Journal of Alzheimer's Disease 79 (2021) 845–861
DOI 10.3233/JAD-200955
IOS Press

Analysis of Risk Factors in Dementia Through Machine Learning

Francisco Javier Balea-Fernandez[a,*], Beatriz Martinez-Vega[b], Samuel Ortega[b], Himar Fabelo[b], Raquel Leon[b], Gustavo M. Callico[b] and Cristina Bibao-Sieyro[c]
[a]*Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain*
[b]*Research Institute for Applied Microelectronics, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain*
[c]*Hospital Universitario de Gran Canaria Dr. Negrín, Las Palmas de Gran Canaria, Spain*

Accepted 6 November 2020
Pre-press 23 December 2020

Abstract.
Background: Sociodemographic data indicate the progressive increase in life expectancy and the prevalence of Alzheimer's disease (AD). AD is raised as one of the greatest public health problems. Its etiology is twofold: on the one hand, non-modifiable factors and on the other, modifiable.
Objective: This study aims to develop a processing framework based on machine learning (ML) and optimization algorithms to study sociodemographic, clinical, and analytical variables, selecting the best combination among them for an accurate discrimination between controls and subjects with major neurocognitive disorder (MNCD).
Methods: This research is based on an observational-analytical design. Two research groups were established: MNCD group ($n = 46$) and control group ($n = 38$). ML and optimization algorithms were employed to automatically diagnose MNCD.
Results: Twelve out of 37 variables were identified in the validation set as the most relevant for MNCD diagnosis. Sensitivity of 100% and specificity of 71% were achieved using a Random Forest classifier.
Conclusion: ML is a potential tool for automatic prediction of MNCD which can be applied to relatively small preclinical and clinical data sets. These results can be interpreted to support the influence of the environment on the development of AD.

Keywords: Alzheimer's disease, machine learning, neurocognitive disorders, risk factors

INTRODUCTION

Sociodemographic data indicate the progressive increase in life expectancy, and Spaniards have the highest among European Union (EU) countries [1]. Spain places it as an average of 83.4 years, while the average in the EU is 80.9 years [1]. In 2018, there were 101.1 million people aged 65 years or more living in the EU and this number is expected to increase until 149.2 million in 2050 [2]. Among the causes of mortality, an increase in the mortality rate due to dementia stands out, while other diseases, such cardiovascular disease, chronic obstructive pulmonary disease, or diabetes, decrease [1]. These data do not only have population significance, but also have economic and health consequences since 35% of men and 65% of women over 65 years old have some type of disability, which represents 52% of the health care costs in specialized care for people over 65, and 40% of the total health expenditure in those over 70. Therefore, both in the present and in the future, it is a priority to effectively and efficiently address the expected changes in both the healthcare system (treatment of multiple diseases and chronicity) and in the social system (social structure) to respond to the current growing demands [3, 4].

*Correspondence to: Francisco Javier Balea-Fernández, PhD, Universidad de Las Palmas de Gran Canaria, Calle Sta. Juana de Arco, 1, 35004 Las Palmas de Gran Canaria, Spain.: E-mail: fbalea@cop.es.

A study carried out by Niu et al. showed that the prevalence of AD in Europe was 5.05% (95% CI, 4.73–5.39) [5], with Spain ranked third in the world with the highest prevalence of dementia, equivalent to 6.3% of Spaniards over 60 years of age [6, 7]. In addition, the data reveals an increasing trend in the number of patients with dementia, predicting an increase of 87% in the European region in the period 2010–2050 [5, 7, 8].

Prevalence of dementias according to data published by the World Health Organization (WHO) show that there are 50 million people in the world who suffer from dementia. Approximately 10 million new cases are registered each year. AD is the most common form of dementia, accounting for 60–70% of all dementias [9].

In recent decades, there has been an evolution in relation to AD diagnostic criteria. Since 1984, the most widely used diagnostic criteria were those of the NINCDS-ADRDA group [10]. In 2007, the NINCDS-ADRDA criteria established a diagnosis based on clinical criteria supported by complementary tests [11]. In 2011, the International Working Group for new research criteria for the diagnosis of AD (IWG-1) and the National Institute of Aging and the Alzheimer's Association (NIA-AA) modified the previous criteria, specifying the probability or possibility in the diagnosis of AD [12]. In 2014, both the IWG-2 and the NIA-AA specified in the diagnostic criteria the presence of an appropriate clinical phenotype (typical or atypical) and a biomarker [13]. Later, in 2018, the NIA-AA criteria defined AD in a strictly biological way using the biomarker profile (ATN) [14]. The criteria for establishing the diagnosis of AD inexorably evolve toward biomarkers in both blood and cerebrospinal fluid. The interest in clinical practice showed greater interest in the diagnosis using biomarkers in blood plasma for better accessibility and less invasiveness than in CSF. On the other hand, the decrease in the cost and availability of diagnoses using biomarkers (CSF, plasma, and image) is urgent because not all the countries or all the hospitals have the necessary material and can assume the high cost in the diagnostic process [14–20].

Risk factors

The unmodifiable and established risk factors for AD are age, sex, and genetic factors, with age being the main one. On the other hand, modifiable risk factors related to healthy lifestyles or educational levels have been identified and their optimization can lead to a significant decrease in the incidence of dementia [5, 20]. Thus, for example, in the Rotterdam study, the results indicate that about 1/4 to 1/3 of dementia cases could potentially be avoided through optimal prevention or treatment of cardiovascular risk factors and improvement of educational level [21].

Prevention data indicate that, in Europe, around 3 million cases of AD could be caused due to potentially modifiable risk factors. Factors such as diabetes, hypertension of adulthood, physical inactivity, depression, smoking, and low educational level have been investigated. Avoiding the risk factors, it has been estimated that 69.5% of the total dementia cases (including data on AD and vascular dementia) are potentially preventable [22].

Sociodemographic risk factors

Depression: In a meta-analysis carried out by Santabárbara et al., depression was statistically associated with AD, with a risk of 54% and a combined RR (Relative Risk) of 1.54 (95% CI: 1.05–2.24). The study concludes that 10.8% of AD cases could be attributed to depression. Currently, depression affects more than 300 million people in the world (with a higher prevalence in women than men), and, in addition to being considered an important risk factor for AD, it generates a high cost both economically and socially [23, 24].

Cognitive reserve: It is considered that an adequate cognitive reserve (analyzed from variables such as education, job occupation, and both cognitive and social activities) is an important protection factor in AD. It is associated with a dementia risk reduction between 23–39% [25]. In a meta-analysis carried out by Meng et al., they concluded that the risk of developing dementia was 2.61-fold higher (95% CI: 2.21–3.07) in those who had a lower educational level in relation to those with a high educational level. On the other hand, when the diagnosis of AD was produced, it was reported in 70% of the analyzed studies, that higher education leads to a faster cognitive decline in relation to the level of basic studies or without studies [26].

Lifestyle:

- Tobacco: Cataldo et al. demonstrated a significant increase in AD development among smokers with a RR of 1.45 (95% CI: 1.16–1.80) [27]. Tobacco, in addition to causing damage at the biochemical level, has been shown to modify the brain structure since it alters the structural

integrity of the gray matter in different brain areas. These structural alterations can affect various neurocognitive domains such as attention, executive skills, learning or memory [28, 29].

- Physical activity: Physical activity can be considered a protection factor in AD. In a meta-analysis carried out by Guure et al., they concluded an OR (Odds Ratio) of 0.62 (95% CI: 0.49–0.75) [30]. Moderate exercise (3,000 MET-minutes/week or 3 weekly sessions of vigorous physical activity up to 1,500 MET-minutes/week) increases $A\beta_{42}$ levels (which means that less $A\beta_{42}$ is deposited in the neurons, and therefore fewer $A\beta$ plaques), decreases the fraction of both total Tau (*T*-Tau) and phosphorylated Tau (*P*-Tau) and improves cognitive performance [31, 32].

Cardiovascular risk factors

Hypertension: Having hypertension (high blood pressure, HBP) during adulthood (not related to the HBP of onset in old age) has been associated with an increased risk of developing AD by 50% in old age. Besides, the adequate treatment of HBP has been related to the reduction of the risk of development of AD. The calculated OR was 1.61 (95% CI: 1.16–2.24) [33, 34].

Diabetes: Diabetes is one of the major pathologies that are arising in the 21st century (especially Type 2 diabetes that is associated with obesity). Type 2 diabetes has an estimated prevalence above 70 years of 30.3% (95% CI: 28.0–32.7). The prevalence figure is double that in other age ranges. It has been estimated that diabetes has a RR of 1.39 (95% CI: 1.17–1.66) for the development of AD [35, 36].

Machine learning and Alzheimer's disease

In recent years, new statistical analysis techniques have been added, until now meta-analytical strategies, both for etiological research, risk factors, and for the process of developing new therapies in AD. Machine learning (ML) is a branch of artificial intelligence that offers methods and techniques that can assist in diagnostic and prognostic problems for several medical applications. ML is generally used for research in medical data analysis since it enables the extraction of meaningful patterns from examples, being potentially useful for computer-aided diagnosis and decision support systems [37]. ML requires a set of training samples which are employed to create a mathematical model. This model is then validated and optimized using another set of samples called validation set. Finally, the validated model is tested against a new set of samples, which have not been previously employed neither to train nor to validate the model. This data partition strategy is employed to evaluate the performance of the model simulating a real scenario where the ML algorithm is used in clinical practice. ML-based generative models provide much more information than specific models, thereby enabling a simultaneous and detailed assessment of different risks [38, 39]. Until now, most studies relied on meta-analytic strategies, but there are no existing studies where these risk factors are analyzed through ML techniques including sociodemographic, clinical, and analytical variables.

Examples of the use of ML to study AD include: the development of specific medications for AD [38]; prediction of transition from mild cognitive impairment to AD dementia using neuroimaging modalities [40]; prediction and classification using Apolipoprotein-E genotype and neuroimaging [41]; classification of frontotemporal dementia and AD [42]; genetically dissect transcriptomic profiles imputed in AD (gene-based per tissue) [43]; among others.

Particularly, ML has been employed in several works related with the analysis of AD clinical variables. In [44], several ML techniques were evaluated for AD screening using sociodemographic and clinical data from different screening tests. In [45], Jammeh et al. also evaluated different ML algorithms for AD screening using clinical data and diagnosis records. Boustani et al. used electronic medical record data to develop a passive digital signature for early identification of AD through logistic regression [46]. In these studies, no variable selection was performed. However, Weakley et al. developed an ML approach based on variable selection using clinical data with the goal of reducing the amount of screening tests required to detect cognitive impairment [47]. In [48], a methodology based on a combination of ML and semi-parametric survival analysis was employed to find the most relevant variables for predicting cognitive impairment and dementia. Finally, Johnson et al. employed a genetic algorithm combined with logistic regression for the prediction of AD progression using clinical data [49]. Regarding to the use of a combination of clinical and analytical data, Fisher et al. employed an unsupervised ML method based on the Conditional Restricted Boltzmann Machine for comprehensive forecasting of AD progression [39]. However, no variable selection was carried out in this study. As it can be seen, there are several publications

in which the ML is applied to the diagnosis and evolution of AD, but there are only a few that introduce the combination of several variable types (sociodemographic, clinical, and analytical) and the use of genetic algorithms and ML techniques to develop a processing framework that explore the most relevant variables for an accurate identification of AD. This study has the aim of demonstrating, as proof-of-concept, the potential of ML for achieving an automatic discrimination between controls and subjects with major neurocognitive disorder (MNCD) by employing the most appropriate variables.

MATERIALS AND METHODS

Participants

This research is based on an observational-analytical design. Two research groups were established: MNCD group (over 65 years of age, having a diagnosis of primary MNCD: AD and frontotemporal dementia) and control group (over 65 years of age, having no diagnosis of dementia, and a Pfeiffer test result less than or equal to 2).

The subject selection was based on a stratified assignment in the following populations: subjects of outpatient consultations at the *Hospital Insular de Lanzarote* (Spain), the *Asociación de Alzheimer Gran Canaria* (Spain), and students of *Peritia et doctrina* at the University of Las Palmas de Gran Canaria (Spain). A strict protocol was followed between March 15 and October 1, 2019. This protocol involved the following steps. First, a presentation of the objectives and methods of the investigation was made to the subjects. In case of acceptance, an informed consent form was signed (in the case of the MNCD group, it was delivered to a family member or to the legal guardian). Next, a structured interview was carried out to gather information (in the case of the MNCD group, it was performed to a family member or to the legal guardian). Finally, the extraction of blood in a coagulation tube was carried out by a specialized nurse in both groups. This blood analysis was not performed in an acute situation (infection, exacerbation, etc.). The same protocol methodology was applied in both groups.

The study protocol and consent procedures were approved by the *Comité de Bioética* of the Hospital Universitario de Gran Canaria Doctor Negrin (reference 2019–054–1).

The sample is made up of 84 subjects, of whom 22 are men (26.2%) and 62 women (73.8%) with a mean age of 79.8 years (SD 8.48, minimum 65 and maximum 100 years). The MNCD group consists of 46 subjects with a mean age of 81.2 years (SD 7.28) and the control group 78.1 years (SD 9.56). The *p*-value for this age difference is 0.088 (using a two-tailed *t*-test), therefore, they do not show statistically significant differences (comparable groups).

Collected participant variables

From each participant, three sets of qualitative and quantitative variables were collected: sociodemographic (qualitative), clinical (qualitative and quantitative), and analytical variables (quantitative). Table 1 shows the qualitative sociodemographic and clinical variables collected from the participants and their respective percentage in each group. Additionally, the Chi-square test was employed in these variables to evaluate if the test rejects the null hypothesis that each variable is independent to the diagnosis at the 5% significance level. Table 2 shows the age and the quantitative clinical and analytical variables collected from the participants and their statistics per group. *p*-values were obtained to evaluate the null hypothesis that the control and NMDC groups have equal means. On one hand, a two-tailed *t*-test at the 5% significance level was computed in the variables where both control and MNCD groups had normal distributions. On the other hand, a two-tailed Wilcoxon Rank-Sum test at the 5% significance level was computed in the variables where at least one group had a non-normal distribution. In total, 38 variables (including age and gender) were collected. In the first column of Table 1 and Table 2, the ratio of subjects with no missing values for each variable with respect to the total of subjects is presented using square brackets. The total number of missing values found in the database is 119. These missing values are randomly distributed among the different subjects and variables. The percentage of distribution within control and MNCD groups presented in both tables were computed with respect to the total of subjects in the dataset.

Machine learning classification approach

The main goal of this work is to demonstrate, as a proof-of-concept, the capabilities of machine learning algorithms to automatically diagnose MNCD in a subject using the previously described variables. In addition, this study will show the potential of employing the combination of genetic algorithms and supervised machine learning approaches to identify

Table 1
Qualitative sociodemographic and clinical variables of participants with MNCD and controls

Variable [Subjects with no missing values/ Total subjects]	Characteristic	MNCD group		Control group		p (Chi-square test*)
		N	%	**N**	%	
Sociodemographic						
Sex [84/84]	Male	13	28.3	9	23.7	0.635
	Female	33	71.7	29	76.3	
Marital status [84/84]	Single	5	10.9	4	10.5	0.577
	Married	15	32.6	10	26.3	
	Widower/widow	23	50.0	18	47.4	
	Divorced	3	6.5	6	15.8	
Work activity prior to retirement [83/84]	Skilled worker	8	17.4	17	44.7	**0.023**
	Unskilled worker	17	37.0	14	36.8	
	Entrepreneur	1	2.2	0	0.0	
	Housewife	19	41.3	7	18.4	
Coexistence format [84/84]	Spouse	14	30.4	12	31.6	**0.002**
	Alone	2	4.3	11	28.9	
	Another relative	23	50.0	15	39.5	
	Nursing home	7	15.2	0	0.0	
Education level [83/84]	University	7	15.2	1	2.6	0.100
	Media	25	54.3	20	52.6	
	Basics	8	17.4	7	18.4	
	No studies	5	10.9	10	26.3	
Previous intellectual activity [82/84]	Reading > 10 books/year	15	32.6	18	47.4	**0.032**
	Between 5–10 books/year	9	19.6	13	34.2	
	Reading < 5 books/year	20	43.5	7	18.4	
Social relationships [83/84]	Good	17	37.0	29	76.3	**< 0.001**
	Normal	15	32.6	8	21.1	
	Nonexistent	13	28.3	1	2.6	
Clinical						
Psychiatric history [84/84]	Depression	21	45.7	4	10.5	**0.002**
	Other background	1	2.2	1	2.6	
	No background	24	52.2	33	86.8	
Cardiovascular history [84/84]	Ischemic heart disease	8	17.4	9	23.7	0.834
	Atrial fibrillation	10	21.7	7	18.4	
	Other background	4	8.7	2	5.3	
	No background	24	52.2	20	52.6	
Neurological history [83/84]	ICTUS / TIA	9	19.6	2	5.3	0.189
	Headache / migraine	2	4.3	1	2.6	
	CET	1	2.2	0	0.0	
	Epilepsy	1	2.2	0	0.0	
	No background	33	71.7	35	92.1	
Kidney history [84/84]	CKD	11	23.9	12	31.6	0.374
	Others	0	0.0	1	2.6	
	No background	35	76.1	25	65.8	
Pneumology history [84/84]	COPD	2	4.3	6	15.8	0.054
	Other pathologies	2	4.3	5	13.2	
	No background	42	91.3	27	71.1	
Family history of dementia [83/84]	Father mother	19	41.3	5	13.2	**0.005**
	Other family	6	13.0	3	7.9	
	No background	20	43.5	30	78.9	
Smoking history [84/84]	Yes	15	32.6	13	34.2	0.877
	No	31	67.4	25	65.8	
Alcoholism history [83/84]	High	3	6.5	1	2.6	0.212
	Moderate	4	8.7	5	13.2	
	Mild	19	41.3	23	60.5	
	No background	19	41.3	9	23.7	
HTA history [84/84]	Yes	25	54.3	5	13.2	**< 0.001**
	No	21	45.7	33	86.8	
Diabetes history [84/84]	Yes	18	39.1	3	7.9	**< 0.001**
	No	28	60.9	35	92.1	

COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease; ICTUS/TIA, stroke/transient ischemic accident; CET, cranioencephalic trauma; HTN, hypertension. *The Chi-square test was employed to evaluate if the test rejects the null hypothesis that each variable is independent to the diagnosis at the 5% significance level.

Table 2
Clinical and analytical quantitative variables of participants with MNCD and controls

Variable [Total without missing values/Total]	MNCD group		Control group		p (t-test*/Wilcoxon Rank-Sum test†)
	Mean	SD	Mean	SD	
Age [84/84]	81.28	7.28	78.11	9.56	0.088*
Clinical					
SBP (mm Hg) [83/84]	124.60	13.39	123.11	15.03	0.422†
DBP (mm Hg) [83/84]	70.27	10.40	68.11	10.16	0.349†
Pfeiffer Test[1] [84/84]	7.54	2.80	0.29	0.69	**< 0.001**†
Barthel Scale[2] [84/84]	59.24	30.80	85.92	25.54	**< 0.001**†
Analytical					
Hemoglobin (g/dL) [79/84]	12.48	1.69	13.27	2.08	0.066*
Average Corpuscular Volume (fL) [79/84]	91.40	4.69	90.88	3.89	0.745†
Corpuscular Hemoglobin Media (pg) [79/84]	30.21	1.95	30.04	1.72	0.681*
Platelets (10^3 μL) [80/84]	244.40	70.37	225.66	65.28	0.258†
Leukocytes (10^3 μL) [80/84]	6.97	1.88	6.60	1.85	0.372*
Neutrophils (10^3 μL) [80/84]	4.27	1.61	3.79	1.46	0.174†
Lymphocytes (10^3 μL) [80/84]	1.87	0.77	1.95	0.82	0.674*
Monocytes (10^3 μL) [80/84]	0.57	0.21	0.65	0.34	0.305†
Glucose (mg/dL) [81/84]	114.93	31.99	104.24	23.88	0.101†
Creatinine (mg/dL) [78/84]	0.94	0.28	1.03	0.43	0.398†
Glomerular Filtering (CKD-EPI) (mL/min) [72/84]	67.55	18.44	61.89	21.82	0.238*
Sodium (mEq/L) [77/84]	141.39	3.04	141.22	4.03	0.986†
Potassium (mEq/L) [77/84]	5.50	6.50	4.50	0.45	0.612†
Alanine Aminotransferase (IU/L) [76/84]	19.99	29.38	16.34	7.97	0.486†
Cholesterol (total) (mg/dL) [72/84]	169.97	37.50	183.61	43.40	0.158*
LDL cholesterol (mg/dL) [64/84]	99.18	34.47	96.70	38.87	0.788*

SBP, systolic blood pressure (in mm Hg); DBP, diastolic blood pressure (in mm HG); SD, standard deviation [1]Pfeiffer Test or Short Portable Mental State Questionnaire: short screening questionnaire made up of ten questions that measures the degree of cognitive impairment: areas evaluated are short and long-term memory, information on everyday events, calculation skills, and orientation. [2]Barthel Scale: instrument used for the functional assessment of a patient. Score from 0 to 100, with 100 being independence and 0 dependence for basic activities of daily living in the following areas: feeding; bathing, grooming, dressing, bowel control, bladder control, toilet use, transfer, mobility on level surfaces and stairs. *The two-tailed t-test at the 5% significance level was computed in the variables where both control and MNCD groups had normal distributions. †The two-tailed Wilcoxon Rank-Sum test at the 5% significance level was computed in the variables where at least one group had a non-normal distribution.

the most relevant variables that offer the best classification results.

For the experiments, all the variables collected from the participants were employed, except the Pfeiffer Test result, since this variable was employed to select the participants of the control group. Therefore, 37 variables were included in the study.

Figure 1 shows the block diagram of the data processing framework developed for this work. This framework is based on two main stages: data preprocessing and supervised machine learning analysis, and variable selection.

Data pre-processing

The data pre-processing approach is based on two main steps as presented in Fig. 1A: missing values replacement and data normalization. First, the variable database is divided in four independent datasets: Training Set (60%; N = 51), Validation Set 1 (10%; N = 8), Validation Set 2 (10%; N = 8), and Test Set (20%; N = 17). The validation set is divided in two sets with the goal of employing the first set to find the optimal variables for data classification and then, employing the second set for the performance evaluation of the different types of classification algorithms. Since the collected database is not large enough, the data partition was randomly repeated 10 times, allowing to obtain more robust results with the proposed processing framework. An additional advantage of this repeated data partition strategy is to avoid bias due to the missing values in data. The results obtained with different repetitions can help to avoid the bias of such missing values, since each repetition will have different missing samples during training, validation, and testing.

In the first step, the missing values replacement is performed using a method based on the k-Nearest-Neighbor (kNN) algorithm for the imputation of the

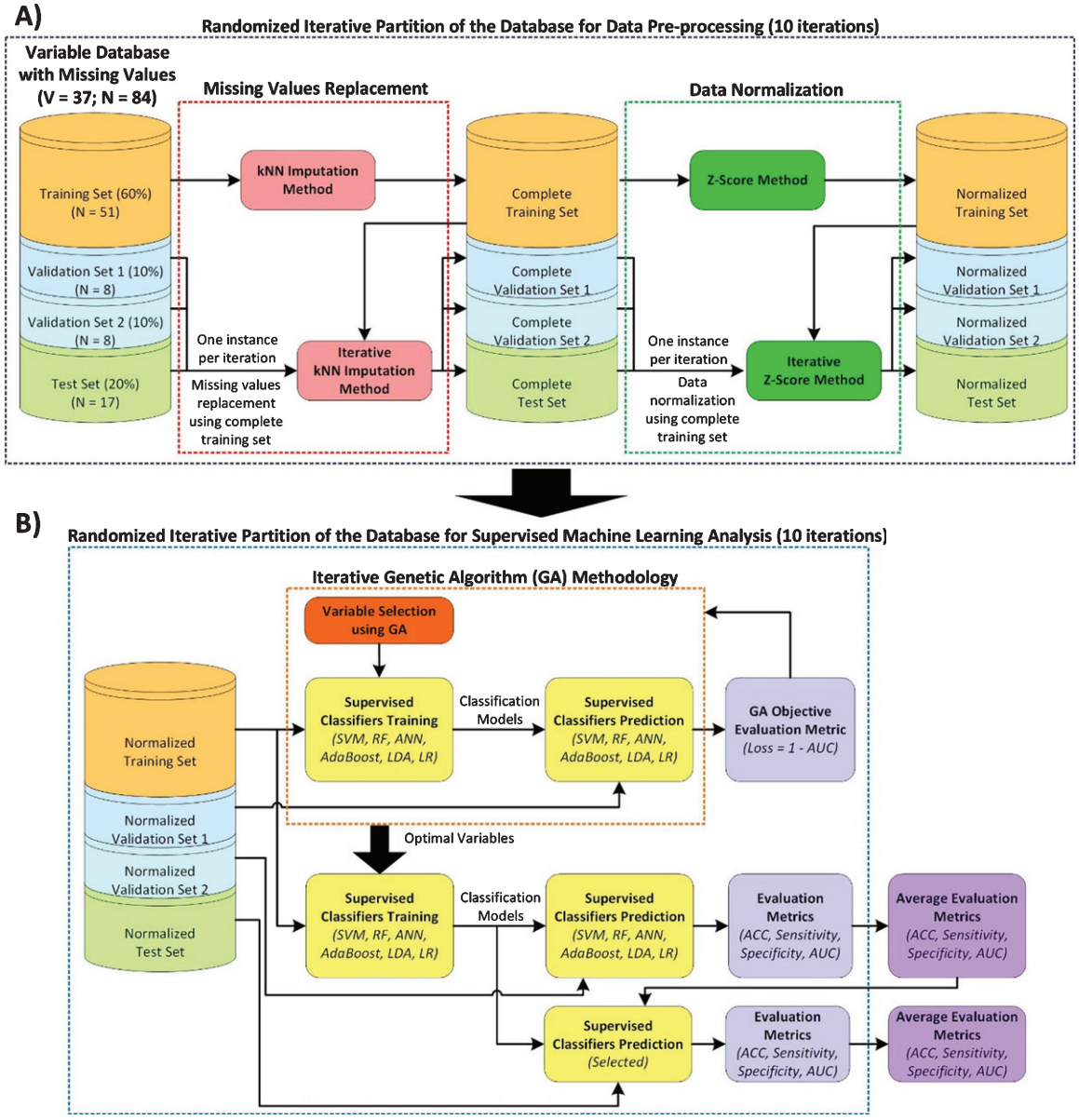


Fig. 1. Block diagram of the data processing framework developed in this study. A) Data pre-preprocessing approach. B) Supervised machine learning analysis and variable selection.

missing data [50]. Initially, this process is performed in the Training Set, since data to train the model should not have missing values. After this, the imputation is applied independently to each instance of the validation and test sets using the complete Training Set for computing the missing values, i.e., only the Training Set is used in the search space for data imputation. This method simulates a real scenario where data from a new patient are obtained, but some of the data variables are missing. Therefore, these values are inferred using data from previous patients available in the database, i.e., the Training Set. This data imputation would not bias the classification results since the labels are not considered in the data imputation. The proposed strategy to impute data from both validation and test sets is intended to avoid possible data leak, since no information outside the Training Set is used either to train the model, or for imputation. The same argument is applied to the next step where the z-score normalization is performed.

In the second step, the numerical variables of the Training Set are normalized using the z-score method [51]. This method is based on the relationship between the mean and the standard deviation of the values presented in a dataset. As in the kNN imputation, each instance of the validation and test sets was normalized independently, using the mean and the standard deviation from the Training Set. In this step, for certain classifiers employed in the supervised machine learning analysis, the categorical values were transformed to dummy variables [52]. This technique makes use of the one-hot encoding where each category is represented by binary values. In this process, only a single value can be 1, while the others are set to 0, indicating the absence or presence of a specific categorical value within a variable.

Supervised machine learning analysis and variable selection

The proposed ML framework to determine the more suitable variables for an automatic discrimination between the NMCD and control groups was based on a combination of supervised ML and genetic algorithms.

The six supervised classifiers employed to perform the experiments were: Support Vector Machines (SVMs), Random Forest (RF), Artificial Neural Networks (ANNs), AdaBoost Ensemble Classifier (AB), Linear Discriminant Analysis (LDA), and Logistic Regression (LR). The SVM classifier was configured with the linear kernel and the cost parameter equal to 1. The RF classifier was set with 500 trees. The AB classifier was configured with the decision tree ensemble and 100 ensemble learning cycles. The ANN was configured with 10 hidden layers, 500 epochs and the scaled conjugate gradient backpropagation as optimizer for training. In the case of the LDA classifier, the discriminant type selected was pseudolinear, which means that all classes have the same covariance matrix. Finally, a binomial distribution was selected in the LR classification. These classifiers were selected because they are some of the more commonly employed in the literature for machine learning data classification in medical applications [53].

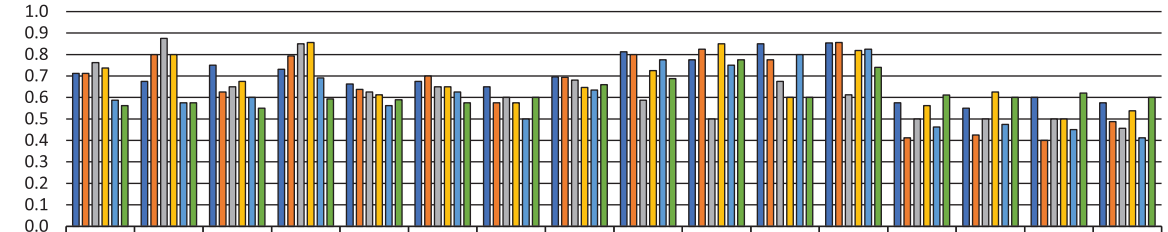
The variable selection process was performed using the Genetic Algorithm (GA) [54]. The GA is an optimization algorithm that mimics the process of natural selection, finding an optimal solution to a problem. In this case, the GA was employed to identify which variables provide the most relevant information to discriminate between the NMCD and control groups. The specific parameters employed in the GA are as follows: a population size of 200; a number of generations equal to the number of variables multiplied by 100 (37×100) with a stop criteria of 50 generations if the result does not change; a binary population type, which indicates whether the variable is relevant or not; a crossover function, where the scattered type was selected; and the gaussian method for the mutation function.

As presented in Fig. 1B, the Training Set and the Validation Set 1 were employed to perform the iterative GA methodology. In this process, the classifiers were trained using the Training Set composed by the initial variables selected by the GA. Then, the classification models generated were employed to classify the Validation Set 1, computing the loss of the prediction based on the AUC metric ($Loss = 1 - AUC$). This process is repeated until finding the minimum loss value, which will represent the best classification models. Moreover, the complete procedure is repeated 10 times with 10 different randomized database partitions to enhance the significance of the results. After this process, each classification model can select different relevant variables in each iteration. The criterion to select the most relevant variables with the proposed methodology is based on the variables which obtained an average repetition value higher than 5 among the six different classification algorithms and the 10 randomized data partitions.

Once the optimal variables were identified, the optimal classification models were generated using the Training Set and the Validation Set 2 was classified to evaluate the performance of the classifiers and the results were averaged due to the 10 dataset partitions. At this point, the classification models were generated and evaluated employing all the selected variables and employing the selected variables for each variable set (Sociodemographic set, Clinical set, and Analytical set).

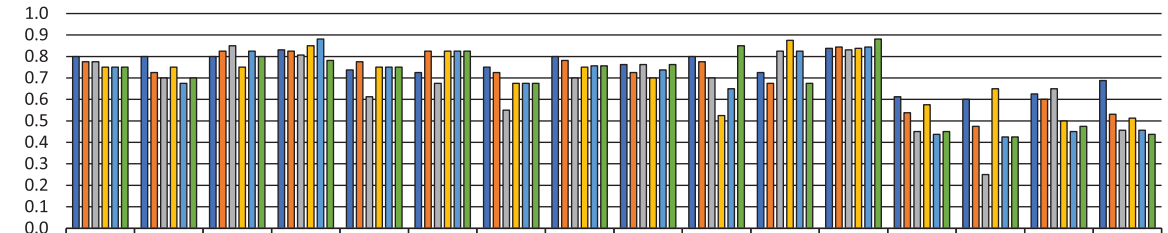
Finally, the best classifiers and the best variable sets were selected to perform the classification on the Test Set. In this process the results were obtained, which were also averaged from the 10 dataset partitions.

For comparison purposes, the same classifiers were trained using the Training Set composed by the 37 variables ($V = 37$) and also employing the three variable subsets independently (Sociodemographic set; $V = 8$, Clinical set; $V = 14$, and Analytical set; $V = 16$). The Validation Set 1 was classified with the models generated to compare the results with the variable selection approach.



	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
	Complete set (V = 37)				Sociodemographic set (V = 8)				Clinical set (V = 14)				Analytical set (V = 16)			
SVM	0.71	0.68	0.75	0.73	0.66	0.68	0.65	0.70	0.81	0.78	0.85	0.85	0.58	0.55	0.60	0.58
RF	0.71	0.80	0.63	0.79	0.64	0.70	0.58	0.69	0.80	0.83	0.78	0.86	0.41	0.43	0.40	0.49
ANN	0.76	0.88	0.65	0.85	0.63	0.65	0.60	0.68	0.59	0.50	0.68	0.61	0.50	0.50	0.50	0.46
AB	0.74	0.80	0.68	0.86	0.61	0.65	0.58	0.65	0.73	0.85	0.60	0.82	0.56	0.63	0.50	0.54
LDA	0.59	0.58	0.60	0.69	0.56	0.63	0.50	0.63	0.78	0.75	0.80	0.83	0.46	0.48	0.45	0.41
LR	0.56	0.58	0.55	0.59	0.59	0.58	0.60	0.66	0.69	0.78	0.60	0.74	0.61	0.60	0.62	0.60

(A)



	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
	Complete set (V=12)				Sociodemographic set (V = 3)				Clinical set (V = 4)				Analytical set (V = 5)			
SVM	0.80	0.80	0.80	0.83	0.74	0.73	0.75	0.80	0.76	0.80	0.73	0.84	0.61	0.60	0.63	0.69
RF	0.78	0.73	0.83	0.83	0.78	0.83	0.73	0.78	0.73	0.78	0.68	0.84	0.54	0.48	0.60	0.53
ANN	0.78	0.70	0.85	0.81	0.61	0.68	0.55	0.70	0.76	0.70	0.83	0.83	0.45	0.25	0.65	0.46
AB	0.75	0.75	0.75	0.85	0.75	0.83	0.68	0.75	0.70	0.53	0.88	0.84	0.58	0.65	0.50	0.51
LDA	0.75	0.68	0.83	0.88	0.75	0.83	0.68	0.76	0.74	0.65	0.83	0.84	0.44	0.43	0.45	0.46
LR	0.75	0.70	0.80	0.78	0.75	0.83	0.68	0.76	0.76	0.85	0.68	0.88	0.45	0.43	0.48	0.44

(B)

Fig. 2. Classification results obtained in the Validation Set 2. A) Reference results using all the variables available in each variable set. B) Results obtained after the selection of the most relevant variables in each variable set.

RESULTS

Overall accuracy (ACC), sensitivity, and specificity metrics were employed to evaluate the classification performance. In addition, the receiver operating characteristic (ROC) curve was employed to find the most relevant variables for the classification using the supervised classifiers, obtaining the best performance using the AUC (Area Under the Curve) metric. AUC was also employed to compare the results among the different approaches.

Statistical analyses were performed to evaluate the results obtained in the experiments. For the statistical comparison between the different pair of variable sets (conformed by the results of the 10 random partitions of the 6 classifiers), we used one-sided *t*-test statistics. With the goal of providing a more intuitive interpretation of the statistical analysis, for each pair of variable sets, e.g., A and B, we performed a right-sided *t*-test to verify the superiority of A with respect to B. Additionally, we used a left-sided *t*-test to verify the superiority of B with respect to A.

Validation results

Reference classification results

Figure 2A shows the reference results obtained in the classification of the Validation Set 2 using all the

Table 3
p-values obtained after testing the two population means of the different reference sets of the Validation Set 2 in a right-tailed and left-tailed *t*-test evaluation performed at the 5% significance level

Variable set	p (Right-tailed *t*-test)				p (Left-tailed *t*-test)			
	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
Complete versus Sociodemographic	0.110	**<0.001**	0.933	**0.042**	0.890	0.999	0.067	0.958
Complete versus Clinical	0.853	0.860	0.664	0.932	0.147	0.140	0.336	0.068
Complete versus Analytical	**0.017**	**<0.001**	0.916	**0.003**	0.983	1.000	0.084	0.997
Sociodemographic versus Clinical	0.995	1.000	0.128	1.000	**0.005**	**<0.001**	0.872	**<0.001**
Sociodemographic versus Analytical	0.123	0.093	0.435	0.089	0.877	0.907	0.565	0.911
Clinical versus Analytical	**<0.001**	**<0.001**	0.843	**<0.001**	1.000	1.000	0.157	1.000

variables available in the database. The complete set ($V=37$) and the other three variable sets were classified using the six proposed classifiers. The ACC, sensitivity, specificity, and AUC metrics were computed. As it can be observed, the best ACC (81%) was achieved with the SVM classifier using the Clinical set ($V=14$), offering the best specificity (85%) and sensitivity and AUC values of 78% and 0.85, respectively. However, the best sensitivity (88%) was obtained with the ANN classifier using the Complete set, but the specificity decreased to 65%, having an ACC of 76% and an AUC of 0.85. Finally, the best trade-off between specificity and sensitivity was obtained by the SVM and RF configurations using the Clinical set.

A statistical analysis has been performed to evaluate the obtained results among the different variable sets and evaluation metrics. Table 3 shows the results of the right-tailed and left-tailed t-test assessment, performed at 5% significance level, to evaluate the alternative hypothesis that the population mean of the first set is greater or lower, respectively, than the population mean of the second set. To compute the *p*-value, the results of the 10 random partitions of the 6 classifiers were considered. It is worth noticing that the mean sensitivity of the Complete set is higher than the means of the Sociodemographic and the Analytical sets, being highly statistically significant. Additionally, the mean ACC, sensitivity, and AUC of the clinical set are higher than the results of the Analytical set, being also highly statistically significant. On the other hand, the mean ACC, sensitivity, and AUC of the Clinical set are higher than the results obtained in the Sociodemographic set, being statistically significant the ACC and highly statistically significant the sensitivity and ACC results. No significant differences were found between the Complete and Clinical sets.

Optimal variable classification results

This section presents the classification results obtained in the Validation Set 2 using the most relevant variables identified with the GA algorithm using the Validation Set 1. Figure 3 shows the boxplot results of the variable selection process where each boxplot represents the number of repetitions of such variable among the 10 random partitions of the database and the six different classifiers employed in the experiments. The most relevant variables were selected if their average repetition value (center cross in the boxplot) were higher than 5. Table 4 shows the average repetition value for each variable and the *p*-value obtained after performing a right-tailed Wilcoxon Rank-Sum test at 5% of significance level to evaluate the alternative hypothesis that the population median of each variable is higher than the population median of the average of the remaining variables. The Wilcoxon Rank-Sum test was computed due to at least one group presented a non-normal distribution. Bold values in the "Average Repetition Value" column indicate the most relevant variables selected by the proposed methodology. The statistical analysis shows that the population median value of the "Coexistence format", "Social relationships", "Average Corpuscular Volume", and "Lymphocytes" variables are statistically significantly higher than their respective population median of the remaining variables. These results were computed considering the average repetition values obtained from the 10 random partitions of the six different classifiers. Additionally, it has been found that there is no statistically significant difference between the variable selection process carried out with the different classifiers and the GA.

Figure 2B shows the classification results using the selected variables that were computed with each proposed classifier for the Complete set ($V=12$) and also for each independent variable set (Sociodemographic set; $V=3$, Clinical set; $V=4$, and Analytical

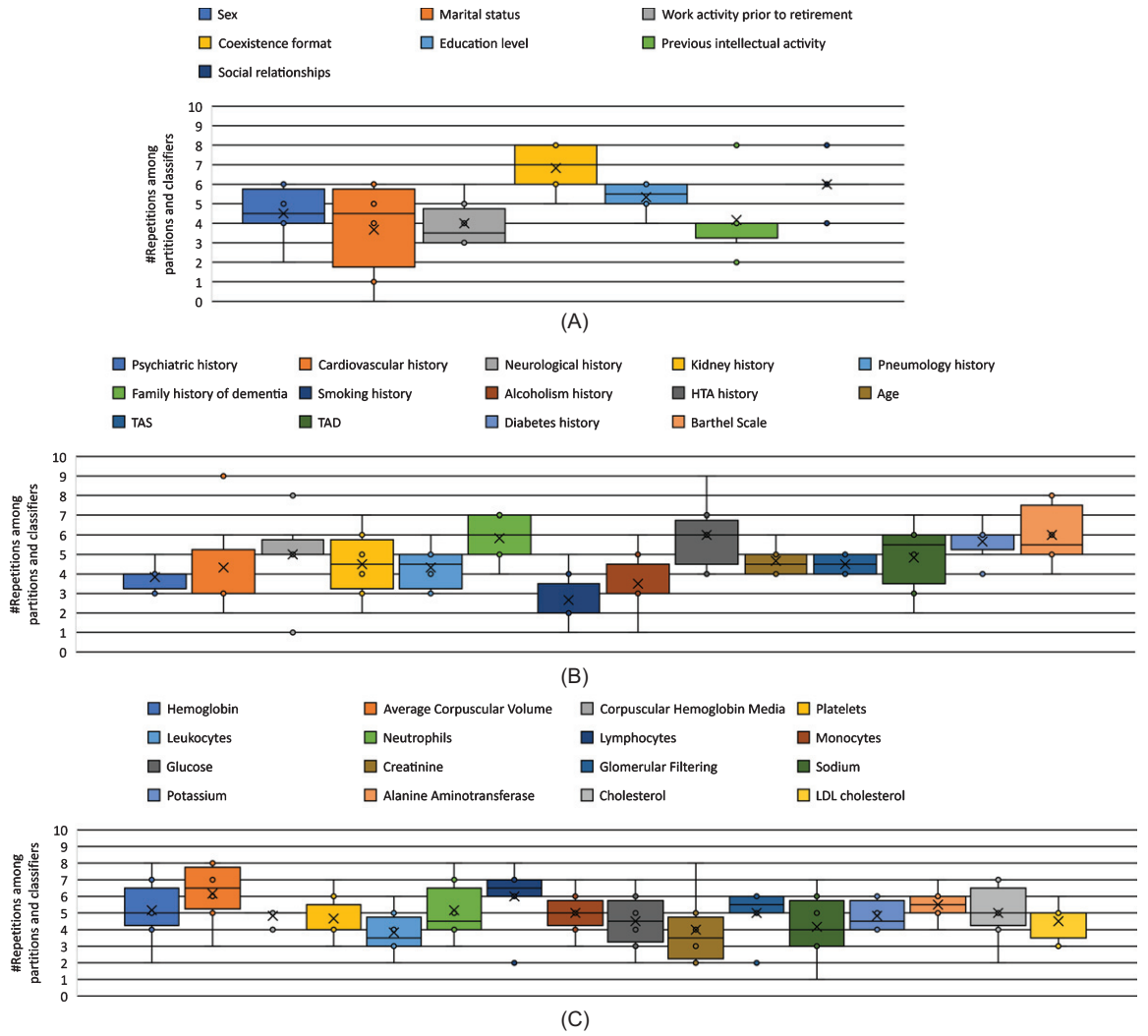


Fig. 3. Boxplots of the number of repetitions of each variable among the partitions and type of classifiers for each variable set. A) Sociodemographic set. B) Clinical set. C) Analytical set. The box boundaries represent the IQR (interquartile range) of the results. Central bars and error bars depict median and minimum/maximum values of repetitions, respectively. The cross represents the average repetition value. The small dots outside the minimum/maximum values represent the outliers.

set; $V = 5$). In general, the results revealed an improvement of the classification when using the selected variables with respect to the reference results, especially in the AUC metric. The most balanced results were obtained with the SVM approach using the Complete set achieving 80% of ACC, sensitivity and specificity, and an AUC of 83%. This configuration achieved the best ACC. However, the best sensitivity (83%) was obtained with several classifiers (RF, AB, LDA, and LR) using the Sociodemographic set. The best specificity (88%) was obtained with the AB classifier using the Clinical set. Respect to the AUC value, the highest results (0.88) were obtained with the LDA and LR classifiers using the Complete and Clinical sets, respectively.

The statistical analysis of these results reveals (Table 5) that the Complete set ($V = 12$), which involves relevant variables from the three different sets, is the most relevant set for performing the classification based on the optimal variables selected using the proposed methodology in the Validation Set 2. According to these results, the Analytical set is the worst one. The population mean is always lower than the population mean of the other three sets, being highly statistically significant in almost all cases for the four metrics. The Sociodemographic set performs better in the sensitivity metric than the Clinical set.

Table 4
Average repetition values (among the ten partitions and six classifiers) for the identification of the most relevant variables using the proposed methodology for NMCD identification. Variables with average repetition values higher than 5 where selected as most relevant (bold numbers in the Average Repetition Value column). Last column indicates the result of the right-tailed Wilcoxon Rank-Sum test performed at the 5% significance level to evaluate the alternative hypothesis that the population median of each variable is higher than the population median of the average of the remaining variables. The Wilcoxon Rank-Sum test was computed due to at least one group had a non-normal distribution. (*) Indicates the variables which were identified as statistically significant in the analysis performed in Tables 1 and 2, where the variables were compared with the diagnostic outcome

Variable	Average Repetition Value	*Selected Variables p* (Right-tailed Wilcoxon Rank-Sum test)
Sociodemographic		
Age	4.7	0.746
Sex	4.5	0.636
Marital status	3.7	0.643
Work activity prior to retirement*	4.0	0.911
Coexistence format*	**6.8**	**0.003**
Education level	**5.3**	0.108
Previous intellectual activity*	4.2	0.979
Social relationships*	**6.0**	**0.026**
Clinical		
Psychiatric history*	3.8	0.992
Cardiovascular history	4.3	0.839
Neurological history	5.0	0.195
Kidney history	4.5	0.634
Pneumology history	4.3	0.739
Family history of dementia	**5.8**	0.108
Smoking history	2.7	0.994
Alcoholism history	3.5	0.911
HTA history*	**6.0**	0.186
SBP	4.5	0.856
DBP	4.8	0.278
Diabetes history*	**5.7**	0.064
Barthel Scale*	**6.0**	0.115
Analytical		
Hemoglobin	**5.2**	0.397
Average Corpuscular Volume	**6.2**	**0.050**
Corpuscular Hemoglobin Media	4.8	0.408
Platelets	4.7	0.846
Leukocytes	3.8	0.904
Neutrophils	**5.2**	0.633
Lymphocytes	**6.0**	**0.027**
Monocytes	5.0	0.392
Glucose	4.5	0.634
Creatinine	4.0	0.905
Glomerular Filtering (CKD-EPI)	5.0	0.108
Sodium	4.2	0.639
Potassium	4.8	0.630
Alanine Aminotransferase	**5.5**	0.113
Cholesterol (total)	5.0	0.397
LDL cholesterol	4.5	0.509

However, the specificity and AUC of the Clinical set are higher than the Sociodemographic set. As in Table 3, these results were computed from 10 random partitions of the 6 classifiers.

Finally, Table 6 presents the statistical analysis of the comparison between using the reference sets, which include all the variables, and the optimal sets, which include the most relevant variables selected with the proposed methodology. These results demonstrate that the Complete, Sociodemographic and Clinical optimal sets outperform their respective reference sets, achieving in most of the metrics higher and highly statistically significant results.

Test results

The Test Set was evaluated using the most relevant configurations obtained in the validation results. These configurations involved the use of the Complete and Clinical variable sets optimized using the GA methodology to evaluate their performance in the Test Set. A statistical analysis was performed using a right-tailed *t*-test to compare the population means of both sets (Complete and Clinical). Furthermore, an additional statistical analysis was performed over each evaluation metric in order to establish a comparative of performance of the different classifiers using a right-tailed *t*-test for each variable set. Figure 4 shows the classification results of the Test Set, where it is possible to observe that the RF classifier using the Complete set achieved the best sensitivity result of 100%, identifying correctly all the subjects with NMDC in the Test Set. This result is highly statistically significant respect to the other classifiers in the Complete set ($p < 0.001$). Furthermore, RF and LDA provided the best AUC results (0.97) using the Clinical set. These results are statistically significant with respect to the SVM, ANN, ADA, and LR results in the Clinical set ($p < 0.002$). The detailed results of this analysis are presented in the Supplementary Table 1. Finally, the statistical analysis of the results obtained after comparing the two population means of the Complete and Clinical sets using the six classifiers (Table 7) reveals that the Complete set performs better than the Clinical set in terms of ACC ($p = 0.012$) and sensitivity ($p < 0.001$). However, no significant differences were found in the specificity and AUC metrics.

DISCUSSION

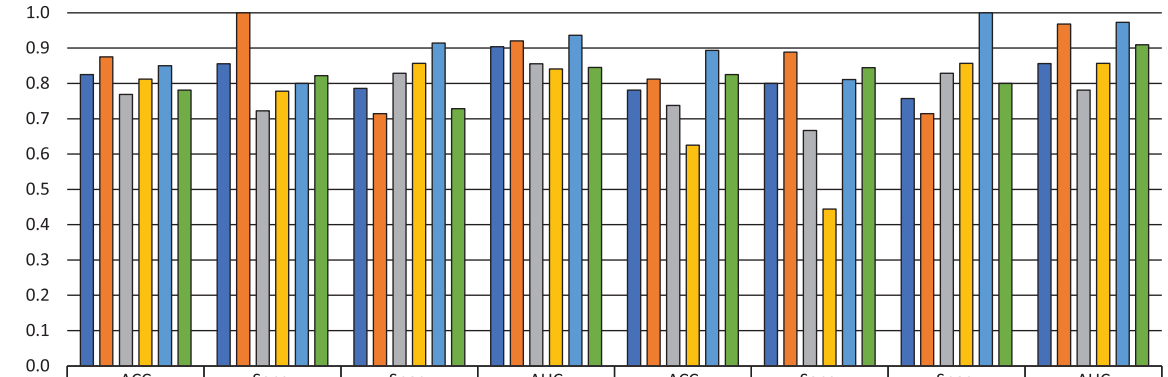
The progressive increase in life expectancy has economic, social, and health consequences. Recent

Table 5
p-values obtained after testing the two population means of the different optimal sets of the Validation Set 2 in a right-tailed and left-tailed *t*-test evaluation performed at the 5% significance level

Variable set pair (set1 versus set2)	p (Right-tailed *t*-test)				p (Left-tailed *t*-test)			
	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
Complete-GA versus Sociodemographic-GA	0.057	0.723	**<0.001**	**0.014**	0.943	0.277	0.999	0.986
Complete-GA versus Clinical-GA	0.146	**0.040**	0.694	0.447	0.854	0.960	0.306	0.553
Complete-GA versus Analytical-GA	**<0.001**	**<0.001**	**<0.001**	**<0.001**	1.000	1.000	1.000	1.000
Sociodemographic-GA versus Clinical-GA	0.750	**0.012**	1.000	0.979	0.250	0.988	**<0.001**	**0.021**
Sociodemographic-GA versus Analytical-GA	**<0.001**	**<0.001**	**0.003**	**<0.001**	1.000	1.000	0.997	1.000
Clinical-GA versus Analytical-GA	**<0.001**	**<0.001**	**<0.001**	**<0.001**	1.000	1.000	1.000	1.000

Table 6
p-values obtained after testing the two population means of the reference and optimal sets of the Validation Set 2 in a right-tailed and left-tailed *t*-test evaluation performed at the 5% significance level. GA indicates that the set belongs to the optimal set

Variable set pair (set1 versus set2)	p (Right-tailed *t*-test)				p (Left-tailed *t*-test)			
	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
Complete versus Complete-GA	1.000	0.940	1.000	1.000	**<0.001**	0.060	**<0.001**	**<0.001**
Sociodemographic versus Sociodemographic-GA	1.000	1.000	0.777	1.000	**<0.001**	**<0.001**	0.223	**<0.001**
Clinical versus Clinical-GA	0.997	0.108	1.000	0.986	**0.003**	0.892	**<0.001**	**0.014**
Analytical versus Analytical-GA	**0.048**	0.466	**0.027**	**0.041**	0.952	0.534	0.973	0.959



	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
	Complete set (V = 12)				Clinical set (V = 4)			
SVM	0.83	0.86	0.79	0.90	0.78	0.80	0.76	0.86
RF	0.88	1.00	0.71	0.92	0.81	0.89	0.71	0.97
ANN	0.77	0.72	0.83	0.86	0.74	0.67	0.83	0.78
AB	0.81	0.78	0.86	0.84	0.63	0.44	0.86	0.86
LDA	0.85	0.80	0.91	0.94	0.89	0.81	1.00	0.97
LR	0.78	0.82	0.73	0.85	0.83	0.84	0.80	0.91

Fig. 4. Classification results obtained in the Test Set using the most relevant variables in the complete and clinical variable sets.

Table 7
p-values obtained after testing the two population means of the Complete and Clinical sets of the Test Set in a right-tailed and left-tailed *t*-test evaluation performed at the 5% significance level. GA indicates that the set belongs to the optimal set

Variable set pair (set1 versus set2)	p (Right-tailed *t*-test)				p (Left-tailed *t*-test)			
	ACC	Sens.	Spec.	AUC	ACC	Sens.	Spec.	AUC
Complete-GA versus Clinical-GA	**0.012**	**<0.001**	0.790	0.646	0.988	0.999	0.210	0.354

data reveal an increasing trend in the number of patients with MNCD, predicting an increase of 87% in the European region in the period 2010–2050 [5]. Health and social costs in aging represent more than 50% of total expenditures today. Of this expenditure, more than 10% is dedicated to neurocognitive disorders (with a significant impact on both the family and the society in direct and indirect costs). A significant increase is expected both in the number of people over 70 years of age with a consequent increase in the incidence and prevalence of MNCD and, the consequent increase in the costs of treating MNCD. Among others, this is an important reason for the application of an efficient diagnostic tool and etiological search (with the consequent effective treatment). Specifically, the most common dementia is AD, and there has been an evolution in relation to the diagnostic criteria in the last decades. As AD is a polygenic and multifactorial disease with complex origins, there is not an single etiology to study and treat [14–20, 38].

The unmodifiable and established risk factors for AD are age, sex, and genetic factors, being age the main one. On the other hand, modifiable risk factors related to healthy lifestyles or educational levels have been identified and their optimization can lead to a significant decrease in the incidence of dementia [5, 20]. Both for diagnosis and to establish an adequate etiology, this research is based on the study of modifiable risk factors through machine learning algorithms. The main goal is to demonstrate, as a proof-of-concept, the capabilities of machine learning algorithms to automatically diagnose MNCD in a subject using the previously described variables.

Three sets of variables have been collected to be analyzed by blocks using the specific techniques of machine learning: sociodemographic, clinical, and analytical variables. These variables (detailed in Tables 1 and 2) separately have been shown as risk factors for the development of MNCD (except analytical variables not previously studied as risk factors), finding RR between 1.16 to 2.24. Using the different combinations of machine learning techniques and optimization algorithms, it has been possible to identify the most relevant variables for the identification of MNCD.

This study has revealed that the use of the 37 variables collected in this study (Complete set) and the use of the 14 variables from clinical data (Clinical set) has no statistically significant differences in the classification results obtained with the six ML classifiers over the Validation Set 2. However, particularly in the sensitivity metric, these two sets are statistically significant ($p < 0.001$) higher than the results obtained with the other two sets (Sociodemographic and Analytical). Regarding to the use of the GA for the identification of the most relevant variables, Table 8 shows the summary of the selected variables from each variable set using the proposed framework, which results are presented in Table 4. In this study, we have demonstrated that the classification results of the Validation Set 2 are improved when using the optimal sets, which include only the most relevant variables. This result was found highly statistically significant in most cases, except for the comparison between the Analytical sets (see Table 6). Finally, the Test Set was evaluated using the most relevant configurations (Complete and Clinical sets with the most relevant variables), revealing that the RF classifier using the Complete set offered the best sensitivity result (100%, $p < 0.001$). According to the results obtained with the proposed methodology, the most relevant set of variables to identify MNCD is given by a combination of variables selected from all types of data, i.e., Sociodemographic, Clinical, and Analytical. These results support the epigenetic theory of the MNCD, that is, the influence of the environment on the AD development.

Table 8
Summary of the most relevant variables identified by the proposed framework

Variable Set	Most Relevant Variable
Sociodemographic	Coexistence format
	Education level
	Social relationships
Clinical	Family history of dementia
	HTA history
	Diabetes history
	Barthel Scale
Analytical	Hemoglobin
	Average Corpuscular Volume
	Neutrophils
	Lymphocytes
	Alanine Aminotransferase

Limitations

The main limitation of this study is found in the relatively small number of samples which may influence the outcomes of this study in two folds. On the one hand, the classification algorithms could generalize better with a higher sample size. On the other hand, the limited number of samples may lead in type-I errors in the statistical hypothesis testing. Despite this, the proposed methodology, where 10 randomized iterative partitions of the database were

performed, allows to deal with the problems of having a very small database. Hence, these preliminary results obtained with the framework and the methodology herein proposed show a promising line of research related to the search for the etiology of AD, demonstrating the influence of the environment on its development and the potential use of ML and algorithm optimization techniques to evaluate the most relevant information for an accurate diagnosis.

Another limitation found in our study is related to the z-score normalization. This method has been shown to be sensitive to outliers compared to other normalization approaches [51]. Hence, the presence of outliers may have influenced our results. The influence of the normalization technique and also further analysis regarding the missing data management will be studied in detail in future works.

Finally, a third limitation is found in the lack of quantitative comparison of results respect to other processing approaches existing in the state-of-the-art using the same dataset. In this sense, it is not possible to know at this time which of many similar methods are actually best, on average and for each subtype of the many relevant small datasets that can help accelerate AD research progress. However, this study presents a relevant comparison through the results of six different ML supervised classifiers commonly used in the literature. Since the comparative regarding the best classifiers for MNCD identification have been performed, the most significant differences with other approaches are based on which method is used to select the most relevant variables for the subsequent classifications. In the future, further comparative with other state-of-art feature selection methods could lead into more robust conclusions about which variables are more important for MNCD identification using ML techniques.

In contrast to the aforementioned limitations, a strength of our approach is that it has been provided successful results with a very small clinical dataset. The methodology presented in this work can be used for many other AD investigations, including small clinical trials, even when datasets and subject availability are limited.

Conclusions and implications

The machine learning techniques have demonstrated to be a suitable tool for studying risk factors in AD, being a potential screening tool for MNCD. The division by blocks (sociodemographic, clinical, and analytical) shows, with the adjustment of machine learning techniques, an important increase in the specificity and sensitivity in the differences between the control and MNCD groups. Moreover, this study has revealed the importance of the inclusion of analytical data as risk factors for the development of MNCD.

ACKNOWLEDGMENTS

We thank the association of relatives of Alzheimer's patients at Gran Canaria and ULPGC (Peritia et Doctrina) for their collaboration and interest in this project, as well as the Insular Hospital of Lanzarote for their involvement. This work was supported by the Canary Islands Government through the ACIISI (*Agencia Canaria de Investigacion, Innovacion y Sociedad de la Información*), ITHaCA project [grant number ProID2017010164]; the Spanish Government through PLATINO project [grant number TEC2017-86722-C4-4-R]. This work was completed while Beatriz Martinez-Vega, Raquel Leon and Samuel Ortega were beneficiary of a pre-doctoral grant given by the ACIISI of the "*Conserjería de Economía, Industria, Comercio y Conocimiento*" of the "*Gobierno de Canarias*", which is part-financed by the European Social Fund (FSE) (*POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)*).

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-0955r1).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-200955.

REFERENCES

[1] European Observatory on Health Systems and Policies (2019) *Spain: Country Health Profile 2019.* https://ec.europa.eu/health/sites/health/files/state/docs/2019_chp_es_english.pdf.

[2] European Union, Ageing Europe - Looking at the lives of older people in the EU. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_looking_at_the_lives_of_older_people_in_the_EU, Last updated 2019.

[3] Mira JJ, Nuño-Solinís R, Fernández-Cano P, Contel JC, Guilabert-Mora M, Solas-Gaspar O (2015) Readiness to tackle chronicity in Spanish health care organisations: A two-year experience with the instrumento de evaluación de modelos de atención ante la cronicidad/assessment of readiness for chronicity in health care organisations instrument. *Int J Integr Care* **15**, e041.

[4] Marcelli S, Gatti C, Rocchi R, Troiani S, Di Tuccio S, Giuli C, Postacchini D, Santarelli A (2017) Chronic care model and cost reduction in initial health: A new approach for satisfaction and improvement of chronicity. *Geriatr Care* **3**, https://doi.org/10.4081/gc.2017.7066.

[5] Niu H, Álvarez-Álvarez I, Guillén-Grima F, Aguinaga-Ontoso I (2017) Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis. *Neurología* **32**, 523-532.

[6] Instituto Nacional de Estadística, Population rate with disabilities diagnosed with certain chronic diseases according to disease, by age and sex. https://www.ine.es/jaxi/Tabla.htm?path=/t15/p418/a2008/hogares/p01/modulo1/l0/&file=02032.px&L=0.

[7] Alzheimer's & Dementia, Alzheimer's Facts and Figures Report. https://www.alz.org/alzheimers-dementia/facts-figures?lang=en-US.

[8] Soto-Gordoa M, Arrospide A, Moreno-Izco F, Martínez-Lage P, Castilla I, Mar J (2015) Projecting burden of dementia in Spain, 2010-2050: Impact of modifying risk factors. *J Alzheimers Dis* **48**, 721-730.

[9] World Health Organization, Dementia: Fact sheets. https://www.who.int/news-room/fact-sheets/detail/dementia.

[10] Mckhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-939.

[11] Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, Meguro K (2007) Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS–ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[12] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 263-269.

[13] Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, Cappa S (2014) Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *Lancet Neurol* **13**, 614-629.

[14] Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Budd S, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Luis J, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM, Sperling R, Elliott C, Masliah E, Ryan L, Silverberg N (2018) NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement* **14**, 535-562.

[15] Abu-rumeileh S, Steinacker P, Polischi B, Mammana A, Bartoletti-Stella A, Oeckl P, Baiardi S, Zenesini C, Huss A, Cortelli P, Capellari S, Otto M, Parchi P (2019) CSF biomarkers of neuroinflammation in distinct forms and subtypes of neurodegenerative dementia. *Alzheimers Res Ther* **12**, 2.

[16] Elahi FM, Casaletto KB, La Joie R, Walters SM, Harvey D, Wolf A, Edwards L, Rivera-Contreras W, Karydas A, Cobigo Y, Rosen HJ, DeCarli C, Miller BL, Rabinovici GD, Kramer JH (2020) Plasma biomarkers of astrocytic and neuronal dysfunction in early- and late-onset Alzheimer's disease. *Alzheimers Dement* **16**, 681-695.

[17] Wimo A, Ali G, Wu Y, Prina AM, Winblad B, Linus J, Liu Z, Prince M (2017) The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimers Dement* **13**, 1-7.

[18] López-álvarez J, Agüera-ortiz LF (2015) Nuevos criterios diagnósticos de la demencia y la enfermedad de Alzheimer: Una visión desde la psicogeriatría. *Psicogeriatría* **5**, 3-14.

[19] Jack CR, Wiste HJ, Weigand SD, Therneau TM, Lowe VJ, Knopman DS, Gunter JL, Senjem ML, Jones DT, Kantarci K, Machulda MM, Mielke MM, Roberts RO, Vemuri P, Reyes DA, Petersen RC (2017) Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement* **13**, 205-216.

[20] Sagrario-Manzano M, Fortea J, Villarejo A, Sanchez del Valle R (2018) *Guías diagnósticas y Terapéuticas de la Sociedad Española de neurología*, Madrid.

[21] de Bruijn RFAG, Bos MJ, Portegies MLP, Hofman A, Franco OH, Koudstaal PJ, Ikram MA (2015) The potential for prevention of dementia across two decades: The prospective, population-based Rotterdam Study. *BMC Med* **13**, 1-8.

[22] Mayer F, Di Pucchio A, Lacorte E, Bacigalupo I, Marzolini F, Ferrante G, Minardi V, Masocco M, Canevelli M, Di Fiandra T, Vanacore N (2018) An estimate of attributable cases of Alzheimer disease and vascular dementia due to modifiable risk factors: The impact of primary prevention in Europe and in Italy. *Dement Geriatr Cogn Dis Extra* **8**, 60-71.

[23] Ownby R, Crocco E, Acevedo A, Loewenstein J (2006) Depression and risk for Alzheimer disease: Systematic review, meta-analysis, and metaregression analysis. *Arch Gen Psyquiatry* **63**, 530-538.

[24] Santabárbara J, Sevil-pérez A, Olaya B, Gracia-garcía P, López-antón R (2019) Depresión tardía clínicamente relevante y riesgo de demencia: Revisión sistemática y metaanálisis de estudios prospectivos de cohortes. *Rev Neurol* **68**, 493-502.

[25] Xu H, Yang R, Qi X, Dintica C, Song R, Bennett DA, Xu W (2019) Association of lifespan cognitive reserve indicator with dementia risk in the presence of brain pathologies. *JAMA Neurol* **76**, 1184.

[26] Meng X, D'Arcy C (2012) Education and dementia in the context of the cognitive reserve hypothesis: A systematic review with meta-analyses and qualitative analyses. *PLoS One* **7**, e38268.

[27] Cataldo JK, Prochaska JJ, Glantz SA (2010) Cigarette smoking is a risk factor for Alzheimer's disease: An analysis controlling for tobacco industry affiliation. *J Alzheimers Dis* **19**, 465-480.

[28] Pan P, Shi H, Zhong J, Xiao P, Shen Y, Wu L, Song Y, He G (2013) Chronic smoking and brain gray matter changes: Evidence from meta-analysis of voxel-based morphometry studies. *Neurol Sci* **34**, 813-817.

[29] Peters R, Poulter R, Warner J, Beckett N, Burch L, Bulpitt C (2008) Smoking, dementia and cognitive decline in the elderly, a systematic review. *BMC Geriatr* **8**, 1-7.

[30] Guure CB, Ibrahim NA, Adam MB, Said S (2017) Impact of physical activity on cognitive decline, dementia, and its subtypes: Meta-analysis of prospective studies. *Biomed Res Int* **2017**, 1-13.

[31] Law LL, Rol RN, Schultz SA, Dougherty RJ, Edwards DF, Koscik RL, Gallagher CL, Carlsson CM, Bendlin BB, Zetterberg H, Blennow K, Asthana S, Sager MA, Hermann BP, Johnson SC, Cook DB, Okonkwo OC (2018) Moderate intensity physical activity associates with CSF biomarkers in a cohort at risk for Alzheimer's disease. *Alzheimers Dement (Amst)* **10**, 188-195.

[32] de Frutos-Lucas J, López-Sanz D, Zuluaga P, Rodríguez-Rojo IC, Luna R, López ME, Delgado-Losada ML, Marcos A, Barabash A, López-Higes R, Maestú F, Fernández A (2018) Physical activity effects on the individual alpha peak frequency of older adults with and without genetic risk factors for Alzheimer's Disease: A MEG study. *Clin Neurophysiol* **129**, 1981-1989.

[33] Barnes DE, Yaffe K (2013) The projected impact of risk factor reduction on Alzheimer's disease prevalence. *Lancet Neurol* **10**, 819-828.

[34] Qiu C, Winblad B, Fratiglioni L (2005) The age-dependent relation of blood pressure to cognitive function and dementia.pdf. *Lancet Neurol* **4**, 487-499.

[35] Lu FP, Lin KP, Kuo HK (2009) Diabetes and the risk of multi-system aging phenotypes: A systematic review and meta-analysis. *PLoS One* **4**, e4144.

[36] Ruiz-García A, Arranz-Martínez E, García-Álvarez JC, García-Fernández ME, Palacios-Martínez D, Montero-Costa A, Ciria-de-Pablo C, López-Uriarte B, García-Pliego RA, Chao-Escuer P, Zafra-Urango C, Alcaraz-Bethencourt A, Redondo-de-Pedro S, Escamilla-Guijarro N, Pascual-Val T, Vieira-Pascual MC, Martínez-Irazusta J, Martínez-Cid-de-Rivera E, Rodríguez-de-Cossío Á, De-Prado-Prieto L, Adrián-Sanz M, Minguela-Puras ME, Blanco-Canseco JM, Rubio-Villar M, Berbil-Bautista ML, Hueso-Quesada R, Plata-Barajas MT, Redondo-Sánchez M, Durán-Tejada MR, García-Redondo MR, Sánchez-Herráiz M, Rey-López AM, García-García-Alcañiz MP, Abad-Schilling C, Hidalgo-Calleja Y, Rivera-Teijido M, En representación del Grupo de Investigación del Estudio SIMETAP. Grupo de Investigación del Estudio SIMETAP (2020) Prevalence of diabetes mellitus in Spanish primary care setting and its association with cardiovascular risk factors and cardiovascular diseases. SIMETAP-DM study. *Clin Investig Arterioscler* **32**, 15-26.

[37] Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine learning for medical imaging. *Radiographics* **37**, 505-515.

[38] Carpenter KA, Huang X (2018) Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: A review. *Curr Pharm Des* **24**, 3347-3358.

[39] Fisher CK, Smith AM, Walsh JR, Simon AJ, Edgar C, Jack CR, Holtzman D, Russell D, Hill D, Grosset D, Wood F, Vanderstichele H, Morris J, Blennow K, Marek K, Shaw LM, Albert M, Weiner M, Fox N, Aisen P, Cole PE, Petersen R, Sherer T, Kubick W (2019) Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci Rep* **9**, 1-14.

[40] Ahmadzadeh M, Christie GJ, Cosco TD, Moreno S (2020) Neuroimaging and analytical methods for studying the pathways from mild cognitive impairment to Alzheimer's disease: Protocol for a rapid systematic review. *Syst Rev* **9**, 4-9.

[41] Gupta Y, Lama RK, Kwon GR (2019) Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers. *Front Comput Neurosci* **13**, 1-18.

[42] Kim JP, Kim J, Park YH, Park SB, Lee JS, Yoo S, Kim EJ, Kim HJ, Na DL, Brown JA, Lockhart SN, Seo SW, Seong JK (2019) Machine learning based hierarchical classification of frontotemporal dementia and Alzheimer's disease. *Neuroimage Clin* **23**, 101811.

[43] Maj C, Azevedo T, Giansanti V, Borisov O, Dimitri GM, Spasov S, Lió P, Merelli I (2019) Integration of machine learning methods to dissect genetically imputed transcriptomic profiles in Alzheimer's Disease. *Front Genet* **10**, 1-16.

[44] So A, Hooshyar D, Park KW, Lim HS (2017) Early diagnosis of dementia from clinical data by machine learning techniques. *Appl Sci* **7**, 651.

[45] Jammeh EA, Carroll CB, Pearson Stephen W, Escudero J, Anastasiou A, Zhao P, Chenore T, Zajicek J, Ifeachor E (2018) Machine-learning based identification of undiagnosed dementia in primary care: A feasibility study. *BJGP Open* **2**, bjgpopen18X101589.

[46] Boustani M, Perkins AJ, Khandker RK, Duong S, Dexter PR, Lipton R, Black CM, Chandrasekaran V, Solid CA, Monahan P (2020) Passive digital signature for early identification of Alzheimer's disease and related dementia. *J Am Geriatr Soc* **68**, 511-518.

[47] Weakley A, Williams JA, Schmitter-Edgecombe M, Cook DJ (2015) Neuropsychological test selection for cognitive impairment classification: A machine learning approach. *J Clin Exp Neuropsychol* **37**, 899-916.

[48] Aschwanden D, Aichele S, Ghisletta P, Terracciano A, Kliegel M, Sutin AR, Brown J, Allemand M (2020) Predicting cognitive impairment and dementia: A machine learning approach. *J Alzheimers Dis* **75**, 717-728.

[49] Johnson P, Vandewater L, Wilson W, Maruff P, Savage G, Graham P, Macaulay LS, Ellis KA, Szoeke C, Martins RN, Rowe CC, Masters CL, Ames D, Zhang P (2014) Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC Bioinformatics* **15**, 1-14.

[50] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525.

[51] Jain A, Nandakumar K, Ross A (2005) Score normalization in multimodal biometric systems. *Pattern Recognit* **38**, 2270-2285.

[52] Hardy MA (1993) *Regression with Dummy Variables*, SAGE Publications, Inc.

[53] Holder LB, Muksitul Haque M, Skinner MK (2017) Machine learning for epigenetics and future medical applications. *Epigenetics* **12**, 505-514.

[54] Sastry K, Goldberg DE, Kendall G (2005) Genetic algorithms. In *Search Methodologies*, Burke EK, Kendall G, eds. Springer US, Boston, MA, pp. 97-125.