

Research Article

Optimizing Neuropsychological Assessments for Cognitive, Behavioral, and Functional Impairment Classification: A Machine Learning Study

Petronilla Battista, Christian Salvatore, and Isabella Castiglioni

Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Segrate, Milano, Italy

Correspondence should be addressed to Isabella Castiglioni; isabella.castiglioni@ibfm.cnr.it

Received 27 May 2016; Revised 7 December 2016; Accepted 21 December 2016; Published 31 January 2017

Academic Editor: Michael Nitsche

Copyright © 2017 Petronilla Battista et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Subjects with Alzheimer's disease (AD) show loss of cognitive functions and change in behavioral and functional state affecting the quality of their daily life and that of their families and caregivers. A neuropsychological assessment plays a crucial role in detecting such changes from normal conditions. However, despite the existence of clinical measures that are used to classify and diagnose AD, a large amount of subjectivity continues to exist. Our aim was to assess the potential of machine learning in quantifying this process and optimizing or even reducing the amount of neuropsychological tests used to classify AD patients, also at an early stage of impairment. We investigated the role of twelve state-of-the-art neuropsychological tests in the automatic classification of subjects with none, mild, or severe impairment as measured by the clinical dementia rating (CDR). Data were obtained from the ADNI database. In the groups of measures used as features, we included measures of both cognitive domains and subdomains. Our findings show that some tests are more frequently best predictors for the automatic classification, namely, LM, ADAS-Cog, AVLT, and FAQ, with a major role of the ADAS-Cog measures of delayed and immediate memory and the FAQ measure of financial competency.

1. Introduction

Dementia is a clinical syndrome which affected more than 35 million people worldwide in 2010, with new estimates of 48.1 million people for 2020 and numbers expected to almost double every 20 years [1]. Alzheimer's disease (AD) represents the primary cause of neurodegenerative dementia [2].

To date, scientists have concentrated on untangling the complex brain changes involved in the onset and progression of AD. However, this pathology is correlated to cognitive impairment, behavioral disturbance, and functional disabilities, which greatly have an impact on the quality of daily life, and is major problem for families, caregivers, and healthcare institutions. It is thus crucial to detect such changes early and to identify the level and the type of impairment in the patients. This could facilitate the provision of optimal support as soon as possible, in order to maintain their quality of life for as long as possible. In addition, early detection enables the disease to be monitored from its initial stage of

disability, possibly administering available treatments when loss of functions is not yet advanced.

Neuropsychological assessment plays a crucial role in detecting loss of cognitive functions and change in behavioral and functional state from normal conditions. Specifically, neuropsychological tests can detect dysfunctions in human "cognitive domains" as a consequence of dysfunctions in different neural networks and subnetworks caused by AD. In 2013, the American Psychiatric Association published the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [3]. DSM-5 defined six key domains of cognitive function, namely, complex attention, executive function, learning and memory, language, perceptual-motor function, and social cognition, and each of these has subdomains. Identifying the domains and subdomains affected in a patient helps in establishing the aetiology and severity of the neurocognitive disorder. Neuropsychological tests can measure different cognitive domains (e.g., language, learning, and memory) and subdomains (e.g., long-term memory and

recognition memory) [4]. However, despite the promising results from several different tests, identifying the best ones, as well as the best combination of tests to be used to classify and diagnose AD, is still a matter of debate, and a large amount of subjectivity continues to exist in the diagnostic process [5, 6]. In fact, even the DSM-5 does not name any proprietary tests. In addition, whether specific tasks are better for detecting impairment than others is still unclear [7]. A long list of neuropsychological tests is thus still considered appropriate and subjects are addressed to intensive testing. Optimizing or even reducing the amount of neuropsychological tests used to classify AD patients, also at an early stage of impairment, may be possible with no additional performance costs, thus reducing the time intensity and cognitive stress of the assessment.

Machine learning (ML) is an advanced computational technique that can be used for automatic classification of subjects with diagnostic purpose. Specifically, methods based on ML are able to learn the relationship between input and response variables of two given classes of subjects (e.g., normal and pathological subjects) and to use the learned model to predict the response variable of a new (independent) subject. ML was first adopted in medicine because of its ability to deal with large and complex datasets [8].

Sophisticated ML methods have been applied in the field of dementia in order to obtain a high level of accuracy in the automatic classification of subject impairment [9]. However, such methods have been extensively used with neuroimaging studies on dementia (e.g., [10]) and there have been few explicit attempts to use ML to assess cognitive, behavioral, and functional measures.

In 2015, Weakley et al. [11] used ML and a combination of twenty-seven measures from a cohort of 272 subjects including cognitive, behavioral, and functional abilities obtained from different neuropsychological tests (e.g., visual and verbal memory and language category fluency) to automatically classify groups of patients with different clinical dementia ratings (CDR), namely, CDR = 0, CDR = 0.5, and CDR = 1+ (i.e., 1 and 2). The CDR Scale is a five-point semistructured interview between the patient and a reliable informant (e.g., caregivers) designed to stage the severity of dementia considering the state of the subject with respect to memory, orientation, judgment and problem solving, community affairs, home/hobbies, and personal care [12]. Weakley et al. [11] envisaged the potential of ML with respect to traditional statistical approaches in fully automating the diagnostic process by reducing time-consuming and subjective manual analyses and producing reliable information on the relationship between input (measures of cognitive, behavioral, and functional abilities) and response variables (CDR score) without the need of defining assumptions on data [13]. The authors used ML (with respect to manual classification) to explore many measure configurations (i.e., many combinations of measures), which was impossible to analyze manually. The ML classifier selected a maximum of six measures able to predict the CDR score (an accuracy of 98%, 82%, and 94% was obtained when classifying CDR = 1+ versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5, resp.). The results showed that ML

was a stable and robust predictive model for a number of approximately 200 participants.

A similar approach was already reported by another study in the literature [14], in which ML was applied to only seven cognitive and behavioral attributes from a database of 765 subjects (together with their educational level and the clinical estimation of the patient insight) to predict CDR scores (0, 0.5, 1, 2, and 3). However, these attributes were selected by the authors and led to poor classification accuracy for the mildly demented severity class (59%).

Given the results obtained by the two above-mentioned studies, it would be interesting to consider the use of more subdomains for the automatic classification, since it could lead to an improvement in the predictive model and classification performance.

In this work, our aim was to assess the potential of automatic classification in optimizing or even reducing the amount of neuropsychological measures to predict cognitive, behavioral, and functional impairment of subjects with AD, even at an early stage. We investigated the role of twelve state-of-the-art neuropsychological tests in the automatic classification of subjects with none, mild, or severe impairment as measured by CDR. Data were obtained from the ADNI database (<http://adni.loni.usc.edu/>). In the groups of measures used as classification features, we included measures of both cognitive domains (language, executive function, memory and learning, and complex attention) and subdomains (perceptual-motor coordination, working memory, and visuoconstructional reasoning). We also included measures of depression and loss of awareness, more specifically, orientation. In order to reduce the dimensionality of the computation, we used the following two approaches: (1) an automated classification following a computational feature reduction method and (2) an automated classification following a feature reduction guided by neuropsychologists.

2. Materials and Methods

2.1. Participants. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu/>). The ADNI was launched in 2003 as a public-private partnership, led by the principal investigator, Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD); see <http://www.adni-info.org/>.

In our study, a total of 324 subjects were considered. This sample is the same as in our previous study [10].

All subjects performed a CDR assessment at follow-up (from at least 18 months to 36 months). The interview included both answers directly obtained from each subject and from the interviewer observing the subjects when performing short and easy tasks. The CDR total score came from the subscores of measures on memory, orientation, capacity of judgment/problem solving, social and business activities, capacity to make home-life/intellectual activities/hobbies,

and personal care. In our work, the CDR score served as a gold standard to classify each participant into one of the following three categories: absence of impairment with CDR = 0, mild impairment with CDR = 0.5, and severe impairment with CDR = 1. In each category, subjects were matched for age and gender. 126 subjects had CDR = 0 at follow-up, 143 subjects had CDR = 0.5, and 55 subjects had CDR = 1 at follow-up.

2.2. Measures Used as Features. All subjects underwent a neuropsychological assessment, starting from the screening visit (visit at time = 0 of the ADNI protocol) up to the baseline (visit at time = 1 month from the screening in the ADNI protocol) and following visits. The ADNI database provided the raw results of this assessment in terms of total and partial test scores.

In our work, we used both total and partial test scores as measures to predict the classification category of subjects.

In order to ensure the independence among features and the gold standard (CDR), we excluded the CDR test from the neuropsychological measures used as features. This avoided circularity in the classification process, thus reducing overfitting.

In the next sections, we present a brief description of the neuropsychological tests and relative measures used as features.

2.2.1. Mini Mental State Examination (MMSE). MMSE is a brief questionnaire developed by Folstein et al. [15] which measures the global cognitive impairment and takes around 15 minutes to complete. It consists of 30 items divided into 6 areas: orientation in time and space; memory (repetition of three words), attention and calculation (serial subtraction or forward/backward spelling, recall of words previously memorized); language (recognition of two objects, repetition of a short sentence; sentence comprehension; sentence writing), and constructional praxis (design copy). In our study we used all of the 32 measures reported in the ADNI database: MMDATE, MMYEAR, MMMONTH, MMDAY, MMSEASON, MMHOSPIT, MMFLOOR, MMCITY, MMAREA, MMSTATE, MMBALL, MMFLAG, MMTREE, MMTRIALS, MMD, MML, MMR, MMO, MMW, MMBALLDL, MMFLAGDL, MMTREEDL, MMWATCH, MMPENCIL, MMREPEAT, MMHAND, MMFOLD, MMONFLR, MMREAD, MMWRITE, MMDRAW, and MMSCORE.

2.2.2. Clock Test. Participants are asked to draw a clock and to set the hands to ten after eleven. Scores are assigned if the symmetry of number placement, correctness of numbers, and hand placement are correct. In our study, we used all of the 12 measures reported in the ADNI database: CLOCKCIRC, CLOCKSYP, CLOCKNUM, CLOCKHAND, CLOCKTIME, CLOCKSCOR, COPYCIRC, COPYSYM, COPYNUM, COPYHAND, COPYTIME, and COPYSCOR.

2.2.3. Logical Memory (LM). This measures declarative/episodic memory by means of a brief story read to the participant

who is asked to retell it from memory immediately. The primary measure of performance is the number of story units recalled. The LM is a subtest of the Wechsler Memory Scale-Revised [16] and one of the most widely used clinical measures of memory. In our study, we used all three measures reported in the ADNI database: LIMMTOTAL, LDELTO-TAL, and LDELTCUE.

2.2.4. Rey Auditory Verbal Learning Test (AVLT). AVLT is a widely used test of anterograde verbal episodic memory. A list of 15 unrelated words is presented orally to the subject [17]. The test consists of 5 consecutive repetitions in order to learn the unstructured verbal material and then a long delay free recall 30 minutes later to verify if subject acquired the words over the course of the 5 trials. Finally, a yes/no recognition trial follows the delayed recall trial. It is possible to obtain a learning score from AVLT using the difference between the last and the first immediate recall trials. The measures that are usually calculated from the AVLT are learning scores (trial 5 minus trial 1), short and long delay recall, and recognition. In our study, we used all of the 18 measures reported in the ADNI database: AVTOT1, AVERR1, AVTOT2, AVERR2, AVTOT3, AVERR3, AVTOT4, AVERR4, AVTOT5, AVERR5, AVTOT6, AVERR6, AVTOTB, AVERRB, AVDEL30MIN, AVDELERRI, AVDELTOT, and AVDELERR2.

2.2.5. Digit Span (DS). DS is a test of working memory with two subscales: forward and backward. In the DS forward, the examiner reads a number sequences of increasing length and asks the participant to repeat them. The total score is the number of sequences correctly repeated. In DS backward, the examiner reads a number sequence of increasing length and then asks the participant to repeat each sequence backward. The primary measure of performance is the number of digit sequences correctly reversed. These two tests are included in the Wechsler Memory Scale-Revised [16]. In our study, we used all five measures reported in the ADNI database: DSPANFOR, DSPANFLTH, DSPANBAC, DSPANBLTH, and DIGITSCOR.

2.2.6. Category Fluency Tests (Animals and Vegetables). This is a widely used measure of the ability to spontaneously generate a set of semantically related words in one minute. The participant is asked to name different examples of a given category and the score is the number of unique examples named. In our study, we used all six measures reported in the ADNI database: CATANIMSC, CATANPERS, CATAN-INTR, CATVEGESC, CATVGPERS, and CATVGINTR.

2.2.7. Trail Making Test A-B (TMT A-B). TMT A-B encompasses two trials, A and B [18]. The first part A is a test of psychomotor processing speed and visual scanning. An array of numbers on a page is presented to the subjects and they are instructed to draw lines connecting the numbers in sequential order within the time allowed. The second part B provides cognitive flexibility measures: psychomotor processing speed, visual scanning, and attentional set shifting. An array of numbers and letters is presented to the subjects and they are asked to draw connecting lines

while alternating between numbers and letters in sequential order. An additional commonly used measure is the time to completion from parts A and part B minus part A. In our study, we used all six measures reported in the ADNI database: TRAASCOR, TRAAERCOM, TRAAERROM, TRABSCOR, TRABERCOM, and TRABERROM.

2.2.8. Boston Naming Test (BNT). BNT assesses naming ability using 30 items [19]. Participants are asked to name a series of visual stimuli (object images) with different frequencies (ranging from high to low). If subjects are not able to come up with the correct answer, they are provided with a cue. A phonemic cue is provided when the participant can recognize the purpose of the object but cannot retrieve the correct name. In our study, we used all six measures reported in the ADNI database: BNTSPONT, BNTSTIM, BNTCSTIM, BNTPHON, BNTCPHON, and BNTTOTAL.

2.2.9. American National Adult Reading Test (ANART). ANART is a widely accepted test to estimate premorbid verbal levels of intelligence in dementing individuals. It consists of 50 orthographically irregular English words. Participants are instructed to pronounce each word aloud, beginning at the top of the list and continuing through to the end. In our study, we used the measure ANARTERR reported within the ADNI database.

2.2.10. Alzheimer's Disease Assessment Scale-Cognitive Behavior (ADAS-Cog). ADAS-Cog is composed of two parts, the noncognitive subscale and the cognitive subscale, and provides a measure index of global cognition. The tests take around 30–40 minutes to administer. Twelve tests are used to evaluate short-term memory (evocation of words; word recognition; learning the instructions of a test); spatial-temporal orientation; language skills (verbal skills, difficulty in naming spontaneous speech, comprehension of spoken language, naming objects and fingers, and execution of commands); praxis; attention and concentration. The rating of the majority of cognitive tests is assigned on the basis of the performance of the patient in each single test, while, in some cases, it is based on clinical estimates carried out by the examiner in the course of conversation and other sessions. The ADAS-Cog scores range from 0, which is equivalent to the absence of problems, to a maximum of 70, which indicates a serious deficit in all tests. For our study, we used all 15 measures reported in the ADNI database: Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q14, TOTAL11, and TOTALMOD.

2.2.11. Geriatric Depression Scale (GDS). GDS is a 30-item self-report assessment used to identify mood changes (i.e., depression) in elderly patients [20]. The examinee has to provide yes/no answers to each item of the GDS. For our study, we used all 16 measures reported in the ADNI database: GDSATIS, GDDROP, GDEMPTY, GDBORED, GDSPRIT, GDAFRAID, GDHAPPY, GDHELP, GDHOME, GDMEMORY, GDALIVE, GDWORTH, GDEN-ERGY, GDHOPE, GDBETTER, and GDTOTAL.

2.2.12. Functional Assessment Questionnaire (FAQ). FAQ is a self-administered functional assessment which provides information on the patient's physical, psychological, social, and role functions. It can be used useful to monitor the patient over time with 0 score corresponding to no impairment and 30 to severely impaired. In our study, we considered all 11 measures reported in the ADNI database: FAQFINAN, FAQFORM, FAQSHOP, FAQGAME, FAQBEVG, FAQMEAL, FAQEVENT, FAQTV, FAQREM, FAQTRAVL, and FAQ total.

Finally, in our work, a total of 131 measures were used. Table 1 shows the entire list of these measures, a short description of what they represent together with the reference tests.

2.3. Feature Normalization. Raw scores and subscores were first normalized as z -scores, using the following formula:

$$z\text{-score} = \frac{(\text{score} - m)}{s}, \quad (1)$$

where score represents the raw score or subscore of a given test and m and s represent the mean and standard deviation, respectively, of the raw score of the subjects.

2.4. Feature Reduction. Feature reduction was applied in order to reduce the number of features to be classified without losing relevant information, which resulted in an improvement in computational performance. Two different approaches were implemented: (a) a computational approach, based on the mathematical discriminatory power of features among classes, and (b) an approach based on our basic understanding of the redundancy of features. More specifically, in this last approach, our cognitive understanding of the disease guided the model.

(a) Computation-Based Feature Reduction. The class discriminatory power of z -scored features was estimated in terms of Fisher's Discriminant Ratio (FDR) as follows:

$$\text{FDR} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \quad (2)$$

where μ_i and σ_i^2 are the mean and the variance of i th class, respectively.

(b) Feature Reduction Guided by the Neuropsychologists. Two experienced neuropsychologists were asked to reduce the number of features on the basis of three primary considerations: (1) redundancy (if the same or similar measures are derived from two or more cognitive tests included in the ADNI database: for example, item Q7 of ADAS overlaps with the MMSE items assessing spatiotemporal disorientation); (2) overlap with CDR (if the same or similar measures are present in the CDR interview which is used as a gold standard for the classification: this could produce bias in the classification performance); (3) poor relevance to AD (based on scientific literature).

The features included in our classification fell within the following domains: global cognitive status, orientation,

TABLE 1: List of neuropsychological measures used as features.

Measure	Description
1. MMDATE	What is today's date?, MMSE
2. MMYEAR	What year is it?, MMSE
3. MMMONTH	What month is it?, MMSE
4. MMDAY	What day of the week is it today?, MMSE
5. MMSEASON	What season are we in?, MMSE
6. MMHOSPIT	What is the name of this hospital (clinic, place)?, MMSE
7. MMFLOOR	What floor are we on?, MMSE
8. MMCITY	What town or city are we in?, MMSE
9. MMAREA	What county (district, borough, area) are we in?, MMSE
10. MMSTATE	What state are we in?, MMSE
11. MMBALL	Ball, MMSE
12. MMFLAG	Flag, MMSE
13. MMTREE	Tree, MMSE
14. MMTRIALS	Enter number of trials, MMSE
15. MMD	D, MMSE
16. MML	L, MMSE
17. MMR	R, MMSE
18. MMO	O, MMSE
19. MMW	W, MMSE
20. MMBALLDL	Ball, MMSE
21. MMFLAGDL	Flag, MMSE
22. MMTREEDL	Tree, MMSE
23. MMWATCH	Show the participant a wrist watch and ask "what is this?", MMSE
24. MMPENCIL	Repeat for pencil, MMSE
25. MMREPEAT	Say "repeat after me: no ifs, ands, or buts.", MMSE
26. MMHAND	Takes paper in right hand, MMSE
27. MMFOLD	Folds paper in half, MMSE
28. MMONFLR	Puts paper on floor, MMSE
29. MMREAD	Present the piece of paper which reads "CLOSE YOUR EYES," and say "read this and do what it says", MMSE
30. MMWRITE	Give the participant a blank piece of paper and say "write a sentence.", MMSE
31. MMDRAW	Present the participant with the Construction Stimulus page. Say "copy this design.", MMSE
32. MMSCORE	MMSE total score, MMSE
33. CLOCKCIRC	Approximately circular face, CLOCK
34. CLOCKSVM	Symmetry of number placement, CLOCK
35. CLOCKNUM	Correctness of numbers, CLOCK
36. CLOCKHAND	Presence of the two hands, CLOCK
37. CLOCKTIME	Presence of the two hands, set to ten after eleven, CLOCK
38. CLOCKSCOR	Total score, CLOCK
39. COPYCIRC	Approximately circular face, CLOCK
40. COPYSVM	Symmetry of number placement, CLOCK
41. COPYNUM	Correctness of numbers, CLOCK

TABLE 1: Continued.

Measure	Description
42. COPYHAND	Presence of the two hands, CLOCK
43. COPYTIME	Presence of the two hands, set to ten after eleven, CLOCK
44. COPYSCOR	Total score, CLOCK
45. LDELFCUE	Use of cue (0/1), LM
46. LDELTOTAL	Total number of story units recalled, Partial Score of LM test
47. LIMMTOTAL	Total number of story units recalled, LM Immediate Recall
48–53. AVTOT1-6	Total of each trial 1, 2, 3, 4, 5, 6, AVLT
54–59. AVERR1-6	Total intrusions of each trial 1, 2, 3, 4, 5, 6, AVLT
60. AVTOTB	Interference, AVLT
61. AVERRB	Total intrusions of List B, AVLT
62. AVDEL30MIN	30 minute delay, AVLT
63. AVDELERR1	Total intrusions, AVLT
64. AVDELTOT	Recognition, AVLT
65. AVDELERR2	Recognition errors, AVLT
66. DSPANFOR	Forward: Total Correct
67. DSPANFLTH	Forward: Length
68. DSPANBAC	Digit Span Backwards, Total Correct
69. DSPANBLTH	Backward: Length
70. DIGITSCOR	Digit Symbol
71. CATANIMSC	Category Fluency—Animals
72. CATANPERS	Category Fluency Animals—Perseverations
73. CATANINTR	Category Fluency (Animals)—Intrusions
74. CATVEGESC	Category Fluency Vegetables—Total Correct
75. CATVGPERS	Category Fluency (Vegetables)—Perseverations
76. CATVGINTR	Category Fluency (Vegetables)—Intrusions
77. TRAAERRCOM	Errors of commission, TMT
78. TRAAERRROM	Errors of omission, TMT
79. TRAASCOR	Part A—time to complete, TMT
80. TRABERRCOM	Error of commission, TMT
81. TRABERRROM	Error of omission, TMT
82. TRABSCOR	Part B—time to complete, TMT
83. BNTSPONT	Number of spontaneously given correct responses, Partial Score of BNT
84. BNTSTIM	Number of semantic cues given, Partial Score of BNT
85. BNTCSTIM	Number of correct responses following a semantic cue, Partial Score of BNT
86. BNTPHON	Number of phonemic cues given, Partial Score of BNT
87. BNTCPHON	Number of correct responses following a phonemic cue, Partial Score of BNT
88. BNTTOTAL	Total Number Correct (1 + 3)
89. ANARTERR	ANART Total Score (total number of errors)

TABLE 1: Continued.

Measure	Description
90. Q1	Word Recall Task, ADAS-Cog
91. Q2	Following commands, ADAS-Cog
92. Q3	Constructional praxis, ADAS-Cog
93. Q4	Delayed Word Recall, ADAS-Cog
94. Q5	Naming objects and fingers, ADAS-Cog
95. Q6	Ideational practice, ADAS-Cog
96. Q7	Orientation, ADAS-Cog
97. Q8	Word recognition, ADAS-Cog
98. Q9	Remembering test instructions, ADAS-Cog
99. Q10	Comprehension of spoken and written language, ADAS-Cog
100. Q11	Word finding difficulty, ADAS-Cog
101. Q12	Language, ADAS-Cog
102. Q14	Number cancellation, ADAS-Cog
103. TOTAL11	Classic 70 points total, excluding Q4 and Q14, ADAS-Cog
104. TOTALMOD	85 points total, including Q4 and Q14, ADAS-Cog
105. GDSATIS	Are you basically satisfied with your life?, Partial Score of GDS
106. GDDROP	Have you dropped many of your activities and interests?, Partial Score of GDS
107. GDEEMPTY	Do you feel that your life is empty?, Partial Score of GDS
108. GDBORED	Do you often get bored?, Partial Score of GDS
109. GDSPIRIT	Are you in good spirits most of the time?, Partial Score of GDS
110. GDAFRAID	Are you afraid that something bad is going to happen to you?, Partial Score of GDS
111. GDHAPPY	Do you feel happy most of the time?, Partial Score of GDS
112. GDHELP	Do you often feel helpless?, Partial Score of GDS
113. GDHOME	Do you prefer to stay at home, rather than going out and doing new things?, Partial Score of GDS
114. GDMEMORY	Do you feel you have more problems with memory than most?, Partial Score of GDS
115. GDALIVE	Do you think its wonderful to be alive now?, Partial Score of GDS
116. GDWORTH	Do you feel pretty worthless the way you are now?, Partial Score of GDS
117. GDENERGY	Do you feel full of energy?, Partial Score of GDS
118. GDHOPE	Do you feel that your situation is hopeless?, Partial Score of GDS
119. GDBETTER	Do you think that most people are better off than you are?, Partial Score of GDS
120. GDTOTAL	Total Score
121. FAQFINAN	Writing checks, paying bills, or balancing checkbook, Partial Score, FAQ

TABLE 1: Continued.

Measure	Description
122. FAQFORM	Assembling tax records, business affairs, or other papers, Partial Score, FAQ
123. FAQSHOP	Shopping alone for clothes, household necessities, or groceries, Partial Score, FAQ
124. FAQGAME	Playing a game of skill such as bridge or chess, working on a hobby, Partial Score, FAQ
125. FAQBEVG	Heating water, making a cup of coffee, turning off the stove, Partial Score, FAQ
126. FAQMEAL	Preparing a balanced meal, Partial Score, FAQ
127. FAQEVENT	Keeping track of current events, Partial Score, FAQ
128. FAQTV	Paying attention to and understanding a TV program, book, or magazine, Partial Score, FAQ
129. FAQREM	Remembering appointments, family occasions, holidays, medications, Partial Score, FAQ
130. FAQTRAVL	Traveling out of the neighborhood, driving, or arranging to take public transportation, Partial Score, FAQ
131. FAQ total	Total Score, FAQ

language, executive functioning, memory, praxis, attention, working memory, visuospatial/constructional ability, functional abilities, and depression, as shown in Table 2.

2.5. The Machine Learning Classifier. In order to automatically classify subjects into different groups through the considered neuropsychological measures, we used an ML classifier based on methods previously published by our group [21].

2.5.1. The Classification Algorithm. The classification algorithm is based on Support Vector Machines (SVMs) [22], which generate a predictive model to perform binary group separation. The predictive model is designed as a hyperplane computed using a training set of data as input to SVM. This set consists of (1) a vector of samples belonging to two different classes and (2) the corresponding vector of labels, which identifies the class of each sample. During this training phase, SVM computes a hyperplane to separate the samples belonging to the two training classes. This hyperplane can then be used as a predictive model to classify a new sample into one or the other of the two classes. The predicted class y for sample x can be computed using the following formula:

$$y(x) = \sum_{n=1}^N w_n \cdot t_n \cdot k(x, x_n) + b, \quad (3)$$

where N is the number of samples included in the training set; w_n is a weight assigned by SVM to each sample n in the training set during the training phase; t_n is the label of the

TABLE 2: Continued.

Status/domains/subdomains	Neuropsychological tests	Reason for exclusion
	GDTOTAL (Total score, GDS)	
	GDHOPE (do you feel that your situation is hopeless?, Partial Score of GDS)	3
	GDDROP (have you dropped many of your activities and interests?, Partial Score of GDS)	3
	GDALIVE (do you think it is wonderful to be alive now?, Partial Score of GDS)	3
	GDHAPPY (do you feel happy most of the time?, Partial Score of GDS)	3
	GDWORTH (do you feel pretty worthless the way you are now?, Partial Score of GDS)	3
	GDENERGY (do you feel full of energy?, Partial Score of GDS)	3
	GDBETTER (do you think that most people are better off than you are?, Partial Score of GDS)	3
	GDSATIS (are you basically satisfied with your life?, Partial Score of GDS)	3
	GDEEMPTY (is life empty?, Partial Score of GDS)	3
	GDBORED (do you often get bored?, Partial Score of GDS)	3
	GDSPIRIT (are you in good spirits most of the time?, Partial Score of GDS)	3
	GDAFRMID (are you afraid that something bad is going to happen to you?, Partial Score of GDS)	3
	GDHELP (do you often feel helpless?, Partial Score of GDS)	3
	GDHOME (do you prefer to stay at home, rather than going out and doing new things?, Partial Score of GDS)	3
	GDMEMORY (do you feel you have more problems with memory than most?, Partial Score of GDS)	3
Depression		

sample n of the training set; $k(x, x_n)$ is a kernel function; and b is a threshold parameter.

We used the Matlab platform to implement the SVM classifier. Our code also included algorithms from the biolearning toolbox of Matlab. Classification was performed using both linear and nonlinear kernels for performance comparison, the latter including a quadratic kernel, a Gaussian Radial Basis Function (RBF) kernel with default sigma = 1, and a Multilayer Perceptron kernel with default scale [1 -1]. For each subject, the CDR score was used as a label for all classifiers.

2.5.2. Cut-Off on Features

(a) *Computation-Based Features.* Z-scored features were sorted in descending order according to their FDR. The 5% features with highest FDR were retained for classification.

(b) *Features Chosen by the Neuropsychologists.* Each of the features chosen by the neuropsychologists was used as input into the classifier, thus obtaining individual feature classification accuracy. Features were sorted in descending order according to their classification accuracy. The top 10 features with the highest accuracy were retained for classification.

2.5.3. *Optimization of Features and Evaluation of Performance.* An optimization of features was performed in order to find the combination of scores and subscores that return the best performance for the classification of the different groups of subjects.

For all kernels of the classifier, we performed a nested 10-fold Cross Validation (nested CV), which consists in (1) splitting the original dataset into 10 subsets of (possibly) equal size; (2) using 10-1 subsets to perform an inner training and validation loop for the optimization of the features; (3) using the 10th held-out subset to perform an outer test loop for the evaluation of the optimized features. In order to test all subsets, this procedure is then repeated 10 times.

Specifically, for each of the 10 rounds of the nested CV, all possible combinations of the reduced features (scores and subscores) were tested, for the two approaches described in Section 2.4. For each of the 10 rounds of the outer loop, accuracy, sensitivity, and specificity were calculated, and results were averaged across all rounds.

In order to avoid problems arising from the use of class-imbalanced datasets, which could lead to the classifier being trained *more* on one class than the other (in a binary-classification framework), the number of subjects in the two classes was kept balanced in both the training and validation sets.

We also evaluated the classification performance using two specific metrics for imbalanced-domain problems, namely, the GM of the true rates and the Dominance [23, 24]. These were computed as follows:

$$\text{GM} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \cdot \frac{\text{TN}}{\text{FP} + \text{TN}}}, \quad (4)$$

$$\text{Dominance} = \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{TN}}{\text{FP} + \text{TN}},$$

where TP (TN) is the number of true positives (negatives) and FP (FN) is the number of false positives (negatives).

The whole nested CV process was repeated for 100 iterations in order to reduce statistical variability of results. In fact, training and validation may depend on a particular random choice for the pair of training and validation sets, which could lead to a wrong classification performance estimate. The use of an iterative procedure helps to prevent this, because classification performances (as well as score frequencies) are averaged across 100 iterations.

A classification was performed for the following three comparisons: (1) CDR = 0.5 versus CDR = 0, (2) CDR = 1 versus CDR = 0.5, and (3) CDR = 1 versus CDR = 0.5.

2.5.4. *Features as Best Predictors.* The optimal combination of features was chosen as the one with the maximum accuracy of classification in the inner validation loop. Hence, we obtained an optimal combination of features for each round of 10 and for each iteration of 100, thus 1000 optimal combinations of features.

In order to determine which features were the most important for the classification, we computed the frequency of each feature in all optimal combinations. The features were sorted in descending order according to their frequency. The top 10 features with the highest frequency were shown as the best predictors.

3. Results

3.1. Feature Reduction

(a) *Computation-Based Features.* Figure 1 shows, as representative example for one round and one iteration (CDR = 1 versus CDR = 0, round #1, iteration #1), features ordered according to their FDR. The cut-off is shown reducing the number of features to the 5% with highest FDR, thus reducing the number from 131 to 7 features. Similar figures and results have been obtained for all the other rounds and iterations.

(b) *Features Chosen by the Neuropsychologists.* Table 2 shows the full list of features available from the ADNI database and those chosen by the neuropsychologists (in bold). Features are grouped into cognitive domains. The neuropsychologists reduced the number of features from 131 to 32. The reasons for exclusion are reported.

Figure 2 shows the features ordered according to their individual classification accuracy for CDR = 1 versus CDR = 0 (a), CDR = 0.5 versus CDR = 0 (b), and CDR = 1 versus CDR = 0.5 (c). Accuracy was computed as an average over 10 rounds and 100 iterations. The cut-off is shown reducing the number of features to the top 10 with highest classification accuracy, thus reducing the number from 32 to 10 features.

3.2. Optimization of Features and Evaluation of Performance

(a) *Computation-Based Features.* The classification performance averaged over all 10 rounds and 100 iterations is shown in Table 3 for CDR = 1 versus CDR = 0, Table 4 for CDR = 0.5 versus CDR = 0, and Table 5 for CDR = 1 versus CDR

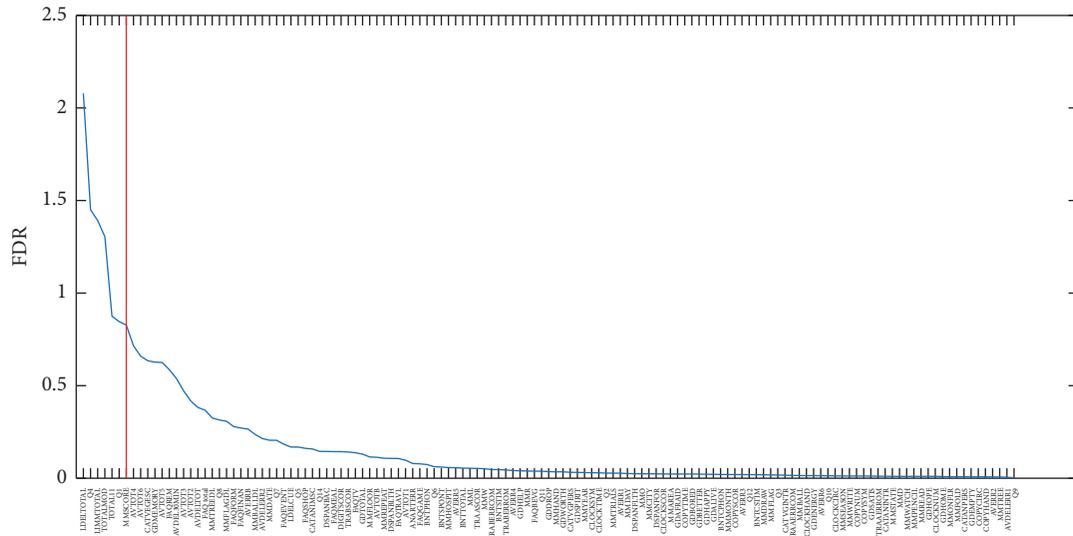


FIGURE 1: Representative example for one round and one iteration (CDR = 1 versus CDR = 0, round #1, iteration #1) of features ordered according to their FDR.

TABLE 3: Performance of ML (accuracy, sensitivity, specificity, GM, and Dominance) in the classification of CDR = 1 versus CDR = 0 using linear, quadratic, Gaussian RBF, and Multilayer Perceptron kernels. Results were obtained using the computation-based feature reduction.

Kernel	Accuracy [mean \pm std]*	Sensitivity [mean \pm std]*	Specificity [mean \pm std]*	Geometric Mean [mean \pm std]*	Dominance [mean \pm std]*
Linear	0.91 \pm 0.07	0.89 \pm 0.15	0.92 \pm 0.08	0.90 \pm 0.09	-0.03 \pm 0.17
Quadratic	0.91 \pm 0.07	0.87 \pm 0.16	0.92 \pm 0.08	0.89 \pm 0.10	-0.05 \pm 0.18
Gaussian RBF	0.92 \pm 0.07	0.90 \pm 0.14	0.93 \pm 0.08	0.91 \pm 0.09	-0.03 \pm 0.16
Multilayer Perceptron	0.91 \pm 0.07	0.87 \pm 0.16	0.93 \pm 0.08	0.90 \pm 0.10	-0.06 \pm 0.18

* Averaged across 10 rounds of the nested CV and across 100 iterations.

= 0.5. Accuracy, sensitivity, specificity, Geometric Mean, and Dominance are reported. Each table shows the classification results for linear, quadratic, Gaussian RBF, and Multilayer Perceptron kernels.

Table 6 shows the classification performance in the inner and outer loops of the nested CV for each of the 10 rounds individually. Results (in terms of accuracy of classification) were averaged over all 100 iterations and are shown for CDR = 1 versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5.

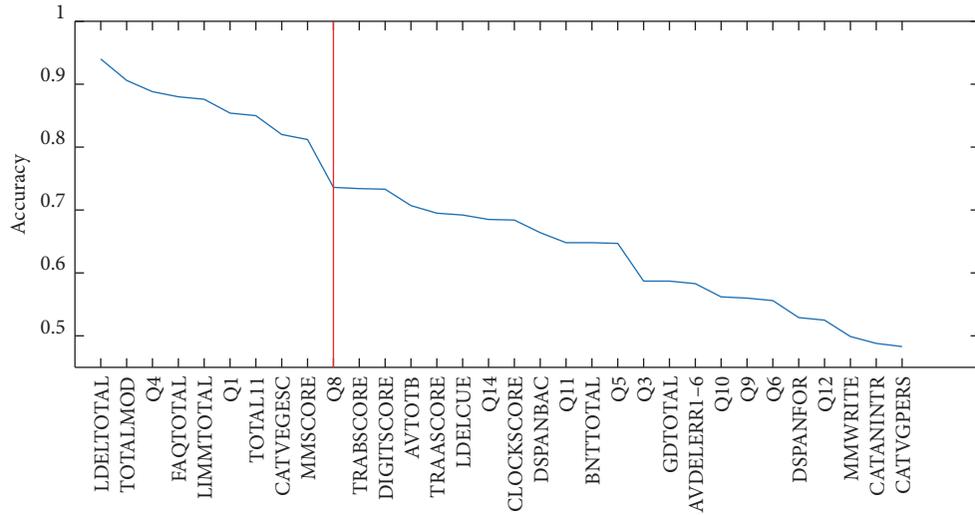
As expected, the best classification performance (accuracy, sensitivity, and specificity > 0.89 for the linear kernel) was obtained when discriminating subjects with moderate problems from normal subjects. However, a good performance (accuracy, sensitivity, and specificity > 0.85 for the linear kernel) was also obtained when discriminating subjects with mild impairment from normal subjects. This result is very important for patients, their families, and caregivers, since it suggests that the detection of changes may already be effective at an early stage of impairment. Thus, optimal support may be established and monitored when the cognitive abilities and independence of the subject have not already been compromised.

The most difficult discrimination was between subjects with mild and severe impairments (accuracy, sensitivity, and specificity ranging from 59% up to 67% for the linear kernel). This, however, has a minor impact on patients, their families, and caregivers, since, if early detection of behavioral changes is effective (see our consideration above), the subject may already be managed and monitored as an early dementia patient, and appropriate assistance and any necessary treatment should have already been started.

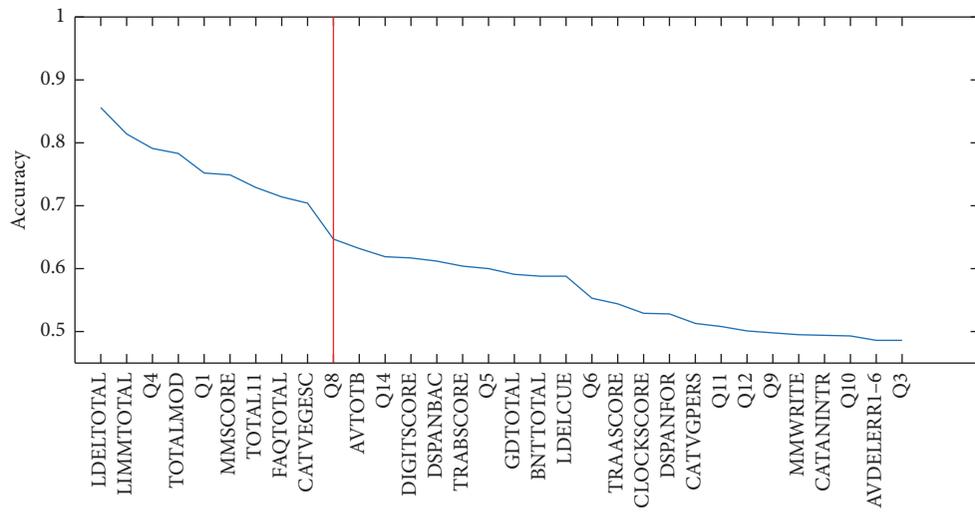
The classification results using nonlinear kernels (i.e., quadratic, Gaussian RBF, and Multilayer Perceptron kernels) are similar with those obtained using a linear kernel.

(b) *Features Chosen by Neuropsychologists.* Table 7 shows the classification performance in terms of accuracy, sensitivity, specificity, Geometric Mean, and Dominance averaged across all 10 rounds and all 100 iterations. Results are reported for CDR = 1 versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5.

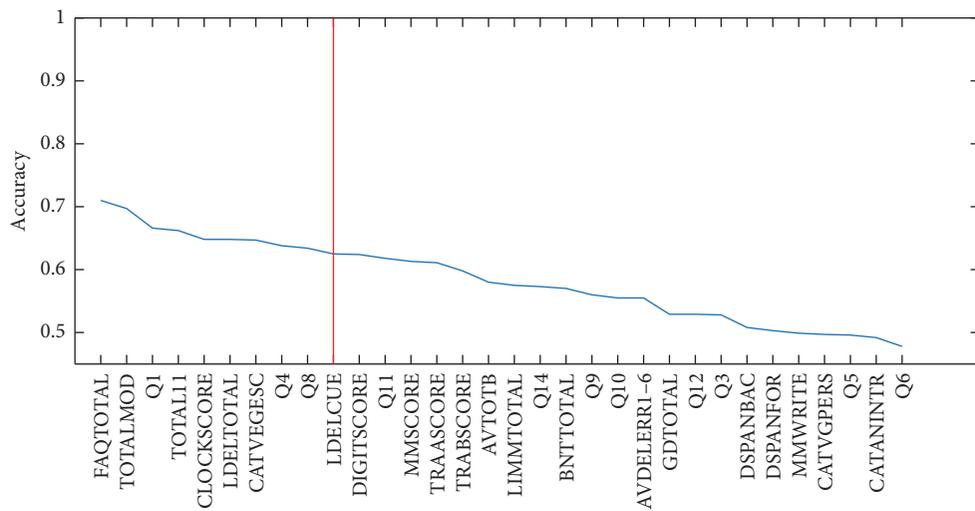
In Table 8, the classification performance in the inner and outer loops of the nested CV for each of the 10 rounds individually is shown. Accuracy of classification was obtained as average over all 100 iterations. The performance for



(a)



(b)



(c)

FIGURE 2: Features ordered according to their individual classification accuracy for CDR = 1 versus CDR = 0 (a), CDR = 0.5 versus CDR = 0 (b), and CDR = 1 versus CDR = 0.5 (c), using the feature reduction guided by the neuropsychologists. Results were obtained as average across all rounds (10) and iterations (100) but for each feature independently. Features are ranked in descending significance with respect to accuracy.

TABLE 4: Performance of ML (accuracy, sensitivity, specificity, GM, and Dominance) in the classification of CDR = 0.5 versus CDR = 0 using linear, quadratic, Gaussian RBF, and Multilayer Perceptron kernels. Results were obtained using the computation-based feature reduction.

Kernel	Accuracy [mean \pm std]*	Sensitivity [mean \pm std]*	Specificity [mean \pm std]*	Geometric Mean [mean \pm std]*	Dominance [mean \pm std]*
Linear	0.86 \pm 0.07	0.85 \pm 0.10	0.87 \pm 0.10	0.86 \pm 0.07	-0.01 \pm 0.15
Quadratic	0.86 \pm 0.07	0.85 \pm 0.11	0.88 \pm 0.09	0.86 \pm 0.07	-0.03 \pm 0.15
Gaussian RBF	0.86 \pm 0.07	0.85 \pm 0.10	0.87 \pm 0.10	0.86 \pm 0.07	-0.02 \pm 0.15
Multilayer Perceptron	0.85 \pm 0.07	0.83 \pm 0.12	0.87 \pm 0.10	0.85 \pm 0.07	-0.04 \pm 0.16

* Averaged across 10 rounds of the nested CV and across 100 iterations.

TABLE 5: Performance of ML (accuracy, sensitivity, specificity, GM, and Dominance) in the classification of CDR = 1 versus CDR = 0.5 using linear, quadratic, Gaussian RBF, and Multilayer Perceptron kernels. Results were obtained using the computation-based feature reduction.

Kernel	Accuracy [mean \pm std]*	Sensitivity [mean \pm std]*	Specificity [mean \pm std]*	Geometric Mean [mean \pm std]*	Dominance [mean \pm std]*
Linear	0.65 \pm 0.12	0.59 \pm 0.22	0.67 \pm 0.16	0.60 \pm 0.16	-0.09 \pm 0.30
Quadratic	0.65 \pm 0.12	0.59 \pm 0.23	0.67 \pm 0.16	0.60 \pm 0.16	-0.07 \pm 0.30
Gaussian RBF	0.64 \pm 0.12	0.61 \pm 0.23	0.65 \pm 0.16	0.60 \pm 0.16	-0.05 \pm 0.31
Multilayer Perceptron	0.63 \pm 0.12	0.62 \pm 0.24	0.63 \pm 0.17	0.60 \pm 0.15	0 \pm 0.33

* Averaged across 10 rounds of the nested CV and across 100 iterations.

TABLE 6: Performance of ML (accuracy) in the inner and outer loops for each of the 10 rounds (averaged across 100 iterations). Results are reported for CDR = 1 versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5 using a linear kernel and the computation-based feature reduction.

Round	CDR = 1 versus CDR = 0		CDR = 0.5 versus CDR = 0		CDR = 1 versus CDR = 0.5	
	Inner loop accuracy*	Outer loop accuracy*	Inner loop accuracy*	Outer loop accuracy*	Inner loop accuracy*	Outer loop accuracy*
1	0.95 \pm 0.02	0.91 \pm 0.06	0.87 \pm 0.03	0.87 \pm 0.07	0.75 \pm 0.03	0.64 \pm 0.14
2	0.95 \pm 0.02	0.90 \pm 0.07	0.87 \pm 0.03	0.85 \pm 0.07	0.75 \pm 0.03	0.65 \pm 0.12
3	0.95 \pm 0.02	0.91 \pm 0.07	0.87 \pm 0.02	0.85 \pm 0.07	0.75 \pm 0.03	0.65 \pm 0.12
4	0.95 \pm 0.02	0.92 \pm 0.07	0.87 \pm 0.03	0.87 \pm 0.07	0.75 \pm 0.03	0.65 \pm 0.10
5	0.95 \pm 0.02	0.92 \pm 0.06	0.87 \pm 0.03	0.85 \pm 0.07	0.76 \pm 0.04	0.62 \pm 0.13
6	0.95 \pm 0.02	0.91 \pm 0.07	0.87 \pm 0.03	0.87 \pm 0.06	0.75 \pm 0.03	0.66 \pm 0.13
7	0.95 \pm 0.02	0.91 \pm 0.07	0.87 \pm 0.02	0.86 \pm 0.06	0.75 \pm 0.03	0.66 \pm 0.10
8	0.95 \pm 0.02	0.91 \pm 0.07	0.86 \pm 0.03	0.86 \pm 0.06	0.75 \pm 0.03	0.64 \pm 0.11
9	0.95 \pm 0.03	0.90 \pm 0.08	0.87 \pm 0.03	0.85 \pm 0.07	0.75 \pm 0.03	0.65 \pm 0.11
10	0.95 \pm 0.03	0.92 \pm 0.07	0.87 \pm 0.03	0.86 \pm 0.08	0.76 \pm 0.03	0.64 \pm 0.12
Total mean**	0.95 \pm 0.01	0.91 \pm 0.07	0.87 \pm 0.01	0.86 \pm 0.07	0.75 \pm 0.01	0.65 \pm 0.12

* mean \pm std averaged across 100 iterations; ** mean \pm std averaged across 100 iterations and 10 rounds.

TABLE 7: Performance of ML (accuracy, sensitivity, specificity, GM, and Dominance) in the classification of CDR = 1 versus CDR = 0.5, CDR = 0.5 versus CDR = 0 and CDR = 1 versus CDR = 0.5. Results were obtained using the feature reduction guided by the neuropsychologists.

Level of impairment	Accuracy [mean \pm std]*	Sensitivity [mean \pm std]*	Specificity [mean \pm std]*	Geometric Mean [mean \pm std]*	Dominance [mean \pm std]*
CDR = 1 vs CDR = 0	0.96 \pm 0.04	0.95 \pm 0.10	0.97 \pm 0.05	0.96 \pm 0.06	-0.03 \pm 0.11
CDR = 0.5 vs CDR = 0	0.86 \pm 0.07	0.84 \pm 0.10	0.89 \pm 0.09	0.86 \pm 0.07	-0.05 \pm 0.13
CDR = 1 vs CDR = 0.5	0.69 \pm 0.10	0.67 \pm 0.21	0.70 \pm 0.13	0.67 \pm 0.13	-0.03 \pm 0.26

* Across 10 rounds of the nested CV and across 100 iterations.

TABLE 8: Performance of ML (accuracy) in the inner and outer loops for each of the 10 rounds (averaged across 100 iterations). Results are reported for CDR = 1 versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5 using a linear kernel and the feature reduction guided by the neuropsychologists.

Round	CDR = 1 versus CDR = 0		CDR = 0.5 versus CDR = 0		CDR = 1 versus CDR = 0.5	
	Inner loop accuracy*	Outer loop accuracy*	Inner loop accuracy*	Outer loop accuracy*	Inner loop accuracy*	Outer loop accuracy*
1	1 ± 0.01	0.96 ± 0.05	0.91 ± 0.02	0.87 ± 0.07	0.81 ± 0.03	0.70 ± 0.11
2	1 ± 0.01	0.96 ± 0.05	0.91 ± 0.02	0.85 ± 0.06	0.81 ± 0.03	0.70 ± 0.10
3	1 ± 0.01	0.96 ± 0.05	0.91 ± 0.02	0.86 ± 0.06	0.81 ± 0.03	0.69 ± 0.10
4	0.99 ± 0.01	0.97 ± 0.04	0.91 ± 0.02	0.85 ± 0.07	0.81 ± 0.03	0.69 ± 0.09
5	0.99 ± 0.01	0.97 ± 0.04	0.91 ± 0.02	0.85 ± 0.06	0.81 ± 0.04	0.69 ± 0.10
6	0.99 ± 0.01	0.96 ± 0.05	0.91 ± 0.02	0.86 ± 0.06	0.81 ± 0.03	0.70 ± 0.11
7	0.99 ± 0.01	0.97 ± 0.04	0.91 ± 0.02	0.86 ± 0.07	0.80 ± 0.04	0.70 ± 0.11
8	0.99 ± 0.01	0.96 ± 0.04	0.91 ± 0.02	0.87 ± 0.05	0.81 ± 0.04	0.69 ± 0.10
9	0.99 ± 0.01	0.97 ± 0.04	0.91 ± 0.02	0.86 ± 0.07	0.80 ± 0.04	0.71 ± 0.10
10	0.99 ± 0.01	0.97 ± 0.05	0.91 ± 0.02	0.86 ± 0.07	0.81 ± 0.04	0.67 ± 0.11
Total mean**	0.99 ± 0.01	0.96 ± 0.04	0.91 ± 0.00	0.86 ± 0.07	0.81 ± 0.01	0.69 ± 0.10

* mean ± std averaged across 100 iterations; ** mean ± std averaged across 100 iterations and 10 rounds.

CDR = 1 versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5 is reported.

As expected, also in this case, the best classification performance (accuracy, sensitivity, and specificity > 0.95) was obtained when discriminating subjects with moderate problems from normal subjects. However, a good performance (accuracy, sensitivity, and specificity > 0.84) was also obtained when discriminating subjects with mild impairment from normal subjects. The most difficult discrimination was distinguishing subjects with mild from moderate impairment (accuracy, sensitivity, and specificity ranging from 67% up to 70%).

The trend of the highest accuracy of classification as a function of the configuration (number of features per combination) is shown in Figure 3. The highest accuracy in the inner loop of the nested CV is shown for CDR = 1 versus CDR = 0 (blue), CDR = 0.5 versus CDR = 0 (red), and CDR = 1 versus CDR = 0.5 (green). Results were obtained using the features chosen by the neuropsychologists and averaged across all 10 rounds and all 100 iterations.

As it can be seen, the performance of the model slightly improves when more features are used as input into the classifier, until reaching a plateau at about 5 features. Although the performance of classification is different for each of the three comparisons (CDR = 1 versus CDR = 0, CDR = 0.5 versus CDR = 0, and CDR = 1 versus CDR = 0.5), the trend as a function of the number of features considered is similar.

3.3. Features More Frequently Found as Best Predictors

(a) *Computation-Based Features.* Table 9 shows the top 10 features most frequently found as best predictors across all 10 rounds and all 100 iterations (i.e., most frequently found in the combinations of features with the best classification accuracy) using the FDR feature reduction. Features are ranked by frequency in a descending order.

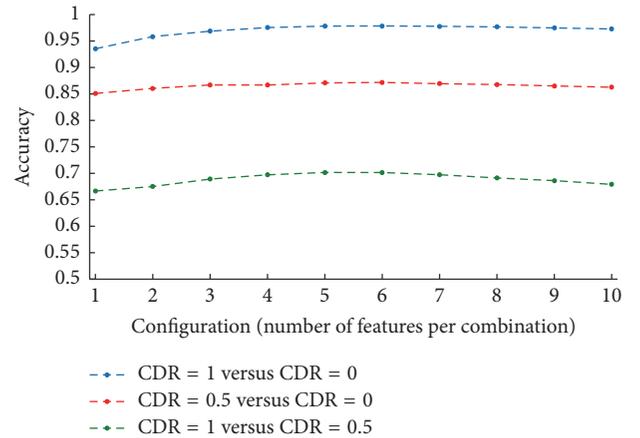


FIGURE 3: Highest accuracy of classification as a function of the configuration (combination of input features) in the inner loop of the nested CV. Performances are reported for CDR = 1 versus CDR = 0 (blue), CDR = 0.5 versus CDR = 0 (red), and CDR = 1 versus CDR = 0.5 (green), using the features chosen by the neuropsychologists. Results were obtained as average across all rounds (10) and iterations (100).

(b) *Features Chosen by the Neuropsychologists.* Table 10 shows the top 10 features most frequently found as best predictors among all 10 rounds and all 100 iterations (i.e., most frequently found in the combinations of features with the best classification accuracy) using the features chosen by the neuropsychologists. Features are ranked by frequency in a descending order.

The two approaches considered for the reduction of features achieved similar results.

In CDR = 1 versus CDR = 0, both approaches found the following features among best predictors: LDELTOTAL, TOTALMOD, LIMMTOTAL, FAQTOTAL, Q4, Q1, and TOTALII. This is not unexpected, since previous studies

TABLE 9: Top 10 features most frequently found as best predictors across all 10 rounds and all 100 iterations using the FDR feature reduction.

Level of impairment	Features	Frequency*
CDR = 1 versus CDR = 0	(1) LDELTOTAL (LM)	71%
	(2) TOTALMOD (ADAS)	10%
	(3) LIMMTOTAL (LM)	4%
	(4) FAQTOTAL (FAQ)	4%
	(5) Q4 (ADAS)	4%
	(6) AVTOT5 (AVLT)	3%
	(7) AVTOT4 (AVLT)	1%
	(8) Q1 (ADAS)	0.8%
	(9) AVDEL30MIN (AVLT)	0.6%
	(10) TOTALI1 (ADAS)	0.5%
CDR = 0.5 versus CDR = 0	(1) LDELTOTAL (LM)	91%
	(2) Q4 (ADAS-Cog)	22%
	(3) LIMMTOTAL (LM)	15%
	(4) TOTALMOD (ADAS-Cog)	12%
	(5) GDHOPE (GDS)	6%
	(6) MMD (MMSE)	2%
	(7) MMSCORE (MMSE)	0.3%
	(8) AVTOT4 (AVLT)	0.1%
	(9) CATVEGESC (Semantic Fluency Test)	0.1%
	(10) TOTALI1 (ADAS)	0.1%
CDR = 1 versus CDR = 0.5	(1) FAQTOTAL (FAQ)	31%
	(2) TOTALMOD (ADAS-Cog)	22%
	(3) AVTOT5 (AVLT)	10%
	(4) FAQFORM (FAQ)	6%
	(5) Q1 (ADAS-Cog)	6%
	(6) FAQREM (FAQ)	6%
	(7) TOTALI1 (ADAS)	5%
	(8) CLOCKSCOR (CLOCK Test)	4%
	(9) CATVEGESC (Semantic Fluency Test)	4%
	(10) Q8 (ADAS)	4%

* Across 10 rounds of the nested CV and across 100 iterations.

have demonstrated that long-term memory reflects the early pathological involvement of the mediotemporal lobe in the course of early AD [25], while functional abilities can occur as a result of cognitive impairment [26].

AVTOT5, AVTOT4, and AVDEL30MIN were found among best predictors when using the FDR-based feature reduction. However, these measures were excluded by the neuropsychologists because of their overlapping with Q1 (Word Recall Task, ADAS) and Q4 (Delayed Word Recall, ADAS) of ADAS.

Q8 and MMSCORE were found among best predictors when using the feature reduction guided by the neuropsychologists. However, these measures were not among the 5% with the highest FDR and thus was excluded in the computation-based classification. MMSCORE has already

TABLE 10: Top 10 features most frequently found as best predictors across all 10 rounds and all 100 iterations using the features chosen by the neuropsychologists.

Level of impairment	Features	Frequency*
CDR = 1 versus CDR = 0	(1) LDELTOTAL (Logical Memory Test)	80%
	(2) TOTALMOD (ADAS)	50%
	(3) FAQ total (FAQ)	29%
	(4) TOTALI1 (ADAS)	18%
	(5) CATVEGESC (Semantic Fluency Test)	13%
	(6) Q4 (ADAS)	13%
	(7) LIMMTOTAL (Logical Memory)	9%
	(8) Q8 (ADAS)	5%
	(9) MMSCORE (MMSE)	5%
	(10) Q1 (ADAS)	3%
CDR = 0.5 versus CDR = 0	(1) FAQ total (FAQ)	81%
	(2) LDELTOTAL (Logical Memory Test)	77%
	(3) Q4 (ADAS)	44%
	(4) TOTALMOD (ADAS)	39%
	(5) CATVEGESC (Semantic Fluency Test)	36%
	(6) LIMMTOTAL (Logical Memory)	30%
	(7) MMSCORE (MMSE)	30%
	(8) Q8 (ADAS)	23%
	(9) TOTALI1 (ADAS)	19%
	(10) Q1 (ADAS)	19%
CDR = 1 versus CDR = 0.5	(1) FAQ total (FAQ)	82%
	(2) CLOCKSCOR (CLOCK Test)	36%
	(3) Q8 (ADAS)	35%
	(4) LDEL30MIN (Logical Memory Test)	33%
	(5) TOTALI1 (ADAS)	30%
	(6) Q4 (ADAS)	29%
	(7) TOTALMOD (ADAS)	22%
	(8) LDELTOTAL (Logical Memory Test)	20%
	(9) CATVEGESC (Semantic Fluency Test)	19%
	(10) Q1 (ADAS)	18%

* Across 10 rounds of the nested CV and across 100 iterations.

been found to discriminate early AD from CN with good accuracy (about 70%) [27], while Q8 has been listed among the useful memory and learning tests for AD detection.

In CDR = 0.5 versus CDR = 0, both approaches found the following features as best predictors: LDELTOTAL, Q4, LIMMTOTAL, TOTALMOD, MMSCORE, CATVEGESC, and TOTALI1. Many studies on MCI patients have found an impairment in long-term memory function [28, 29]. The

measures that best predict the conversion to AD range from immediate and delayed words recall [30] to verbal fluency. The number of different vegetable names produced is a fluency task that provides measures of language and executive processes, including self-initiated activity, categorization, and mental flexibility that are often involved in MCI patients (e.g., [31, 32]). GDHOPE, MMD, and AVTOT4 were found among the best predictors when using the computation-based feature reduction, but not when using the feature reduction guided by the neuropsychologists, because they were excluded from the list of input features by these neuropsychologists. These measures were excluded because AVTOT4 overlaps with Q1 (Word Recall Task, ADAS), while the other two measures were considered by the neuropsychologists not highly significant in the MCI literature and were not used to further our understanding of cognitive impairment. FAQTOTAL, Q8, and Q1 were found among the best predictors when the feature reduction was based on neuropsychological expertise but not when using the computation-feature reduction, because they were not among the 5% of features with the highest FDR retained for classification. This result is in line with the current literature, since Q1 and Q8 are both measures of long-term memory, and FAQ total has been recently reported as an important feature of MCI [33].

In CDR = 1 versus CDR = 0.5, both approaches found the following features among the best predictors: FAQTOTAL, TOTALMOD, Q1, TOTAL11, CLOCKSCOR, CATVEGESC, and Q8. Most of these features provide information on episodic memory (Q1 and Q8), global cognitive status (the global scores of ADAS and CLOCKSCOR), and language/executive functions (CATVEGESC). Given the primacy of memory in age-related cognitive impairment [34], ML algorithms highlighted measures of episodic memory (recalling information learned previously) as well as recognition memory (a nonverbal task that requires the recognition of the words learned in a longer list of words presented with distracter words). AVTOT5, FAQFORM, and FAQREM were found among best predictors when using the computation-based feature reduction, but they were excluded from the list of input features by the two neuropsychologists. The reason for the exclusion was related to the overlap of AVTOT5 with Q1 (Word Recall Task, ADAS), while the other two measures were considered by the neuropsychologists not highly significant in the MCI/AD literature. LDELCUE, Q4, and LDELTOTAL were found among best predictors when the neuropsychologists guided the feature reduction but not when using the automatic feature reduction because not among the 5% features with the highest FDR retained for the classification. Consistently with the literature, many studies [34–36] have shown anterograde episodic memory as the best marker to predict the conversion to AD in subjects with MCI.

4. Discussion

The ability of neuropsychological measures to help in discriminating between different degrees of cognitive impairment has been widely recognized (e.g., [37, 38]). Despite the existence of clinical measures that are used to classify

patients and diagnose clinical disorders showing cognitive impairment, as AD, a large amount of subjectivity affects the diagnostic process. Machine learning is emerging in clinical neuropsychology and neurology as a credible technique to support this process in a quantitative way.

Our work explored the contribution of cognitive, behavioral, and functional measures in the automatic classification of different stages of cognitive impairment using ML. Both total and partial scores as measures of cognitive domains and subdomains from a group of different neuropsychological tests available from the ADNI database were considered as potential predictors of cognitive impairment, even at an early stage.

Our results showed that, among 131 measures considered, it is possible to use a subset of measures in which the classification accuracy is higher than 90% for severe (CDR = 1) versus no impairment (CDR = 0) and higher than 85% for mild (CDR = 0.5) versus no impairment (CDR = 0). Our performances for these two comparisons are similar to those obtained by Logie et al. [6]. These findings could have an impact on the clinical process to perform early diagnosis, with different benefits for patients. However, although we did not include CDR = 2 in the group of patients with severe impairment as Weakley et al. did, accuracy of our classification for patients with CDR = 0.5 versus CDR = 1 was found limited (65–69%) suggesting that our work is in progress and needs improvements. The automatic diagnosis of AD is not challenging and we would expect that ML, by using many features, would be able to predict mild impairment better than subjective techniques. Our model should be modified for this specific comparison in order to achieve best accuracy.

From a methodological point of view, we used a lower number of features than the smallest number of subjects in each group (55 subjects with CDR = 1). This warrants that no curse-of-dimensionality problems occurred. In addition, the feature reduction strategies adopted in our study reduced the number of features used as inputs for the classifier to (at most) 10 (using the feature reduction guided by the neuropsychologists). This number is much lower than the number of subjects in the smallest group, thus avoiding overfitting problems, especially considering the use of SVMs, which are *designed* to handle high-dimensional data.

Concerning the final cognitive profile, when comparing the fully automated classification to the classification guided by the neuropsychologists, a good overlap of results between the two classifications was found. Some tests were found more frequently among best predictors for the automatic classification, namely, LM, ADAS-Cog, AVLT, and FAQ, with a major role of ADAS-Cog measures of delayed and immediate memory and the FAQ measure of financial competency.

There are some measures known to be highly implicated in AD that this model fails to identify. There are some tests that are specifically recommended, that is, the Free and Cued Selective Reminding Test (FCSRT), which is considered as a valid clinical marker for AD [39]. This test was not included in our model since it was not available in the ADNI dataset. These tests provide specific episodic memory measures that correlate with the hippocampal dysfunction, suggesting that

a poor performance is typically registered in AD patients [40].

In the field of neurodegenerative diseases, several other studies have applied ML for the identification of subsets of “optimal” classification predictors (e.g., [41, 42]), resulting in a subset of optimal measures, in particular measures of decline in episodic memory. A similar approach has been used for the evaluation of linguistic features within a language test for patients with AD [43–46]. They show the potential of ML in determining the best predictors for language impairment in these patients. Some other studies have investigated this contribution in the classification of AD.

Promising results on similar approaches have been also published in the field of psychiatric disorders. An interesting study by Costafreda et al. [47] showed that pattern recognition algorithms coupled with verbal fluency have a reliable diagnostic power in differentiating schizophrenic patients from healthy controls and patients affected by bipolar disorders. Pina-Camacho et al. [48] used SVM to study predictors of schizophrenia in early-onset first episodes of psychosis. They found that, among the variables included, neuropsychological measures (impaired attention, motor coordination, and global cognition) showed the highest predictive value for a diagnostic outcome of the disease. A study on autism spectrum disorder [49] showed that cognitive measures could be a useful aid to the diagnostic process when assessed by an SVM classifier. However, limitations of these studies were related to the use of cognitive, behavioral, and functional measures during the diagnostic classification of subjects used as a gold standard of analysis, thus inducing overfitting (e.g., [50–52]). In our study, gold standard measures (CDR in our case) were independent from measures used as features. This excluded any potential bias due to circularity in the classification process.

Overall, the results of our work show that ML provides an effective technique to quantify the process to classify patients and to diagnose clinical disorders by using neuropsychological measures. Combined with multicenter databases that collect information from a substantial number of patients (as the ADNI database used in our work), ML was proven to be suitable to conduct statistically meaningful machine learning calculations and to assist clinicians in the optimization of the best measures (and submeasures) to be used, reducing at the same time the subjectivity of the process. For example, we have shown how ML can support the selection of important subscores for the classification of subjects. This could be useful in the optimization of current neuropsychological tests or in the design of new tests for next-generation cognitive assessment.

Notably, cognitive measures found in our works cannot be considered sufficient to perform a neuropsychological classification of patients and we cannot conclude that they are the best predictors from a cognitive perspective, for many reasons. For computation reasons, we have adopted two different strategies to reduce the number of features as input to classification; both strategies have limited the measures used by the machine learning to select the optimal predictors. Furthermore, other measures not included in the considered studies could be behavioral data providing

a thorough description of the disease (e.g., loss of empathy, disinhibition, and apathy). These measures could improve the model. In addition, our findings need to be verified in other independent databases and by using other machine learning algorithms, thus proving the generality of our results.

In conclusion, ML approaches can be a useful tool for clinical neuropsychologists especially when they need to deal with a huge amount of data and the level of subjectivity of the process must be reduced. Given that a higher level of accuracy in classification still needs to be achieved and that some questions remain unanswered, further validations and verifications are needed in future research. However, we believe that this study represents a step towards achieving the goal of automatic classification of AD patients by means of clinical neuropsychological assessment.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Authors' Contributions

Petronilla Battista and Christian Salvatore contributed equally to the paper.

Acknowledgments

This work was supported by the CNR Research Project “Aging: Molecular and Technological Innovations for Improving the Health of the Elderly” no. DSB.AD009.001, Activity no. DSB.AD009.001.043. Data collection and sharing were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense, Award no. W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research provide funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<http://www.fnih.org/>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the

Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California. The Alzheimer's Disease Neuroimaging Initiative data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu/>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

- [1] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: a systematic review and metaanalysis," *Alzheimer's & Dementia*, vol. 9, no. 1, pp. 63–75, 2013.
- [2] H. W. Querfurth and F. M. LaFerla, "Alzheimer's disease," *New England Journal of Medicine*, vol. 362, no. 4, pp. 329–344, 2010.
- [3] *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, American Psychiatric Association, 2013.
- [4] M. D. Lezak, D. B. Howieson, D. W. Loring, H. J. Hannay, and J. S. Fischer, *Neuropsychological Assessment*, 4th edition, 2004.
- [5] H. C. Fichman, R. M. Oliveira, and C. S. Fernandes, "Neuropsychological and neurobiological markers of the preclinical stage of alzheimer's disease," *Psychology and Neuroscience*, vol. 4, no. 2, pp. 245–253, 2011.
- [6] R. H. Logie, M. A. Parra, and S. Della Sala, "From cognitive science to dementia assessment," *Policy Insights from the Behavioral and Brain Sciences*, vol. 2, no. 1, pp. 81–91, 2015.
- [7] G. Gainotti, D. Quaranta, M. G. Vita, and C. Marra, "Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 38, no. 3, pp. 481–495, 2014.
- [8] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.
- [9] C. Salvatore, P. Battista, and I. Castiglioni, "Frontiers for the early diagnosis of AD by means of MRI brain imaging and support vector machines," *Current Alzheimer Research*, vol. 13, no. 5, pp. 509–533, 2016.
- [10] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach," *Frontiers in Neuroscience*, vol. 9, article 307, 2015.
- [11] A. Weakley, J. A. Williams, M. Schmitter-Edgecombe, and D. J. Cook, "Neuropsychological test selection for cognitive impairment classification: a machine learning approach," *Journal of Clinical and Experimental Neuropsychology*, vol. 37, no. 9, pp. 899–916, 2015.
- [12] W. J. Burke, J. P. Miller, E. H. Rubin et al., "Reliability of the Washington University Clinical Dementia Rating," *Archives of Neurology*, vol. 45, no. 1, pp. 31–32, 1988.
- [13] L. Breiman, "Statistical modeling: the two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.
- [14] W. R. Shankle, S. Mania, M. B. Dick, and M. J. Pazzani, "Simple models for estimating dementia severity using machine learning," *Studies in Health Technology and Informatics*, vol. 52, part 1, pp. 472–476, 1998.
- [15] M. F. Folstein, L. N. Robins, and J. E. Helzer, "The mini-mental state examination," *Archives of General Psychiatry*, vol. 40, no. 7, p. 812, 1983.
- [16] D. Wechsler, *Manual for Wechsler Memory Scale—Revised*, 1987.
- [17] A. Rey, *The Clinical Psychological Examination*, Presses Universitaires de France, Paris, France, 1964.
- [18] R. M. Reitan, "Trail making test results for normal and brain-damaged children," *Perceptual and Motor Skills*, vol. 33, no. 2, pp. 575–581, 1971.
- [19] E. F. Kaplan, H. Goodglass, and S. Weintraub, *The Boston Naming Test*, Lea Febiger, Philadelphia, Pa, USA, 2nd edition, 1983.
- [20] J. A. Yesavage, T. L. Brink, T. L. Rose et al., "Development and validation of a geriatric depression screening scale: a preliminary report," *Journal of Psychiatric Research*, vol. 17, no. 1, pp. 37–49, 1982.
- [21] C. Salvatore, A. Cerasa, I. Castiglioni et al., "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy," *Journal of Neuroscience Methods*, vol. 222, pp. 230–237, 2014.
- [22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [23] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.
- [24] V. García, R. A. Mollineda, and J. S. Sánchez, "A new performance evaluation method for two-class imbalanced problems," in *Structural, Syntactic, and Statistical Pattern Recognition*, vol. 5342 of *Lecture Notes in Computer Science*, pp. 917–925, Springer, Berlin, Germany, 2008.
- [25] G. A. Carlesimo, R. Perri, and C. Caltagirone, "Category cued recall following controlled encoding as a neuropsychological tool in the diagnosis of Alzheimer's disease: a review of the evidence," *Neuropsychology Review*, vol. 21, no. 1, pp. 54–65, 2011.
- [26] V. G. Wadley, O. Okonkwo, M. Crowe, and L. A. Ross-Meadows, "Mild cognitive impairment and everyday function: evidence of reduced speed in performing instrumental activities of daily living," *American Journal of Geriatric Psychiatry*, vol. 16, no. 5, pp. 416–424, 2008.
- [27] S. Freitas, M. R. Simões, L. Alves, and I. Santana, "Montreal cognitive assessment (MoCA): normative study for the portuguese population," *Journal of Clinical and Experimental Neuropsychology*, vol. 33, no. 9, pp. 989–996, 2011.
- [28] J. A. Small, S. Kemper, and K. Lyons, "Sentence repetition and processing resources in Alzheimer's disease," *Brain and Language*, vol. 75, no. 2, pp. 232–258, 2000.
- [29] K. L. Lange, M. W. Bondi, D. P. Salmon et al., "Decline in verbal memory during preclinical Alzheimer's disease: examination of the effect of APOE genotype," *Journal of the International Neuropsychological Society*, vol. 8, no. 7, pp. 943–955, 2002.
- [30] M. Wagner, S. Wolf, F. M. Reischies et al., "Biomarker validation of a cued recall memory deficit in prodromal Alzheimer disease," *Neurology*, vol. 78, no. 6, pp. 379–386, 2012.
- [31] J. Brandt, E. Aretouli, E. Neijstrom et al., "Selectivity of executive function deficits in mild cognitive impairment," *Neuropsychology*, vol. 23, no. 5, pp. 607–618, 2009.

- [32] K. E. Nutter-Upham, A. J. Saykin, L. A. Rabin et al., "Verbal fluency performance in amnesic MCI and older adults with cognitive complaints," *Archives of Clinical Neuropsychology*, vol. 23, no. 3, pp. 229–241, 2008.
- [33] S. T. Farias, D. Mungas, B. R. Reed, D. Harvey, D. Cahn-Weiner, and C. DeCarli, "MCI is associated with deficits in everyday functioning," *Alzheimer Disease and Associated Disorders*, vol. 20, no. 4, pp. 217–223, 2006.
- [34] M. W. Bondi, A. J. Jak, L. Delano-Wood, M. W. Jacobson, D. C. Delis, and D. P. Salmon, "Neuropsychological contributions to the early identification of Alzheimer's disease," *Neuropsychology Review*, vol. 18, no. 1, pp. 73–90, 2008.
- [35] M. J. Baek, H. J. Kim, H. J. Ryu et al., "The usefulness of the story recall test in patients with mild cognitive impairment and Alzheimer's disease," *Aging, Neuropsychology, and Cognition*, vol. 18, no. 2, pp. 214–229, 2011.
- [36] M. S. Albert, M. B. Moss, R. Tanzi, and K. Jones, "Preclinical prediction of AD using neuropsychological tests," *Journal of the International Neuropsychological Society*, vol. 7, no. 5, pp. 631–639, 2001.
- [37] H. Christensen, D. Hadzi-Pavlovic, and P. Jacomb, "The psychometric differentiation of dementia from normal aging: a meta-analysis," *Psychological Assessment*, vol. 3, no. 2, pp. 147–155, 1991.
- [38] É. M. L. Pimentel, "Role of neuropsychological assessment in the differential diagnosis of Alzheimer's disease and vascular dementia," *Dementia e Neuropsychologia*, vol. 3, no. 3, pp. 214–221, 2009.
- [39] B. Dubois, M. D. Feldman, C. Jacova et al., "Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria," *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014.
- [40] M. Sarazin, V. Chauviré, E. Gerardin et al., "The amnesic syndrome of hippocampal type in Alzheimer's disease: an MRI study," *Journal of Alzheimer's Disease*, vol. 22, no. 1, pp. 285–294, 2010.
- [41] Y. Cui, B. Liu, S. Luo et al., "Identification of conversion from mild cognitive impairment to alzheimer's disease using multivariate predictors," *PLoS ONE*, vol. 6, no. 7, Article ID e21896, 2011.
- [42] Y. Cui, W. Wen, D. M. Lipnicki et al., "Automated detection of amnesic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach," *NeuroImage*, vol. 59, no. 2, pp. 1209–1217, 2012.
- [43] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2015.
- [44] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. De Jager, and P. Garrard, "Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology," *Journal of Alzheimer's Disease*, vol. 42, pp. S3–S17, 2014.
- [45] C. Reverberi, P. Cherubini, S. Baldinelli, and S. Luzzi, "Semantic fluency: cognitive basis and diagnostic performance in focal dementias and Alzheimer's disease," *Cortex*, vol. 54, no. 1, pp. 150–164, 2014.
- [46] A. König, A. Satt, A. Sorin et al., "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.
- [47] S. G. Costafreda, C. H. Y. Fu, M. Picchioni et al., "Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder," *BMC Psychiatry*, vol. 11, article 18, 2011.
- [48] L. Pina-Camacho, J. Garcia-Prieto, M. Parellada et al., "Predictors of schizophrenia spectrum disorders in early-onset first episodes of psychosis: a support vector machine model," *European Child and Adolescent Psychiatry*, vol. 24, no. 4, pp. 427–440, 2015.
- [49] C. E. Wilson, F. Happé, S. J. Wheelwright et al., "The neuropsychology of male adults with high-functioning autism or asperger syndrome," *Autism Research*, vol. 7, no. 5, pp. 568–581, 2014.
- [50] J. Dukart, F. Sambataro, and A. Bertolino, "Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers," *Journal of Alzheimer's Disease*, vol. 49, no. 4, pp. 1143–1159, 2015.
- [51] J. Koikkalainen, H. Pölönen, J. Mattila, M. van Gils, H. Soininen, and J. Lötjönen, "Improved classification of alzheimer's disease data via removal of nuisance variability," *PLoS ONE*, vol. 7, no. 2, Article ID e31112, 2012.
- [52] Y. Cui, B. Liu, S. Luo et al., "Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors," *PLoS ONE*, vol. 6, no. 7, Article ID e21896, 2011.