko, X. Ding, C. Canavan, T. Chen, A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia, Expert Syst. Appl. 217 (2023) 119541, http://dx.doi.org/10.1016/j.eswa.2023.119541.

[42] R.P. Adelson, A. Garikipati, J. Maharjan, M. Ciobanu, G. Barnes, N.P. Singh, F.A. Dinenno, Q. Mao, R. Das, Machine learning approach for improved longitudinal prediction of progression from mild cognitive impairment to Alzheimer's Disease, Diagnostics 14 (1) (2024) 13, http://dx.doi.org/10.3390/diagnostics14010013, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[43] R. Franciotti, D. Nardini, M. Russo, M. Onofrj, S.L. Sensi, Comparison of machine learning-based approaches to predict the conversion to Alzheimer's disease from mild cognitive impairment, Neuroscience 514 (2023) 143–152, http://dx.doi.org/10.1016/j.neuroscience.2023.01.029.

[44] J. Amunts, J.A. Camilleri, S.B. Eickhoff, K.R. Patil, S. Heim, G.G. von Polier, S. Weis, Comprehensive verbal fluency features predict executive function performance, Sci. Rep. 11 (1) (2021) 6929, http://dx.doi.org/10.1038/s41598-021-85981-1, Number: 1 Publisher: Nature Publishing Group.

[45] M. Kulmanov, F.Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Brief. Bioinform. 22 (4) (2021) bbaa199, http://dx.doi.org/10.1093/bib/bbaa199.

[46] S.-E. Choi, S. Mukherjee, L.E. Gibbons, R.E. Sanders, R.N. Jones, D. Tommet, J. Mez, E.H. Trittschuh, A. Saykin, M. Lamar, L. Rabin, N.S. Foldi, S. Sikkes, R.J. Jutten, E. Grandoit, C. Mac Donald, S. Risacher, C. Groot, R. Ossenkoppele, P.K. Crane, Development and validation of language and visuospatial composite scores in ADNI, Alzheimer's Dementia Trans. Res. Clin. Interv. 6 (1) (2020) e12072, http://dx.doi.org/10.1002/trc2.12072.

[47] I.M. Lorentzen, J. Espenes, E. Hessen, K. Waterloo, G. Bråthen, S. Timón, D. Aarsland, T. Fladby, B.-E. Kirsebom, Regression-based norms for the FAS phonemic fluency test for ages 40–84 based on a Norwegian sample, Appl. Neuropsychol. Adult 30 (2) (2023) 159–168, http://dx.doi.org/10.1080/23279095.2021.1918128, Publisher: Routledge.

[48] S. García-Herranz, M.C. Díaz-Mardomingo, J.C. Suárez-Falcón, R. Rodríguez-Fernández, H. Peraita, C. Venero, Normative data for verbal fluency, trail making, and rey-osterrieth complex figure tests on monolingual spanish-speaking older adults, Arch. Clin. Neuropsychol.: Off. J. Natl. Acad. Neuropsychol. 37 (5) (2022) 952–969, http://dx.doi.org/10.1093/arclin/acab094.

[49] P. Battista, C. Salvatore, I. Castiglioni, Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study, Behav. Neurol. 2017 (2017) 1850909, http://dx.doi.org/10.1155/2017/1850909.

[50] H. Donovan, E. Ellis, L. Cole, E. Townsend, A. Cases, Reducing time to complete neuropsychological assessments within a memory assessment service and evaluating the wider impact, BMJ Open Qual. 9 (3) (2020) e000767, http://dx.doi.org/10.1136/bmjoq-2019-000767.

[51] D.A. Loewenstein, M.P. Rubert, T. Argüelles, R. Duara, Neuropsychological test performance and prediction of functional capacities among Spanish-speaking and English-speaking patients with dementia, Arch. Clin. Neuropsychol.: Off. J. Natl. Acad. Neuropsychol. 10 (2) (1995) 75–88.

[52] J. Wang, Z. Wang, N. Liu, C. Liu, C. Mao, L. Dong, J. Li, X. Huang, D. Lei, S. Chu, J. Wang, J. Gao, Random forest model in the diagnosis of dementia patients with normal mini-mental state examination scores, J. Pers. Med. 12 (1) (2022) 37, http://dx.doi.org/10.3390/jpm12010037.

[53] F. Garcia-Gutierrez, A. Delgado-Alvarez, C. Delgado-Alonso, J. Díaz-Álvarez, V. Pytel, M. Valles-Salgado, M.J. Gil, L. Hernández-Lorenzo, J. Matías-Guiu, J.L. Ayala, J.A. Matias-Guiu, Diagnosis of Alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms, Int. J. Geriatr. Psychiatry 37 (2) (2022) _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gps.5667.

[54] M. Díaz-Mardomingo, S. García-Herranz, R. Rodríguez-Fernández, C. Venero, H. Peraita, Problems in classifying mild cognitive impairment (MCI): One or multiple syndromes? Brain Sci. 7 (12) (2017) 111, http://dx.doi.org/10.3390/brainsci7090111.

[55] M.C. Díaz-Mardomingo, H. Peraita, Detección precoz del deterioro cognitivo ligero de la tercera edad, Psicothema 20 (3) (2008) 438–444.

[56] H. Peraita, S. García-Herranz, M.C. Díaz-Mardomingo, Evolution of specific cognitive subprofiles of mild cognitive impairment in a three-year longitudinal study, Curr. Aging Sci. 4 (2011) 171–182, http://dx.doi.org/10.2174/1874609811104020171.

[57] A. Lobo, J. Ezquerra, F. Gómez Burgada, J.M. Sala, A. Seva Díaz, [Cognocitive mini-test (a simple practical test to detect intellectual changes in medical patients)], Actas Luso Esp. Neurol. Psiquiatr. Cienc. Afines 7 (3) (1979) 189–202.

[58] J.A. Yesavage, T.L. Brink, T.L. Rose, O. Lum, V. Huang, M. Adey, V.O. Leirer, Development and validation of a geriatric depression screening scale: A preliminary report, J. Psychiatr. Res. 17 (1) (1982) 37–49, http://dx.doi.org/10.1016/0022-3956(82)90033-4.

[59] B. Tessaro, A. Hermes-Pereira, L.P. Schilling, R.P. Fonseca, R. Kochhann, L.C. Hübner, Verbal fluency in Alzheimer's disease and mild cognitive impairment in individuals with low educational level and its relationship with reading and writing habits, Dementia Neuropsychol. 14 (2020) 300–307, http://dx.doi.org/10.1590/1980-57642020dn14-030011, Publisher: Academia Brasileira de Neurologia, Departamento de Neurologia Cognitiva e Envelhecimento.

[60] T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.C. Chatzisavvas, A comparison of machine learning techniques for customer churn prediction, Simul. Model. Pract. Theory 55 (2015) 1–9, http://dx.doi.org/10.1016/j.simpat.2015.03.003.

[61] B. Ozenne, F. Subtil, D. Maucort-Boulch, The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases, J. Clin. Epidemiol. 68 (8) (2015) 855–859, http://dx.doi.org/10.1016/j.jclinepi.2015.02.010.

[62] Z. DeVries, E. Locke, M. Hoda, D. Moravek, K. Phan, A. Stratton, S. Kingwell, E.K. Wai, P. Phan, Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability, Spine J. 21 (7) (2021) 1135–1142, http://dx.doi.org/10.1016/j.spinee.2021.02.007.

[63] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830, http://dx.doi.org/10.48550/arXiv.1201.0490.