García-Gutiérrez *et al.*
Alzheimer's Research & Therapy (2024) 16:26
https://doi.org/10.1186/s13195-024-01394-y

Alzheimer's
Research & Therapy

RESEARCH

Open Access

Unveiling the sound of the cognitive status: Machine Learning-based speech analysis in the Alzheimer's disease spectrum

Fernando García-Gutiérrez[1], Montserrat Alegret[1,2], Marta Marquié[1,2], Nathalia Muñoz[1], Gemma Ortega[1,2], Amanda Cano[1,2], Itziar De Rojas[1,2], Pablo García-González[1], Clàudia Olivé[1], Raquel Puerta[1], Ainhoa García-Sanchez[1], María Capdevila-Bayo[1], Laura Montrreal[1], Vanesa Pytel[1], Maitee Rosende-Roca[1], Carla Zaldua[3], Peru Gabirondo[3], Lluís Tárraga[1,2], Agustín Ruiz[1,2], Mercè Boada[1,2] and Sergi Valero[1,2*]

Abstract

Background Advancement in screening tools accessible to the general population for the early detection of Alzheimer's disease (AD) and prediction of its progression is essential for achieving timely therapeutic interventions and conducting decentralized clinical trials. This study delves into the application of Machine Learning (ML) techniques by leveraging paralinguistic features extracted directly from a brief spontaneous speech (SS) protocol. We aimed to explore the capability of ML techniques to discriminate between different degrees of cognitive impairment based on SS. Furthermore, for the first time, this study investigates the relationship between paralinguistic features from SS and cognitive function within the AD spectrum.

Methods Physical-acoustic features were extracted from voice recordings of patients evaluated in a memory unit who underwent a SS protocol. We implemented several ML models evaluated via cross-validation to identify individuals without cognitive impairment (subjective cognitive decline, SCD), with mild cognitive impairment (MCI), and with dementia due to AD (ADD). In addition, we established models capable of predicting cognitive domain performance based on a comprehensive neuropsychological battery from Fundació Ace (NBACE) using SS-derived information.

Results The results of this study showed that, based on a paralinguistic analysis of sound, it is possible to identify individuals with ADD (F1 = 0.92) and MCI (F1 = 0.84). Furthermore, our models, based on physical acoustic information, exhibited correlations greater than 0.5 for predicting the cognitive domains of attention, memory, executive functions, language, and visuospatial ability.

Conclusions In this study, we show the potential of a brief and cost-effective SS protocol in distinguishing between different degrees of cognitive impairment and forecasting performance in cognitive domains commonly affected within the AD spectrum. Our results demonstrate a high correspondence with protocols traditionally used to assess cognitive function. Overall, it opens up novel prospects for developing screening tools and remote disease monitoring.

*Correspondence:
Sergi Valero
svalero@fundacioace.org
Full list of author information is available at the end of the article

Keywords Alzheimer's disease, Mild cognitive impairment, Early diagnosis, Neuropsychological tests, Machine Learning, Speech acoustics, Automated pattern recognition

Introduction

The digitization in the healthcare field experienced in the last few years has led to the emergence of new technologies with great potential for Alzheimer's disease (AD) management [1]. This neurodegenerative disease is the most frequent type of dementia worldwide. With no effective treatment yet available, AD poses a significant challenge to the sustainability of existing healthcare systems [2]. As a result, the development of tools capable of detecting the disease in its early stages has become one of the most active research areas [3].

From a neuropsychological standpoint, AD typically manifests with impairments in memory, attention, language, executive and visuospatial functions, and behavior [2]. These alterations worsen as the disease progresses, eventually limiting the individual's independence. At the pathophysiological level, AD is characterized by a cascade of brain-level events, including the accumulation of amyloid-β plaques (Aβ), the formation of hyperphosphorylated tangles of tau protein (p-tau), and neuroinflammation [2, 4]. These neuropathological signatures ultimately lead to atrophy, decreased brain metabolism, and disruptions in brain connectivity, which cause the observed cognitive alterations [3–5].

In this context, it is well known that the entire landscape of neurological events in AD initiates several years before the onset of clinical symptoms and progresses silently until the first cognitive changes emerge [6]. Consequently, numerous diagnostic criteria have been proposed, incorporating information from biomarkers such as Aβ and p-tau, measured by positron emission tomography (PET) or detected on cerebrospinal fluid (CSF), along with evidence of neurodegeneration assessed using magnetic resonance imaging (MRI) [4]. However, the detection of biomarkers through neuroimaging or lumbar puncture requires specialized equipment and trained personnel, rendering these procedures expensive and inaccessible for most healthcare centers. Furthermore, most patients consult a memory unit when their cognitive impairment is already evident and only a minority when the initial neuropsychological symptoms arise [2]. Collectively, these factors indicate that while these techniques have significant value for disease diagnosis, their use as population screening tools to identify individuals at risk of developing AD is limited.

Conversely, neuropsychological batteries have traditionally served as the initial assessment protocol for suspected cognitive impairment associated with AD [7, 8]. Nevertheless, these batteries still rely on the presence of clinicians in a specialized memory unit, making it a time-consuming procedure. Recognizing these limitations, numerous online abbreviated protocols have been proposed [1, 9]. Among these, neuropsychological assessment through the spontaneous speech (SS) analysis represents one of the most promising approaches [10]. Previous research supports the presence of language deficits in AD patients several years before the progression to the dementia stage, rendering it a valuable tool for detecting individuals in the early stages of mild cognitive impairment (MCI) [11, 12]. Additionally, language abilities exhibit associations with other cognitive domains such as memory, attention, and executive functions, suggesting that SS analysis has the potential to offer an approximate representation of an individual's cognitive performance [12].

Moreover, the increasing interest in speech analysis using voice recordings stems from the versatility and abundance of information that can be extracted from this type of data. By employing modern natural language processing (NLP) techniques applied to automatic transcriptions [13] or based on the information derived from the raw waveform [14], it is possible to obtain information of great interest for assessing the cognitive performance of an individual. Within this wealth of information, the physical-acoustic features drawn directly from the raw sound represent an agnostic, standardized, and widely available resource [15, 16]. These features encompass parameters such as formants, pitch, and prosody, which are affected in numerous neurodegenerative diseases and affective disorders [15].

To date, most studies analyzing SS using Machine Learning (ML) techniques have focused on developing diagnostic tools to differentiate clinical phenotypes within the AD continuum. For instance, different research groups have shown that it is possible to obtain accuracies ranging from 80 to 90% for discriminating between healthy controls (HC) and AD dementia (ADD) using speech-derived information [17–22]. Other authors have extended their efforts beyond distinguishing ADD and HC, aiming to identify individuals in the early stages of MCI. In this context, results vary considerably based on the technique or sample used, with accuracies from 65 to 80% [18, 21, 23–25]. In contrast, very few authors have focused on addressing the relationship between SS and other aspects of the disease [26, 27]. For example, the investigation of the

connection between SS and neuropsychological impairment has received limited attention [10, 28]. This relevant aspect is not widely available since demonstrating that SS can be a reliable proxy for an individual's cognition requires its evaluation against standardized neuropsychological measures. However, most research has only utilized the Mini-Mental State Examination (MMSE) to explore the association of SS with cognitive status [29–31], typically within the context of the ADReSS challenge [32]. Moreover, the sample sizes employed, usually consisting of a few hundred subjects, severely restrict the ability to draw conclusive findings regarding the expected performance of the models [28].

The present study aims to extend previous research investigating how information obtained from a paralinguistic analysis of various SS tests can differentiate among clinical phenotypes and predict cognitive performance. For this purpose, firstly, we applied ML techniques to differentiate individuals with preserved cognition (subjective cognitive decline, SCD), patients with MCI, and those with ADD. Secondly, we developed models to predict changes in cognitive performance based on SS over the neuropsychological domains of memory, attention, visuospatial and executive functions, and language. As input variables for the models, we focused on information extracted from SS using standardized physical acoustic features obtained from the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [15]. Compared to previous works, our study utilizes a significantly larger population from a real-world setting, comprising SCD individuals, and patients with MCI, and ADD. In addition, as a novelty, our research delves into the connection between speech and changes in cognitive domain performance across the AD spectrum using ML and physical-acoustic features.

Methods

Study participants

This study comprised 1500 individuals who underwent evaluation at the Memory Clinic of Ace Alzheimer Center Barcelona (Ace) (single site) between March 2022 and April 2023. The participants were referred to the Memory Clinic either by their General Health practitioner due to subjective cognitive complaints, or they attended the Open House Initiative without a previous referral from a physician [33]. All subjects completed neurological, neuropsychological, and social evaluations at Ace. The cognitive assessment included the Spanish version of the Mini-Mental State Examination (MMSE) [34], the memory test of the Spanish version of the 7 Minute screening neurocognitive battery [35], the short form of the Neuropsychiatric Inventory Questionnaire (NPI-Q) [36], the Hachinski's Ischemia Scale [37], the Blessed Dementia Rating Scale [38], the Clinical Dementia Rating (CDR) [39], and the complete Neuropsychological Battery of Fundació Ace (NBACE) [7]. The final diagnosis for each participant was determined through consensus by a multidisciplinary team, including neurologists, neuropsychologists, and social workers, at a consensus diagnostic conference [40].

The 1500 participants included 135 individuals with SCD (CDR = 0) with no objective functional or cognitive impairments [41], 826 patients with MCI (CDR = 0.5) [42], and 539 with ADD (CDR > 0.5) [43]. Within the subjects with ADD, 398 had mild dementia (CDR = 1), and 141 had moderate dementia (CDR = 2). Table 1 summarizes the clinical and sociodemographic characteristics of the sample used for this study.

Ethical considerations

This study and its informed consent were approved by the ethics committees of the Hospital Universitari de Bellvitge (Barcelona) (ref. PR007/22) under Spanish biomedical laws (Law 14/2007, 3 July, regarding biomedical research; Royal Decree 1716/2011, 18 November) and followed the recommendations of the Declaration of Helsinki. All participants signed an informed consent for the SS protocol.

Acquisition and preprocessing of speech data

The SS protocol was carried out using the acceXible platform on a tablet. The assessments were conducted in Spanish within a calm and controlled environment. Initially, the participants were presented The Cookie Theft Picture and were instructed to provide a comprehensive description of the image [44]. Subsequently, they were given one minute to name as many different animals as possible. The participants' voices were recorded during the administration of these two tests, and the collected data was utilized for further analyses. The average duration of the protocol was 109.07 ± 15.54 s, with

Table 1 Clinical and sociodemographic variables of the sample used for this study

Variable	All sample	SCD	MCI	ADD
Sample size	1500	135	826	539
Age (mean (SD))	76.17 (9.41)	67.06 (10.42)	74.40 (9.01)	81.15 (6.67)
Sex (% female)	63.47	63.70	62.23	65.31
Years of formal education (mean (SD))	8.58 (4.56)	12.30 (3.79)	8.61 (4.42)	7.62 (4.48)
MMSE (mean (SD))	25.18 (3.94)	29.33 (0.88)	26.85 (2.60)	21.62 (3.36)

Abbreviations: *SCD*, subjective cognitive decline; *MCI*, mild cognitive impairment; *ADD*, Alzheimer's disease dementia; *MMSE*, Mini-Mental State Examination; *SD*, standard deviation

a minimum and maximum duration of 74.6 and 158.9 s respectively.

The audio recordings were standardized to a frequency of 16 KHz, removing initial and final silent portions and applying the noise reduction model presented in [45] to eliminate potential environmental artifacts. The resulting audio data were manually reviewed. Then, features were extracted from the standardized feature set eGeMAPS [15] using the Python interface of the open-source toolkit OpenSMile [46]. The set of features from the eGeMAPS are oriented to provide a simplified and standardized selection of relevant acoustic parameters for detecting physiological changes in voice production guided by findings of previous related studies [15]. Based on prior research [19, 20], we adopted the default configuration provided by the OpenSMILE library [15, 46]. The variables were calculated from the eighteen low level descriptors using a symmetric moving average filter of three frames long. For a more comprehensive understanding of the computation of the paralinguistic variables, readers are directed to [15]. In total 176 variables were extracted from the voice recordings.

Calculation of cognitive composites

Neuropsychological composites created from the NBACE battery were used to determine the cognitive status of participants. Composite scores are a widely employed approach to capture common factors of variance across different neuropsychological tests. Their purpose is to simplify the information by eliminating redundancy between variables offering a more comprehensive characterization of the cognitive domain being assessed [47]. In this study, five cognitive domains, typically examined in AD [48, 49], were considered: memory, attention, visuospatial functions, executive functions, and language. Structural equation models (SEM) were used to calculate the composite scores based on the neuropsychological structure described in [7], and defined by an expert panel of neuropsychologists.

Briefly, the SEM framework defines a measurement model in which observed items $\boldsymbol{y} \in \mathbb{R}^i$ (e.g., i neuropsychological tests) are determined by unobservable factors $\boldsymbol{\eta} \in \mathbb{R}^p$ (e.g., p higher cognitive functions) according to $\boldsymbol{y} = \boldsymbol{\upsilon} + \Delta \cdot \boldsymbol{\eta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\upsilon} \in \mathbb{R}^i$ corresponds to the intercepts of the regression paths, $\Delta \in \mathbb{R}^{i \times p}$ to the measurement slopes, and $\boldsymbol{\epsilon} \in \mathbb{R}^i$ to the residual error. Additionally, the measurement model is subjected to a structural part that defines the relationship between observable and latent variables $\boldsymbol{\eta} = \alpha + B \cdot \boldsymbol{\eta} + \boldsymbol{\zeta}$, where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a parameter vector, $B \in \mathbb{R}^{p \times p}$ is a non-singular matrix with $\text{diag}(B) = 0$ indicating the relationship between latent variables, and $\boldsymbol{\zeta} \in \mathbb{R}^p$ represents the latent variable residuals.

In the present study, the model parameters were adjusted using the robust maximum likelihood estimator, and the variances of the latent variables were fixed to 1 for model identification (unit variance method [47]). The model coefficients were estimated considering the baseline evaluations of the entire Ace database (N = 23,987)——including individuals HC/SCD, MCI, and ADD——using the R package lavaan [50]. The composite scores were adjusted for age, sex, and years of formal education effects using linear regression models. Further details on the composites and their calculation can be found in the Appendix A.

The memory composite was created considering the variables long-term and recognition memory of the Word List subtest from the Wechsler Memory Scale, third version (WMS-III) [51]. The attention composite included the Digit Forward and Digit Backwards from the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) [52]. To define the visuospatial functions, the 15-Objects Test [53], the Poppelreuter-type overlap figures [54], and the Luria's Clock test [55] were considered. The executive functions were calculated from the Phonetic and Semantic Verbal fluencies [56, 57] and the Automatic Inhibition subtest of the Syndrom Kurtz Test (SKT) [58]. Finally, language function included the abbreviated 15-item naming test from the Boston Naming Test (BNT) [59] and the Verbal Comprehension and Repetitions [7].

Machine Learning modeling

In the present study, two different problems were addressed. Firstly, classification models were developed to differentiate between clinical phenotypes. Secondly, regression models were implemented to predict the cognitive composites outlined in the "Calculation of cognitive composites" section. The models used in each problem are described below.

For the classification problems, the following models were used: random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM), and k-nearest neighbor (KNN). Due to the high dimensionality of the input data, the SVM and KNN algorithms were combined with a previous feature selection step. The feature selection aims to identify the optimal combination of variables by eliminating those that are irrelevant/redundant. We performed feature selection using a wrapper-based approach involving two sequential stages: candidate subset generation (SG) and subset evaluation (SE) [60]. For the SG component, we utilized genetic algorithms (GA), a population-based metaheuristic optimization strategy inspired by the natural selection process [61]. For the SE, we considered the mean balanced accuracy ($\text{BA} = 0.5 \cdot [\text{sensitivity} + \text{specificity}]$) obtained through

fivefold cross-validation (CV) from predictions derived from a Gaussian naive Bayes classifier. Feature selection was not applied for the RF and XGB models because these types of algorithms, not based on distance like SVM or KNN, are more robust to the high dimensionality.

Moreover, the hyperparameters for RF, XGB, and SVM models were determined using a hyperparameter optimization (HPO) framework. A nested ten-fold CV was applied to the training set to perform the HPO. The HPO was implemented using the Optuna open-source library [62], applying a Bayesian optimization (BO) using a tree-structured Parzen estimator (TPE) as a surrogate model [63]. Since the KNN model had a small number of hyperparameters, they were optimized using a grid search. In addition, given the high-class imbalance in the classification problems, the KNN was combined with the synthetic minority over-sampling technique (SMOTE) applied to the minority class [64].

For the regression tasks, the same models adapted for predicting quantitative variables (RF, XGB, GA-SVM, and GA-KNN) were used. In this case, the SE step of the feature selection strategy used in the GA-SVM and GA-KNN was performed using a KNN regressor minimizing the mean squared error (MSE). For the BO-TPE, the MSE evaluated by a nested tenfold CV applied on the training set was minimized. The optimized hyperparameters for each model and the configuration details for BO-TPE and the GAs are provided in Appendix B.

All models were implemented using Python (v3.9.16). The scikit-learn library was utilized for RF, KNN, and SVM algorithms [65]. The XGB package [66] was employed for the XGB models. Finally, for the GAs, the home-made library pywinEA2 available on GitHub was used.

Experimental setup

Among the main objectives of this study was to differentiate clinical phenotypes. For this purpose, the following problems were addressed: differentiation of individuals with a preserved cognitive state (SCD) from those with cognitive impairment (MCI and ADD), discrimination between SCD and ADD, classification of MCI and ADD, and the distinction between SCD and MCI. For the prediction of the cognitive composites, models were fitted on each of the five composites mentioned in the "Calculation of cognitive composites" section.

All models were evaluated using a ten-fold CV. Performance metrics were reported as the mean values obtained on the test set. The HPO and feature selection techniques, as described in the "Machine Learning modeling" section, were implemented using a nested CV approach on the training set to prevent overfitting. For classification tasks, CV was conducted with class stratification. Figure 1 illustrated the training and model validation pipeline applied for all the algorithms.

The following performance metrics were considered for the classification problems:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,, \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}\,, \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}\,, \tag{3}$$

$$\text{Balanced accuracy (BA)} = \frac{\text{Sensitivity} + \text{Specificity}}{2}\,, \tag{4}$$

$$\text{F1-score (F1)} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}\,, \tag{5}$$

where TP and TN represent true positives and negatives and FN and FP stand for false positives and negatives. For the regression problems, the correlation coefficient between model predictions ($\hat{Y} \in \mathbb{R}^n$) and true values ($Y \in \mathbb{R}^n$) was considered, as well as the following metrics:

$$\text{Mean absolute error (MAE)} = \frac{1}{n} \cdot \sum_{i=1}^{n} |Y_i - \hat{Y}_i|\,, \tag{6}$$

$$\text{Explained variance (EV)} = 1 - \frac{\text{Var}[\, Y - \hat{Y}\,]}{\text{Var}[\, Y\,]}\,, \tag{7}$$

$$\text{Relative MEA (RMEA)} = \frac{\text{MAE}}{|P^{v}_{95\%} - P^{v}_{5\%}|}\,, \tag{8}$$

with $\text{Var}[\,\cdot\,]$ being the variance and $|P^{v}_{95\%} - P^{v}_{5\%}|$ representing the range of the variable v between the 95% and 5% percentiles. Consequently, the RMEA allows contextualizing the magnitude of the MEA in relation to the magnitude of the analyzed variable.

As input variables for all models, the 176 variables extracted from the voice recordings (the "Acquisition and preprocessing of speech data" section) and sociodemographic variables including age, years of education, and sex were considered. For the distance-based models such as SVM and KNN, the variables were standardized to z-scores based on the statistics of the training data.

All experiments were conducted on the Ace computing's cluster, composed of 368 CPU cores and 1280 GB of RAM running on Rocky Linux OS (v8.6).

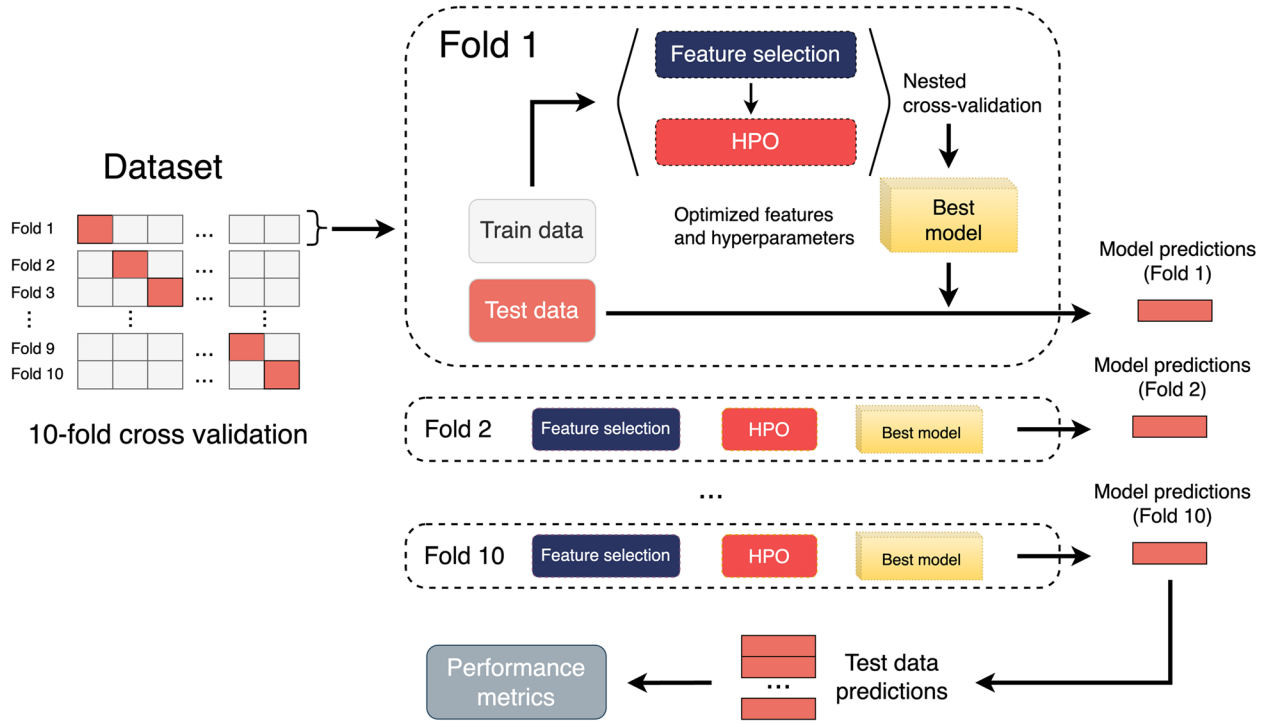

Fig. 1 Pipeline applied to train and evaluate the different models used in this study. The input data were divided into ten folds according to a cross-validation (CV) scheme. Feature selection and hyperparameter optimization (HPO) were conducted using a nested CV applied to the training data. The resulting model from this nested CV was then used to make the predictions on the test set. The final performance metrics were calculated based on the average performance of the predictions obtained on the test set

Results

Cognitive composites analysis

As mentioned in the "Calculation of cognitive composites" section, five cognitive composites were derived from the neuropsychological variables of the NBACE [7]. These composites encompassed the cognitive domains of attention, executive functions, language, memory, and visuospatial functions. The SEM used for their estimation showed good fit indices supporting the consistency of the proposed factorial structure (comparative fit index = 0.93, Tucker-Lewis index = 0.92, mean square error of approximation = 0.07) [47].

Figure 2 illustrates the composite values based on clinical diagnosis. Using linear regression models, significant differences in all composite scores across different diagnostic groups were found (Bonferroni adjusted p-value < 0.05) (refer to Appendix C for further details). Across all cognitive domains, as expected, SCD subjects exhibited the highest composite scores, while individuals with MCI displayed intermediate ones, and the ADD group showed the lowest composite values.

Table 2 presents the discriminatory performance achieved using a logistic regression model for each composite. The mean values obtained on the test set from ten repetitions of ten-fold CV are displayed. As expected, all composites exhibited a strong discriminatory ability between SCD and ADD individuals. They also offered a clear, although lower, differentiation of SCD and MCI subjects. Notably, the composites of executive functions, language, memory, and visuospatial functions provided the best discrimination between SCD and ADD individuals (AUC > 0.98). The attention composite showed a lower predictive performance (AUC ≈ 0.95). In the detection of MCI individuals, the values were slightly lower. The executive function composite showed the highest predictive capacity (AUC ≈ 0.93). The visuospatial, language, and memory composite scores also demonstrated good discriminatory ability (AUC > 0.90). The attention composite showed the poorest performance (AUC ≈ 0.84).

Spontaneous speech for differentiating clinical phenotypes

The results of the top-performing models for distinguishing clinical phenotypes are presented in Table 3. Due to significant imbalances between classes, the models with the highest F1-score value were selected. The receiver operating characteristic (ROC) curves

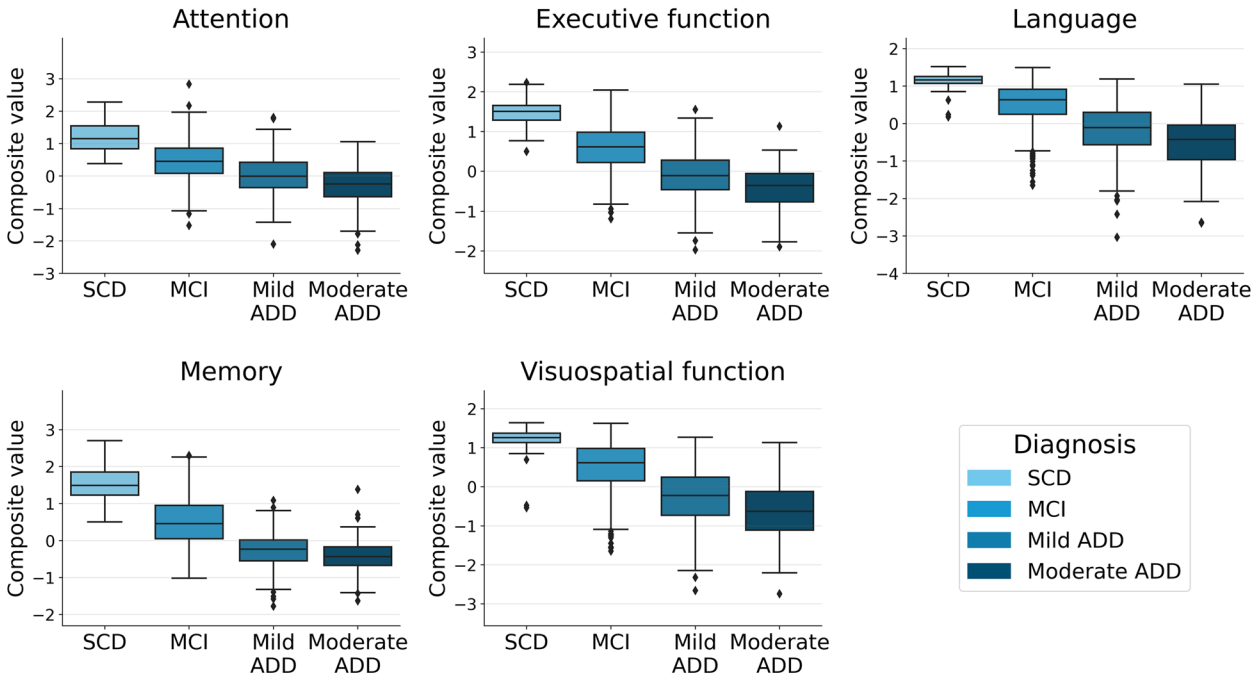

Fig. 2 Cognitive composite values were obtained through structural equation models (SEM) and adjusted for age, sex, and years of education. The scores were represented based on the diagnostic group. Abbreviations: SCD, subjective cognitive decline; MCI, mild cognitive impairment; ADD, Alzheimer's disease dementia

Table 2 Discriminatory capacity of cognitive composites for differentiating clinical phenotypes using subjects with subjective cognitive decline (SCD) as reference

Composite	Accuracy		Sensitivity		Specificity		AUC	
	ADD	MCI	ADD	MCI	ADD	MCI	ADD	MCI
Attention	0.87	0.74	0.86	0.74	0.89	0.75	0.95	0.84
Executive function	0.97	0.84	0.97	0.83	0.98	0.91	0.99	0.93
Language	0.96	0.81	0.95	0.79	0.98	0.91	0.99	0.91
Memory	0.98	0.81	0.98	0.80	0.97	0.87	0.99	0.90
Visuospatial function	0.95	0.79	0.95	0.77	0.98	0.92	0.99	0.91

Mean value obtained over the test set from ten repetitions of tenfold cross-validation. A logistic regression model was employed to predict clinical phenotypes. The composite scores were considered as independent variables

Abbreviations: *AUC*, area under the curve; *MCI*, mild impairment; *ADD*, Alzheimer's disease dementia

for each problem, incorporating all algorithms, are depicted in Fig. 3. For the models that integrated GAs, the selected variables are detailed in the Supplementary material. Appendix E provides a summary of the features that were consistently chosen by the GAs during the CV.

Regarding the algorithms implemented, the tree-based models showed the best performance for most of the problems. Specifically, the RF obtained the highest F1 for the differentiation between SCD and MCI/ADD, the XGB achieved the best results in SCD-ADD and MCI-ADD problems, and the GA-SVM model outperformed the rest of the algorithms for distinguishing SCD and MCIs. The GA-KNN combination consistently demonstrated the poorest performance across all the problems (refer to Appendix D for further details).

When distinguishing between SCD and subjects with cognitive impairment (MCI/ADD), an F1-score of 0.85 was achieved. In this comparison, sensitivity and specificity were close to 0.75. The performance notably improved when differentiating between SCD and ADD, reaching an F1-score of 0.92, with a sensitivity of approximately 0.90 and specificity of around 0.80. In contrast, in all cases involving the identification of MCI subjects,

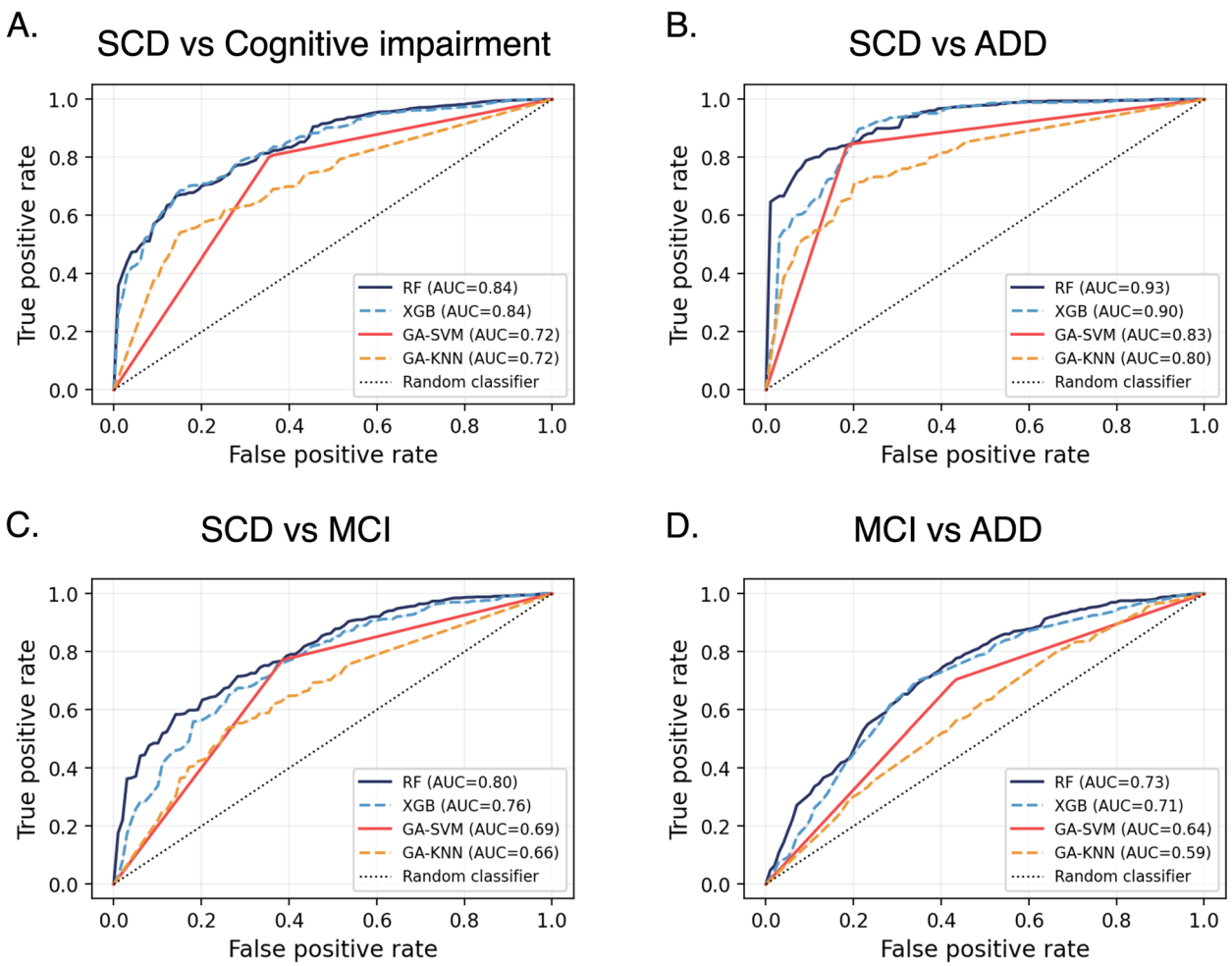

Fig. 3 The average values of the receiver operating characteristic (ROC) curves, obtained for the models discussed in the "Machine learning modeling" section, are presented. The ROC curves were calculated on the test set from a ten-fold cross-validation. The analyzed classification tasks encompassed: **A** discrimination between subjective cognitive decline (SCD) and patients with mild cognitive impairment (MCI) or Alzheimer's disease dementia ADD, **B** SCD vs ADD, **C** SCD vs MCI, **D** and MCI vs ADD. Abbreviations: RF, random forest; XGB, extreme gradient boosting; GA, genetic algorithm; SVM, support vector machine; and KNN, k-nearest neighbor

Table 3 Average values of the classification metrics computed on the test set for the models achieving the highest F1-score in each of the problems

Problem	Best model	BA	F1	Precision	Sensitivity	Specificity
SCD vs cognitive impairment	RF	0.75	0.85	0.97	0.76	0.73
SCD vs ADD	XGB	0.84	0.92	0.94	0.90	0.79
SCD vs MCI	GA-SVM	0.69	0.84	0.93	0.77	0.62
MCI vs ADD	XGB	0.67	0.63	0.56	0.72	0.63

Abbreviations: *BA*, balanced accuracy; *F1*, f1-score; *RF*, random forest; *XGB*, extreme gradient boosting; *GA-SVM*, genetic algorithm-support vector machine; *SCD*, subjective cognitive decline; *MCI*, mild cognitive impairment; *ADD*, Alzheimer's disease dementia

the performance decreased. When discerning between MCI and SCD, a specificity of 0.77 was reached, but at the expense of a high rate of false positives (sensitivity of 0.62). Moreover, the specificity for distinguishing ADD from MCI decreased to 0.71 while maintaining a low sensitivity (< 0.65).

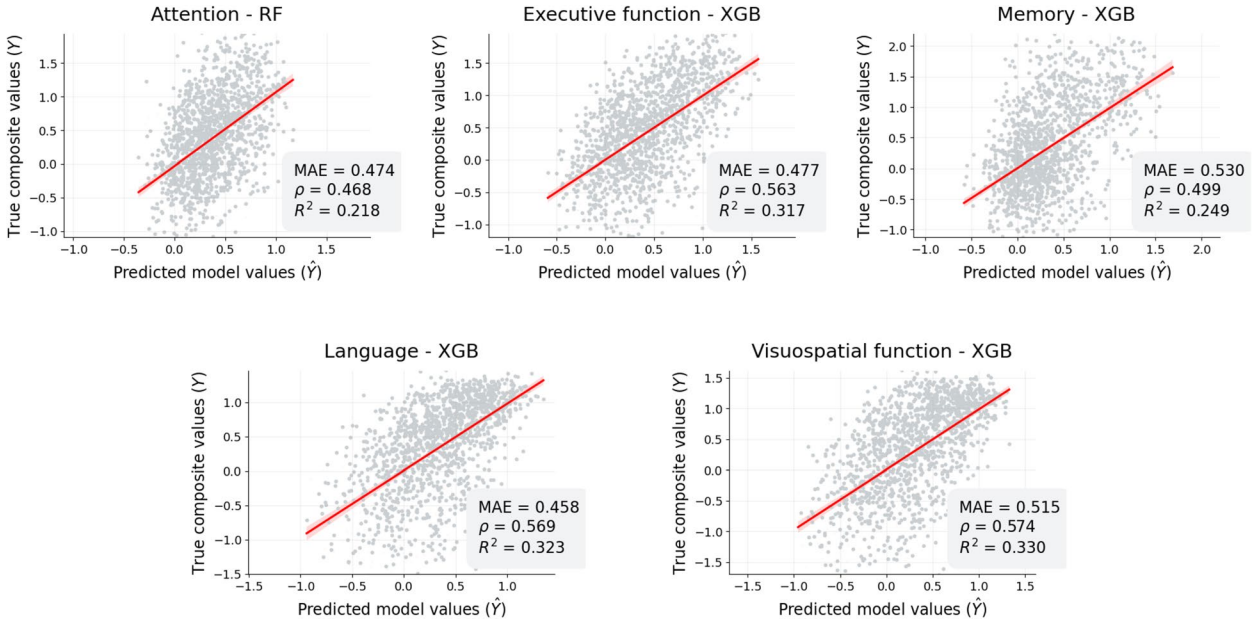

Fig. 4 Correlation between the predicted values generated by the models and the actual values for each cognitive domain under examination. The predictions from the model with the lowest mean absolute error (MAE) were represented. The values depicted in the figure correspond to the predictions made on the test set from a ten-fold cross-validation. To minimize the influence of outliers in the representation, the scales of both the *X* and *Y* axes were adjusted considering the 95th percentile of the true values. Abbreviations: ρ, correlation between model predictions ($\hat{Y}$) and true values (Y); R^2, coefficient of determination; RF, random forest; XGB, extreme gradient boosting

Finally, we conducted a subanalysis excluding sociodemographic information and fitting the top-performing models based purely on physical-acoustic variables. In this scenario, we observed a decline in performance compared to the previous models. Specifically, for distinguishing SCD from individuals with cognitive impairment, the F1-score dropped to 0.80 (from 0.85). In the discrimination between SCD and ADD, the F1-score decreased to 0.81 (from 0.92). For identifying SCD and MCI, the F1-score was similar (0.78 vs 0.77). Finally, in the classification of MCI and ADD, the F1-score moved to 0.55 (from 0.63).

Spontaneous speech for predicting cognitive domains

The predictions given by the best regression models used for estimating the five cognitive scores outlined in the "Calculation of cognitive composites" section, are collected in Fig. 4. Alternatively, the different regression metrics described in the "Experimental setup" section are presented in Table 4. The values of the former table were stratified by clinical phenotype.

Overall, the tree-based models exhibited superior performance. The RF model achieved the lowest MAE in predicting the attention score, while the XGB outperformed the other algorithms for predicting the remaining cognitive domains. Detailed results of all the models can be found in Appendix D. Furthermore, the variables selected by the GAs for the SVM and KNN models are listed in the Supplementary material and Appendix E.

A consistent correspondence between the model predictions and the actual values for each cognitive domain was observed. The language composite regression exhibited the best predictive performance, with a correlation coefficient of 0.57, an EV of 32.3%, and an RMAE of 17.8%. Similarly, for the executive and visuospatial functions, comparable results were achieved, with correlations above 0.56, EVs greater than 31.0% and RMAEs below 18.0%. The models also performed well for the memory and attention composites, although the correspondence between the predictions and the actual values decreased slightly (correlation < 0.5). When stratified by clinical diagnosis, the models performed better, with a fewer MAE, in subjects with MCI across all cognitive domains. The predictions for the SCD group showed the highest errors in attention, executive function, and memory composites. In contrast, the ADD group exhibited the highest error in the composites of language and visuospatial functions. Figure 5 illustrates the prediction distributions generated by the models stratified by the diagnostic group. The regression models predicted higher values in all the composites for the SCD, lower values for the MCIs, and a reduced estimate for the ADD individuals.

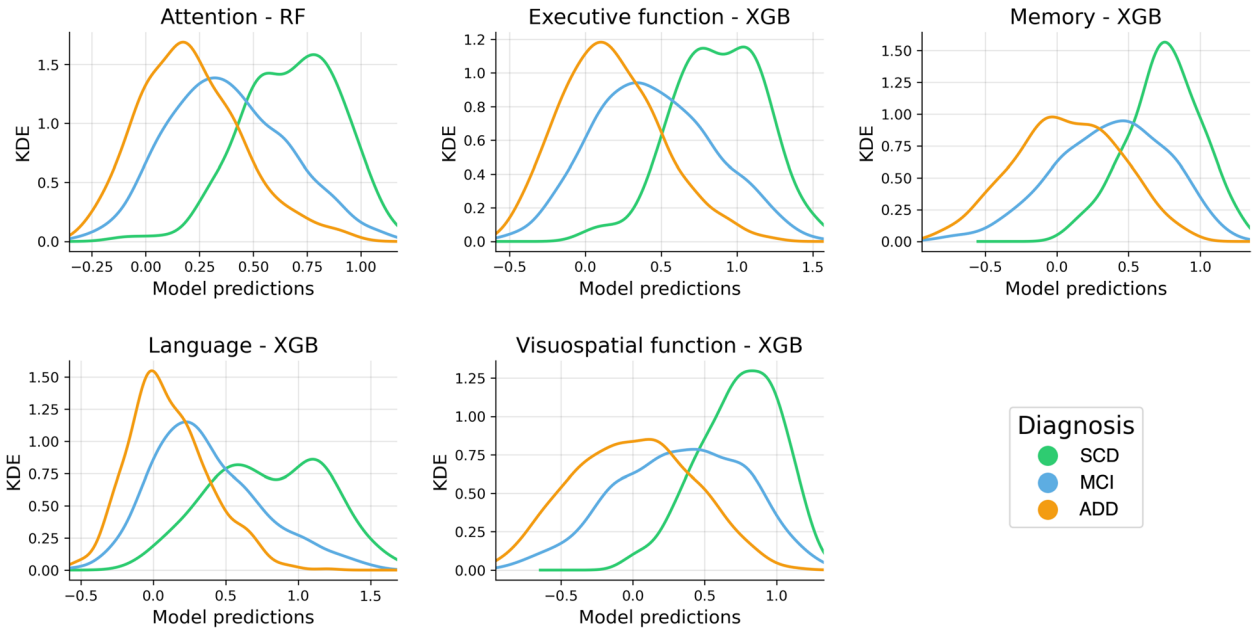

Fig. 5 Distribution of predictions given by the models stratified by clinical diagnosis for each of the cognitive functions. The *Y*-axis represents a kernel density estimation (KDE) of the model predictions distribution (*X*-axis). Abbreviations: SCD, subjective cognitive decline; MCI, mild cognitive impairment; ADD, Alzheimer's disease dementia; RF, random forest; XGB, extreme gradient boosting

Table 4 Regression metrics obtained by the best models in the prediction of each cognitive domain

Sample	Metric	Attention	Executive function	Language	Memory	Visuospatial function
All	RMAE (%)	18.007	17.421	17.801	18.330	17.880
	MAE	0.474	0.477	0.458	0.530	0.515
	Correlation	0.468	0.563	0.569	0.499	0.574
	EV[a] (%)	21.778	31.696	32.345	24.924	32.964
SCD	RMAE (%)	22.063	23.295	16.736	26.588	18.401
	MAE	0.580	0.638	0.431	0.769	0.530
	Correlation	0.062	0.082	0.009	0.076	0.002
	EV[a] (%)	–	–	–	–	–
MCI	RMAE (%)	16.282	15.339	15.573	17.023	16.125
	MAE	0.428	0.420	0.401	0.492	0.464
	Correlation	0.370	0.477	0.487	0.383	0.485
	EV[a] (%)	12.494	17.759	17.559	10.321	18.114
ADD	RMAE (%)	19.634	19.141	21.483	18.265	20.440
	MAE	0.517	0.524	0.553	0.528	0.589
	Correlation	0.243	0.205	0.330	-0 036	0 347
	EV[a] (%)	3.668	–	6.192	–	6.560

For each cognitive domain, the regression metrics obtained by the best models are presented. The random forest (RF) was the best model at predicting the attention score, while the extreme gradient boosting (XGB) performed better on all other scores. The metrics were calculated for the entire sample and stratified by clinical phenotype

Abbreviations: *MAE*, mean absolute error; *RMAE*, relative MAE (described in the "Experimental setup" section); *EV*, explained variance; *SCD*, subjective cognitive decline; *MCI*, mild cognitive impairment; *ADD*, Alzheimer's disease dementia

[a] The EV is not shown when the variance of the true values was lower than the variance of the residuals

Similar to the classification problems (see the "Spontaneous speech for differentiating clinical phenotypes" section), we investigated the influence of demographic variables on the best models. For this purpose, we eliminated the variables of age, years of formal education, and sex from the models. In this context, we observed a marginal decline in their performance, while still maintaining correlations between predicted and true values that closely mirrored those obtained when demographic information was integrated. Specifically, for the attention composite, the correlation and EV were at 0.43 and 18.3%, respectively. Concerning executive function, the correlation and EV values shifted to 0.51 and 26.0%. The performance scores for the memory composite declined to 0.42 and 18.0%. For the language composite, the new values were 0.50 and 25.3%. Finally, for the visuospatial function, the correlation and EV dropped to 0.51 and 26.3% respectively.

Discussion

The present study shows that the automated analysis of a brief SS test using ML techniques consistently detects cognitive alterations along the AD spectrum. To the best of our knowledge, this is the first study with a large sample from a clinical setting exploring the association between SS and cognitive performance across neuropsychological domains.

Firstly, several ML models were applied to distinguish clinical phenotypes associated with AD. Our models focused on differentiating between individuals diagnosed with SCD, MCI, and ADD. The results demonstrated a good differentiation between SCD individuals from those with already manifest cognitive impairment (MCI/ADD) (AUC = 0.84) as well as between SCD and ADD patients (AUC = 0.93) (Table 3). These findings are consistent with previous studies [17–22, 67–71]. For example, in [20], they achieved an accuracy of 80.28% on the test set for detecting subjects with dementia using information derived from *The Cookie Theft Picture* description task. In the same line, in [17], they reached an AUC of 0.86 for discriminating between HC and ADD subjects. Similarly, the authors of [21] obtained an AUC of 0.93 for distinguishing HC individuals from those with dementia and an AUC of 0.88 for differentiating dementia and non-dementia subjects. Therefore, our findings provide new evidence supporting the potential of SS for the identification of individuals with ADD.

In contrast, the models exhibited a moderate predictive performance for identifying individuals with MCI (MCI vs SCD: AUC = 0.80, MCI vs ADD: AUC = 0.73). These findings are consistent with the results previously reported in the literature [18, 21, 23–25]. In their study, [18] identified MCI subjects with a BA of 0.65, slightly lower than the results presented in this work (BA = 0.69, see Table 3). With subtly better performance, in [23], they used linguistic features extracted from transcripts, reporting a specificity and sensitivity of 78% and 74%, respectively. However, it is important to note that their study was supported by a considerably smaller sample size than ours, limiting conclusions about the model's accuracy. Consistent with the findings of this work, in [21], they observed that the identification of subjects with MCI was notably more challenging compared to detecting dementia, obtaining an AUC of 0.74. Overall, the difficulties in identifying MCI stem from its inherent heterogeneity, encompassing different subtypes and potential underlying etiologies. Furthermore, the cognitive deficits in this population are subtle and can be influenced by factors such as age, education level, and individual cognitive abilities, often masking the underlying MCI status. Consequently, the boundary between normal cognition and MCI remains inherently fuzzy, limiting the performance of the models [42]. Future studies should strive to overcome these limitations and enhance the ability to identify the cognitive changes that emerge during the MCI stage. Nevertheless, our results, based exclusively on the physical-acoustic properties of the sound, demonstrate that the application of Artificial Intelligence (AI) techniques to SS data offers valuable insights for identifying patients with MCI.

This study also investigated the potential to estimate cognitive performance using the paralinguistic variables derived from SS tests. For this purpose, neuropsychological tests from a standardized battery of neuropsychological measures were grouped into five composite scores representing the cognitive domains of attention, executive functions, language, memory, and visuospatial functions [7]. These neurocognitive composites were the target variables of the ML models. As outlined in the "Cognitive composites analysis" section, the resulting composites discriminate between the different diagnostic groups. This discriminatory ability was expected, as the neuropsychological tests forming the composites are partially instrumental in defining the diagnoses. Nevertheless, these results show that their grouping into cognitive domains is consistent and effectively summarizes the information from the individual neuropsychological tests into higher cognitive functions. Therefore, this subanalysis supports the use of these neuropsychological groupings to describe the different cognitive functions analyzed.

The algorithms used to infer the neurocognitive composites based on the SS tests and physical-acoustic features showed a strong predictive ability (Fig. 4). In general, we observed that the models were proficient in predicting the different cognitive scores with a correlation close to 0.5 relative to the actual values (Table 4). Nevertheless, our findings indicated a significant increase in model errors within the SCD and ADD groups. For example, the correlation between model predictions and actual values in SCD subjects remained poor, likely due to the ceiling effect present in many of the neuropsychological tests used to construct the composites or an overrepresentation of MCI subjects within the sample. However, despite this, we observed a meaningful alignment of the model predictions with the different disease stages, associating higher values for SCD subjects, declining scores in the MCI stage, and lower estimates for ADD patients (Fig. 5). This result becomes especially relevant for the remote detection of cognitive impairment in the general population, as the SS test can be completed within an average duration of 110 s, and the correspondence between model predictions and disease stages is consistent. In addition, the increased predictive ability obtained for MCI patients (see Table 4) is particularly significant, as this stage holds great importance for developing screening tools, recruiting patients for clinical trials, and monitoring disease progression [2].

Regarding the models used, our results also provided insightful findings. First, we noted that tree-based algorithms, specifically RF and XGB, consistently surpassed distance-based algorithms (i.e., SVM and KNN) across tasks. We hypothesize that this superiority can be attributed to the high dimensionality of the input data and the robust nature of tree-based algorithms in handling non-smooth distributions [66, 72]. Moreover, despite incorporating a prior feature selection step to mitigate the sensitivity of distance-based algorithms to the high dimensionality, the GAs consistently selected a high number of variables. This elevated number of input features is presumably responsible for the poorer performance achieved by these models. It suggests that the effectiveness of feature selection techniques may not scale optimally with the increasing number of input variables [60].

On the other hand, we observed that excluding demographic variables from the models harmed their performance. These findings underscore the potential advantages of incorporating contextual information about the patient in SS and AI-based screening tools, enabling the algorithms to uncover nonlinear interactions among input variables [19, 21]. For instance, it is reasonable to expect that a specific response pattern on an SS test has distinct implications for an older non-literate individual compared to a young person with a high level of education. However, additional investigation is needed to evaluate the standalone predictive capacity of SS and the consequences of integrating contextual information into the models. In this regard, special attention should be directed towards variables that can be readily and efficiently collected through a remotely administered protocol, such as self-reported information on comorbidities or a family history of neurodegenerative diseases.

Our study shows that an automated analysis of speech based on paralinguistic features and ML techniques has the potential for detecting and assessing AD stages. Compared to other modalities, such as neuroimaging or plasma biomarkers, SS-based protocols represent a fast, cost-effective, and accessible tool for evaluating the patient's cognitive status despite lower diagnostic accuracy [10, 73]. The SS protocol is non-invasive, does not require expensive equipment or highly trained personnel, and is a patient-friendly procedure. Furthermore, beyond the implementation of SS as an early screening tool, periodic assessment of SS could provide valuable insights into the decline in different cognitive domains on a simple and rapid basis. Overall, it opens up new opportunities for implementing novel and widely accessible to the general population screening strategies.

From a methodological standpoint, our study benefits from using standardized SS tests and paralinguistic features, facilitating the replication of our findings. Moreover, unlike previous studies, we developed models capable of inferring cognitive performance from SS data using a large sample of subjects at different disease stages. In addition, we employed optimized and transparently evaluated ML models, providing detailed fit indices that enable easy comparison with other studies. Collectively, our work presents a promising avenue for leveraging automated speech analysis in AD, offering potential benefits for the early detection and monitoring of cognitive decline.

Nevertheless, this study has certain limitations. Firstly, despite using a large sample of subjects, especially compared with most current studies [17, 19, 20, 23, 68, 71], the generalization of our results should be confirmed by future prospective analyses involving larger samples and individuals from different cohorts and in different languages. This recurrent issue is commonly encountered in studies employing ML techniques and represents

a significant challenge for potential translation to the clinical setting. Secondy, our study was conducted using data generated in a controlled clinical environment, ensuring the acquisition of high-quality audio data. However, the remote administration of the SS protocol is expected to introduce noise and other factors that may potentially impact the performance of the models. Consequently, the evaluation of the effectiveness of the proposed approach with remotely generated data will be an aspect of interest to be explored in future studies. Moreover, our study relies solely on a restricted set of paralinguistic features. As demonstrated by other researchers [18, 19, 21], using more diverse data derived from speech analysis, such as employing NLP techniques [13], could substantially enhance the performance of predictive models. For future investigations, it may be worthwhile to consider including broader and more diverse SS parameters to improve the performance established in this study. Finally, this work has a cross-sectional design due to the restricted access to follow-up information. Future research endeavors will be necessary to explore how longitudinal analysis of SS data can provide relevant insights into predicting aspects related to the AD continuum, and differentiate ADD from other types of dementia.

Conclusion

In conclusion, the convergence of Artificial Intelligence advancements and the rapid digitization witnessed in recent years has set the stage for developing new technologies capable of monitoring AD in a simple and increasingly accessible manner. Among these innovative technologies, SS protocols, such as the one employed in this study, stand out. These technical breakthroughs hold great potential in enabling early and precise detection of cognitive changes within the AD continuum, ultimately facilitating remote access to specialists and personalized therapies. Our study provides new evidence to the field, demonstrating the feasibility of inferring cognitively impaired performance across different cognitive functions from SS data, establishing a solid foundation for future developments in predictive models.

Appendix A. Calculation of cognitive composites

Machine Learning (ML) models were developed for the prediction of neuropsychological composites. This appendix details the calculation of the composite scores generated from the Neuropsychological Battery from Fundació Ace (NBACE) battery [7].

As mentioned in the main manuscript, the memory composite was created considering the variables long-term and recognition memory of the Word List subtest from the Wechsler Memory Scale, third version (WMS-III) [51]. The attention composite included the Digit Forward and Digit Backwards from the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) [52]. To define the visuospatial functions, the 15-Objects Test [53], the Poppelreuter-type overlap figures [54], and the Luria's Clock test [55] were considered. The executive functions were calculated from the Phonetic and Semantic Verbal fluencies [56, 57] and the Automatic Inhibition subtest of the Syndrom Kurtz Test (SKT) [58]. Finally, language function included the abbreviated 15-item naming test from the Boston Naming Test (BNT) [59] and the Verbal Comprehension and Repetitions [7]. The subtest measuring the time taken to complete the SKT was transformed to a logarithmic scale for structural equation models (SEM) fitting.

All available baseline observations collected in Ace were used for the composite scores calculation. Healthy controls (HC) and individuals with subjective cognitive decline (SCD) with a diagnosis of preserved cognition (CDR = 0), patients with mild cognitive impairment (MCI) (CDR = 0.5), and subjects with dementia due to Alzheimer's disease (ADD) (CDR $\geq$ 1) were selected. Table 5 lists the demographic characteristics of the subjects used to calculate the composite scores. The syntax of the structural equation model used to calculate the composite scores is given in Listing 1. The model fitted with the entire Ace database was subsequently used to infer the composite scores for the sample of subjects used for the study.

Table 5 Clinical and sociodemographic characteristics of the sample used to estimate the parameters of the structural equation models used to calculate the composite scores

	All sample	HC/SCD	MCI	ADD
Sex (% females)	67.09	69.73	62.79	71.04
Age (mean (SD))	75.54 (9.93)	64.51 (9.66)	73.85 (9.23)	80.65 (7.05)
Years of formal education (mean (SD))	7.28 (4.60)	11.08 (4.32)	7.54 (4.42)	5.88 (4.18)
MMSE (mean (SD))	24.00 (4.72)	29.13 (1.25)	26.21 (2.94)	20.08 (3.85)

Abbreviations: *SD*, standard deviation; *MMSE*, Mini-Mental State Examination

```
# neuropsychological tests nomenclature
# M_WMS1 -> WMS-III learning trial 1
# M_WMS_total -> WMS-III learning (sum of all trials)
# M_f_rec -> WMS-III recognition task (false recognitions)
# M_recon -> WMS-III recognition task (total score)
# M_ret -> WMS-III delayed recall
# M_digtspan_direct -> WAIS-III Digit Span Forward
# M_digtspan_invers -> WAIS-III Digit Span Backwards
# G_15obj -> The 15-Objects test (correct answers)
# G_15obj_nf -> The 15-Objects test (errors)
# G_pop_total -> Two Poppelreuter-type overlap figures
#                (correct answers)
# G_Pop_Neoformes -> Two Poppelreuter-type overlap figures
#                (errors)
# G_Luria -> Luria's Clock test
# FE_anflu -> Semantic verbal fluency
# FE_R_abstracte -> WAIS-III Similarities
# FE_SKTtemps -> SKT (time)
# FE_SKTerrors -> SKT (errors)
# LL_Namingtotal -> 15-BNT free evoked correct answers
# LL_comprensio -> Verbal comprehension
# LL_R_total -> Language repetitions
# P_Ecopraxiatotal -> Imitation praxis
# P_ideo_total -> Ideomotor praxis
# P_constr_total -> WAIS-III constructional praxis

# measurement model
memory =~ NA*M_WMS_total + M_f_rec + M_recon + M_WMS1 + M_ret
attworkmem =~ NA*M_digtspan_direct + M_digtspan_invers
visuospatial =~ NA*G_15obj_nf + G_pop_total + G_Pop_Neoformes
                + G_15obj + G_Luria
execf =~ NA*FE_anflu + FE_R_abstracte + FE_SKTtemps
        + FE_SKTerrors + FE_Pflu
language =~ NA*LL_Namingtotal + LL_comprensio + LL_R_total
praxis =~ NA*P_Ecopraxiatotal + P_ideo_total + P_constr_total

# residual correlations
FE_SKTtemps ~~ FE_SKTtemps
M_WMS_total ~~ M_WMS1
M_f_rec ~~ M_recon
G_pop_total ~~ G_Pop_Neoformes
G_15obj_nf ~~ G_15obj

# setting the variances of the latent factors to 1
attworkmem ~~ 1*attworkmem
execf ~~ 1*execf
language ~~ 1*language
memory ~~ 1*memory
praxis ~~ 1*praxis
visuospatial ~~ 1*visuospatial
```

Listing 1 Syntax (according to the laavan library) used to calculate composite scores

Appendix B. Machine Learning model hyperparameters

This appendix details the hyperparameters of all the ML models used in the main manuscript. Table 6 lists the hyperparameters of the algorithms used for the classification and regression problems. The library with the implementation of the genetic algorithm (GA) used is available on GitHub. Table 7 summarizes the hyperparameters of the GA. The Scikit-learn [65] library was used for models: random forest, support vector machines, and k-nearest neighbors. For the XGBoost algorithm, the xgboost [66] library was employed. The hyperparameter optimization (HPO) was carried out using the optuna [62] library. The definition of the hyperparameters listed in the table can be found in the documentation of the respective libraries.

Table 6 Hyperparameters of classification and regression models

Model	Parameter[a]	HPO[b]	Value/search space[c]
Random forest	Number of estimators	-	200
	Class weight[d]	-	Balanced
	Max depth	TPE	{2, . . ., 10}
	Min samples split	TPE	{2, . . ., 40}
	Min samples leaft	TPE	{2, . . ., 30}
	Max samples	TPE	[0.5, 1.0]
	Max features	TPE	[0.5, 1.0]
XGBoost	Number of estimators	-	200
	Max depth	TPE	{2, . . ., 10}
	Learning rate	TPE	[0.01, 0.3]
	Gamma	TPE	[0.0, 100.0]
	Min child weight	TPE	[0.0, 100.0]
	Subsample	TPE	[0.2, 1.0]
	colsample_bytree	TPE	[0.2, 1.0]
	colsample_bynode	TPE	[0.2, 1.0]
	L1 regularization	TPE	[0.1, 10.0]
	L2 regularization	TPE	[0.1, 10.0]
	Scale positive weight[d]	TPE	[0.1, 10.0]
Support vector machines	Kernel	-	Polynomial
	C	TPE	[1e-05, 1e02]
	Degree	TPE	{1, . . ., 10}
	Class weight[d]	TPE	[0.05, 0.95]
	$Coef_0$	TPE	[0.0, 10.0]
K-nearest neighbors	Number of neighbors	Grid search	{4, . . ., 30}
	Weights	Grid search	{Uniform, Distance}

For the hyperparameter optimization conducted using tree-structured Parzen estimator (TPE), a total of 1000 configurations were sampled, with the initial 500 being randomly selected

[a] Hyperparameters not listed in this table were selected at their default value

[b] Hyperparameters that were left fixed are specified by "-"

[c] $\{\cdot\}$ indicates sampling of categorical values, while $[\cdot]$ indicates sampling of real values

[d] Parameter only considered for classification problems

Table 7 Hyperparameters of the genetic algorithm used to perform feature selection

Parameter	Value	Description
Generations	1,000	Number of algorithm iterations
Population size	300	Number of candidate solutions (aka individuals) in the population
Individual representation	Binary	The candidate solutions were represented as a binary array where a value of 1 indicated the presence of a feature and a value of 0 its absence
Selection	Tournament selection ($k = 2$)	Stochastically, k individuals are selected from the whole population. From this selection, the individual with the best fitness value is selected for the next generation. This process is repeated until fill established population size
Elitism	30	Individuals who will pass to the next generation without undergoing the selection process
Mutation	Bit-flip (probability = 0.05)	In each generation, each position of the candidate solution inverts its value according to the specified probability
Cross-over	One-point (probability = 0.5)	In each generation, two individuals are selected with a certain probability and their information is combined by splitting the solution by a certain cut-off point. The resulting fragments are used to generate the offspring

Appendix C. Cognitive composites analysis

This appendix collects the results of the regression models used to analyze the composites generated in the studied sample. Table 8 contains the values represented in Fig. 2 of the main manuscript.

Table 8 Regression model results for the study sample, controlling for sex, age, and years of formal education

Composite	Condition	CDR	Coef.[a]	*p*-value	95%CI
Attention	SCD	0.0	1.19	< 0.001	[1.10, 1.29]
	MCI	0.5	− 0.73	< 0.001	[− 0.83, − 0.62]
	Mild ADD	1.0	− 1.15	< 0.001	[− 1.26, − 1.04]
	Moderate ADD	2.0	− 1.47	< 0.001	[− 1.60, − 1.33]
Executive function	SCD	0.0	1.48	< 0.001	[1.39, 1.57]
	MCI	0.5	− 0.91	< 0.001	[− 1.01, − 0.81]
	Mild ADD	1.0	− 1.58	< 0.001	[− 1.69, − 1.48]
	Moderate ADD	2.0	− 1.90	< 0.001	[− 2.03, − 1.77]
Language	SCD	0.0	1.15	< 0.001	[1.05, 1.24]
	MCI	0.5	− 0.63	< 0.001	[− 0.73, − 0.53]
	Mild ADD	1.0	− 1.31	< 0.001	[− 1.42, − 1.20]
	Moderate ADD	2.0	− 1.67	< 0.001	[− 1.81, − 1.54]
Memory	SCD	0.0	1.53	< 0.001	[1.43, 1.62]
	MCI	0.5	− 1.01	< 0.001	[− 1.12, − 0.91]
	Mild ADD	1.0	− 1.78	< 0.001	[− 1.89, − 1.67]
	Moderate ADD	2.0	− 1.95	< 0.001	[− 2.08, − 1.82]
Visuospatial function	SCD	0.0	1.22	< 0.001	[1.12, 1.33]
	MCI	0.5	− 0.73	< 0.001	[− 0.84, − 0.62]
	Mild ADD	1.0	− 1.49	< 0.001	[− 1.61, − 1.37]
	Moderate ADD	2.0	− 1.88	< 0.001	[− 2.02, − 1.73]

Abbreviations: *CDR*, Clinical Dementia Rating; *CI*, confidence interval

[a]Coefficients of the regression models considering the composite value as the dependent variable and the dummy coding of the clinical diagnosis as the independent variable. The models were adjusted for sex, age, and educational level. The coefficients indicate the average value of the composite associated with each group adjusting for the covariates

Appendix D. Machine Learning model results

This appendix summarizes the results obtained by the classification (Table 9) and regression models (Table 10) applied for the differentiation of clinical phenotypes and the prediction of composite cognitive scores, respectively. For each model, the average results obtained on the test set of a ten-fold cross-validation are shown.

Table 9 Average values of the classification metrics computed on the test set for the binary classification problems

Problem	Model	BA	F1	Sen	Spe	Pre
SCD vs altered	RF	0.749	0.854	0.764	0.733	0.967
	XGB	0.739	0.884	0.819	0.659	0.960
	GA-SVM	0.725	0.875	0.805	0.644	0.958
	GA-KNN	0.680	0.742	0.605	0.756	0.961
SCD vs ADD	RF	0.820	0.916	0.900	0.741	0.933
	XGB	0.842	0.920	0.898	0.785	0.944
	GA-SVM	0.830	0.893	0.844	0.815	0.948
	GA-KNN	0.734	0.774	0.661	0.807	0.932
SCD vs MCI	RF	0.692	0.840	0.770	0.615	0.924
	XGB	0.680	0.807	0.715	0.644	0.925
	GA-SVM	0.693	0.841	0.772	0.615	0.925
	GA-KNN	0.628	0.656	0.508	0.748	0.925
MCI vs ADD	RF	0.668	0.614	0.669	0.668	0.569
	XGB	0.670	0.625	0.715	0.625	0.555
	GA-SVM	0.635	0.595	0.704	0.567	0.516
	GA-KNN	0.564	0.532	0.654	0.475	0.449

Abbreviations: *BA*, balanced accuracy; *F1*, f1-score; *RF*, random forest; *XGB*, extreme gradient boosting; *GA-SVM*, genetic algorithm-support vector machine; *GA-KNN*, genetic algorithm-K-nearest neighbors

Table 10 Regression metrics obtained in the prediction of the composite cognitive scores

Composite	Model	RMAE (%)	MAE	Correlation	EV (%)
Attention	RF	18.0	0.474	0.468	21.8
	XGB	18.0	0.475	0.464	21.6
	GA-SVM	18.8	0.495	0.407	14.9
	GA-KNN	19.2	0.506	0.352	11.4
Executive function	RF	17.7	0.486	0.552	30.1
	XGB	17.4	0.477	0.563	31.7
	GA-SVM	18.3	0.502	0.517	26.1
	GA-KNN	19.2	0.527	0.450	19.9
Language	RF	18.1	0.466	0.548	29.9
	XGB	17.8	0.458	0.569	32.3
	GA-SVM	19.7	0.508	0.493	21.4
	GA-KNN	19.9	0.511	0.418	17.1
Memory	RF	18.5	0.534	0.494	24.3
	XGB	18.3	0.530	0.499	24.9
	GA-SVM	19.6	0.568	0.407	14.9
	GA-KNN	20.2	0.584	0.354	10.2
Visuospatial function	RF	18.1	0.521	0.565	31.7
	XGB	17.9	0.515	0.574	33.0
	GA-SVM	20.1	0.578	0.498	22.3
	GA-KNN	20.0	0.577	0.430	18.1

Abbreviations: *MAE*, mean absolute error; *RMAE*, relative MAE (described in the "Experimental setup" section); *EV*, explained variance; *RF*, random forest; *XGB*, extreme gradient boosting; *GA-SVM*, genetic algorithm- support vector machine; *GA-KNN*, genetic algorithm-K-nearest

Appendix E. Features selected by the genetic algorithms

This appendix collects the features selected by the models that underwent a preliminary feature selection step using GAs (i.e., SVM and KNN). Table 11 shows the percentage of times in which each variable was selected in each fold for each problem. To simplify the information, only variables consistently chosen across different folds are included, encompassing those that appeared more than 75% of the time in each of the problems examined in the study. The total number of features selected in each of the folds by each of the algorithms in each problem is provided in a separate Excel document (see *Selected-features-GA.xlsx*).

Table 11 Percentage of times each feature was selected across cross-validation folds for the different problems by the genetic algorithms

Feature	P1	P2	P3	P4	P5	P6	P7	P8	P9
F0	100	100	100	100	100	100	100	100	100
F1	100	100	100	-	100	100	100	100	100
F2	100	100	100	-	100	100	95	100	90
F3	100	100	100	-	100	95	90	-	95
F4	-	-	-	-	100	90	75	95	90
F5	-	100	-	100	-	-	85	80	80
F6	75	100	80	85	-	-	-	-	85
F7	75	-	-	-	-	85	75	95	90
F8	-	-	80	-	85	75	-	90	80
F9	-	-	-	-	80	95	95	-	85
F10	-	-	-	-	-	85	95	90	80
F11	-	-	-	-	75	85	80	-	95
F12	-	-	-	-	90	75	-	75	95
F13	-	-	-	-	-	75	80	90	85
F14	-	75	-	-	-	80	-	85	80
F15	95		100	85	-	-	-	-	-
F16	95	75	100	-	-	-	-	-	-
F17	-	-	-	-	80	-	-	75	90
F18	-	-	-	90	-	-	80	75	-
F19	-	85	-	-	-	-	80	-	75
F20	75	-	-	-	-	75	85	-	-
F21	-	-	-	-	75	-	80	-	75
F22	75	-	-	-	-	80	75	-	-
F23	-	80	-	100	-	-	-	-	-
F24	-	-	-	-	90	-	85	-	-
F25	-	-	-	80	-	-	90	-	-
F26	-	-	-	-	-	-	-	80	90
F27	-	-	-	80	-	-	90	-	-
F28	-	-	-	90	-	-	80	-	-
F29	-	-	-	-	-	75	90	-	-
F30	-	-	-	-	90	-	75	-	-
F31	-	-	-	-	-	-	-	75	90
F32	75	85	-	-	-	-	-	-	-

Feature	P1	P2	P3	P4	P5	P6	P7	P8	P9
F33	-	-	80	80	-	-	-	-	-
F34	-	-	-	-	80	-	-	-	75
F35	80	-	75	-	-	-	-	-	-
F36	75	-	-	-	-	75	-	-	-
F37	-	75	-	-	-	-	75	-	-
F38	75	75	-	-	-	-	-	-	-
F39	-	75	75	-	-	-	-	-	-
F40	-	-	-	-	100	-	-	-	-
F41	-	95	-	-	-	-	-	-	-
F42	-	-	-	-	90	-	-	-	-
F43	-	-	-	-	-	-	-	-	90
F44	-	-	-	-	-	-	90	-	-
F45	-	-	-	-	-	-	85	-	-
F46	85	-	-	-	-	-	-	-	-
F47	-	-	-	-	-	-	-	85	-
F48	-	-	-	-	85	-	-	-	-
F49	-	-	85	-	-	-	-	-	-
F50	-	-	-	-	-	-	80	-	-
F51	-	-	-	-	-	80	-	-	-
F52	-	-	-	-	-		-	80	-
F53	-	-	-	-	-	80	-	-	-
F54	-	-	-	-	-	-	80	-	-
F55	-	-	-	80	-	-	-	-	-
F56	-	-	-	-	-	-	-	-	80
F57	-	-	-	-	80	-	-	-	-
F58	-	-	-	-	80	-	-	-	-
F59	-	-	80	-	-	-	-	-	-
F60	-	-	-	80	-	-	-	-	-
F61	-	-	-	80	-	-	-	-	-
F62	-	-	-	-	80	-	-	-	-
F63	-	-	80	-	-	-	-	-	-
F64	-	-	-	-	-	-	-	-	80
F65	-	-	-	-	-	-	-	-	75
F66	-	-	-	-	-	-	-	75	-
F67	-	-	-	-	-	-	75	-	-
F68	-	-	-	-	-	-	75	-	-
F69	-	-	-	-	-	-	-	75	-
F70	-	-	-	-	-	-	75	-	-
F71	-	-	-	-	-	-	-	75	-
F72	-	-	-	-	75	-	-	-	-
F73	-	-	-	-	-	75	-	-	-
F74	-	-	-	-	-	75	-	-	-
F75	-	-	-	-	75	-	-	-	-
F76	-	-	-	-	75	-	-	-	-
F77	-	-	-	-	75	-	-	-	-
F78	-	-	-	-	75	-	-	-	-
F79	75	-	-	-	-	-	-	-	-
F80	75	-	-	-	-	-	-	-	-
F81	75	-	-	-	-	-	-	-	-
F82	-	75	-	-	-	-	-	-	-
F83	-	-	75	-	-	-	-	-	-
F84	-	-	75	-	-	-	-	-	-
F85	-	75	-	-	-	-	-	-	-
F86	-	-	-	-	-	-	-	-	75

[a] The names of the features are listed in Table 12

Abbreviations: *P1*, SCD vs cognitive impairment; *P2*, SCD vs ADD; *P3*, SCD vs MCI; *P4*, MCI vs ADD; *P5*, attention; *P6*, executive function; *P7*, memory; *P8*, language; *P9*, visuospatial function

Table 12 Abbreviations of the feature names listed in Table 11

Code[a]	Feature name	Code[a]	Feature name
F0	Age	F44	(AE) Frequency F2-bandwidth (voiced) AMean
F1	Years of formal education	F45	(ID) Frequency F0-semitonefrom-27.5hz (voiced) 20-80th percentile range
F2	(AE) Spectral flux (voiced) CoV	F46	(AE) Frequency F3-bandwidth (voiced) CoV
F3	(ID) Energy/Amplitude Loudness CoV	F47	(ID) Frequency F2-amplitudelogrelf0 (voiced) AMean
F4	(ID) Temporal-feature Loudness-Peak/second	F48	(ID) Frequency F1-frequency (voiced) CoV
F5	(AE) Ceptral MFCC4 (voiced) CoV	F49	(ID) Ceptral MFCC4 CoV
F6	(AE) Frequency F0-semitonefrom-27.5hz (voiced) Mean rising slope	F50	(AE) Spectral Alpha-ratio (voiced) CoV
F7	(ID) Energy/Amplitude Shimmer-localdb (voiced) CoV	F51	(ID) Frequency F0-semitonefrom-27.5hz (voiced) Std falling slope
F8	(AE) Frequency F0-semito nefrom-27.5hz (voiced) CoV	F52	(AE) Frequency F3-frequency (voiced) CoV
F9	(AE) Spectral flux (unvoiced) Amean	F53	(AE) Ceptral MFCC3 (voiced) CoV
F10	(AE) Frequency F0-semitonefrom-27.5hz (voiced) 20-80th percentile range	F54	(AE) Spectral Hammarberg index (unvoiced) AMean
F11	(ID) Energy/Amplitude Loudness 20th percentile	F55	(AE) Ceptral MFCC2 CoV
F12	Energy/Amplitude Loudness 50th percentile	F56	Frequency F0-semitonefrom-27.5hz (voiced) 20th percentile
F13	Spectral flux (unvoiced) Amean	F57	Spectral Hammarberg index (unvoiced) AMean

Code[a]	Feature name	Code[a]	Feature name
F14	Spectral Alpharatio (unvoiced) AMean	F58	(AE) Energy/Amplitude Loudness Mean falling slope
F15	(AE) Spectral Slope500-1500 (unvoiced) Amean	F59	(AE) Frequency F0-semitonefrom-27.5hz (voiced) Std rising slope
F16	(AE) Frequency F3-bandwidth (voiced) AMean	F60	(AE) Spectral Slope0-500 (unvoiced) Amean
F17	Temporal-feature Unvoiced-Segment-Length/second Mean	F61	(AE) Ceptral MFCC4 AMean
F18	(AE) Frequency F0-semitonefrom-27.5hz (voiced) Mean falling slope	F62	(AE) Ceptral MFCC1 (voiced) AMean
F19	(AE) Frequency F2-bandwidth (voiced) CoV	F63	(AE) Frequency F2-frequency (voiced) CoV
F20	(AE) Frequency F2-frequency (voiced) AMean	F64	Frequency F2-bandwidth (voiced) AMean
F21	Frequency F0-semitonefrom-27.5hz (voiced) CoV	F65	Frequency F3-amplitudelogrelf0 (voiced) AMean
F22	(AE) Spectral Slope500-1500 (voiced) AMean	F66	Frequency F0-semitonefrom-27.5hz (voiced) Mean rising slope
F23	(AE) Temporal-feature Voiced-Segment-Length/second Mean	F67	Spectral flux (voiced) CoV
F24	Spectral Slope500-1500 (voiced) CoV	F68	(AE) Energy/Amplitude Loudness 20-80th percentile range
F25	(AE) Spectral Alpharatio (unvoiced) AMean	F69	(ID) Spectral flux AMean
F26	(AE) Ceptral MFCC1 CoV	F70	(AE) Ceptral MFCC4 (voiced) AMean
F27	(AE) Spectral Harmonic difference H1-H2 (voiced) CoV	F71	(AE) Energy/Amplitude Loudness Std rising slope
F28	(AE) Frequency F0-semitonefrom-27.5hz (voiced) Std falling slope	F72	Spectral Slope500-1500 (unvoiced) Amean
F29	(AE) Energy/Amplitude Loudness 50th percentile	F73	(AE) Spectral flux AMean
F30	(AE) Energy/Amplitude Loudness Mean rising slope	F74	Energy/Amplitude Shimmer-localdb (voiced) AMean
F31	Frequency F3-frequency (voiced) CoV	F75	(AE) Others Equivalent-Sound-Level (dB)
F32	Ceptral MFCC1 (voiced) CoV	F76	Temporal-feature Voiced-Segments/second
F33	(AE) Frequency Jitter-local (voiced) CoV	F77	(AE) Energy/Amplitude Loudness AMean
F34	Spectral Slope500-1500 (voiced) AMean	F78	Frequency F2-frequency (voiced) AMean
F35	(AE) Spectral Slope0-500 (voiced) CoV	F79	Frequency F3-bandwidth (voiced) AMean
F36	(AE) Temporal-feature Loudness-Peak/second	F80	(ID) Ceptral MFCC3 AMean
F37	Ceptral MFCC1 (voiced) AMean	F81	(AE) Ceptral MFCC1 (voiced) CoV
F38	Ceptral MFCC3 (voiced) AMean	F82	Frequency F0-semitonefrom-27.5hz (voiced) 50th percentile
F39	Ceptral MFCC4 (voiced) CoV	F83	(AE) Frequency F1-frequency (voiced) AMean
F40	Ceptral MFCC4 (voiced) AMean	F84	(AE) Spectral Harmonic difference H1-H2 (voiced) AMean
F41	(AE) Energy/Amplitude Shimmer-localdb (voiced) CoV	F85	(ID) Ceptral MFCC2 CoV
F42	(AE) Energy/Amplitude Loudness Std falling slope	F86	Frequency F0-semitonefrom-27.5hz (voiced) Std rising slope
F43	Frequency F2-frequency (voiced) CoV		

[a] Feature code used to identify the variable in Table 11

Abbreviations: *ID*, image description task; *AE*, animal enumeration task; *MFCC*, mel-frequency cepstral coefficient; *AMean*, arithmetic mean; *CoV*, coefficient of variation; *HNR*, harmonics-to-noise ratio; *Std*, standard deviation

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13195-024-01394-y.

Additional file 1. Supplementary material associated with this article is provided.

Acknowledgements
The authors are grateful to all patients who consented to participate for their significant contribution to making this project possible.

Authors' contributions
SV, FG-G, LT, AR, and MB: conceptualization. SV and FG-G: methodology and formal analysis. SV, FG-G, MA, NM, GO, AC, IR, PG-G, CO, RP, AG-S, MC-B, LM, and AR: data curation. SV, FG-G, MA, MM, NM, VP, MR-R, and AR: writing——review and editing. MM, NM, MA, VP, CZ, PG, GO, and NL: project administration. SV: supervision. All authors contributed to interpretation of the findings, the critical review of the manuscript, and have read and agreed to the published version of the manuscript.

Funding
This project has received funding from R &D Missions in the Artificial Intelligence program, which is part of the Spain Digital 2025 Agenda and the National Artificial Intelligence Strategy and financed by the European Union through Next Generation EU funds (project TARTAGLIA, exp. MIA.2021.M02.0005). This project has also received funding from the Instituto de Salud Carlos III (ISCIII) Acción Estratégica en Salud, integrated in the Spanish National RCDCI Plan and financed by ISCIII Subdirección General de Evaluación and the Fondo Europeo de Desarrollo Regional (FEDER-Una manera de hacer Europa) grant PI19/00335 awarded to M.M., grant PI17/01474 awarded to M.B., grants AC17/00100, PI19/01301 and PI22/01403 awarded to A.R. and by the European Union Joint Programme-Neurodegenerative Disease Research (JPND) Multinational research projects on Personalized Medicine for Neurodegenerative Diseases/Instituto de Salud Carlos III grant AC19/00097 awarded to A.R. and grant FI20/00215 from the Instituto de Salud Carlos III (ISCIII) awarded to I.d.R. For CSF biomarker research, A.R. and M.B. received support from the European Union/EFPIA Innovative Medicines Initiative Joint undertaking ADAPTED and MOPEAD projects (grant numbers 115975 and 115985, respectively). A.C. received support from the Instituto de Salud Carlos III (ISCIII) under the grant Sara Borrell (CD22/00125) and the Spanish Ministry of Science and Innovation, Proyectos de Generación de Conocimiento grant PID2021-122473OA-I00.

Availability of data and materials
The datasets generated and/or analyzed during the current study are not publicly available due to they contain human privacy-sensitive data but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Hospital Universitari de Bellvitge (Barcelona) (ref. PR007/22) on 10 March 2022. Informed consent was obtained from all subjects involved in the study. The informed consent was reviewed and approved by the Ethics committee above mentioned.

Competing interests
The authors declare no competing interests.

Author details
[1]Ace Alzheimer Center Barcelona, Universitat Internacional de Catalunya, Barcelona, Spain. [2]Networking Research Center on Neurodegenerative Diseases (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain. [3]Accexible Impacto s.l., Urduliz, Bizkaia, Spain.

Received: 7 November 2023 Accepted: 18 January 2024
Published online: 02 February 2024

References

1. Mok VC, Pendlebury S, Wong A, Alladi S, Au L, Bath PM, et al. Tackling challenges in care of Alzheimer's disease and other dementias amid the COVID-19 pandemic, now and in the future. Alzheimers Dement. 2020;16(11):1571–81.
2. Alzheimer's & Dementia. 2023 Alzheimer's disease facts and figures. Alzheimers Dement. 2023;19(4):1598–695.
3. Rafii MS, Aisen PS. Detection and treatment of Alzheimer's disease in its preclinical stage. Nat Aging. 2023;3(5):520–31.
4. Scheltens P, De Strooper B, Kivipelto M, Holstege H, Chételat G, Teunissen CE, et al. Alzheimer's disease. Lancet. 2021;397(10284):1577–90.
5. DelEtoile J, Adeli H. Graph theory and brain connectivity in Alzheimer's disease. Neuroscientist. 2017;23(6):616–26.
6. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7(3):280–92.
7. Alegret M, Espinosa A, Vinyes-Junqué G, Valero S, Hernández I, Tárraga L, et al. Normative data of a brief neuropsychological battery for Spanish individuals older than 49. J Clin Exp Neuropsychol. 2012;34(2):209–19.
8. Espinosa A, Alegret M, Valero S, Vinyes-Junqué G, Hernández I, Mauleón A, et al. A longitudinal follow-up of 550 mild cognitive impairment patients: evidence for large conversion to dementia rates and detection of major risk factors involved. J Alzheimers Dis. 2013;34(3):769–80.
9. Alegret M, García-Gutiérrez F, Muñoz N, Espinosa A, Ortega G, Lleonart N, et al. FACEmemory®, an innovative online platform for episodic memory pre-screening: findings from the first 3,000 participants. J Alzheimers Dis. 2024; Pre-press:1–15. https://doi.org/10.3233/JAD-230983.
10. Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. Front Psychol. 2021;12:620251.
11. Sperling RA, Karlawish J, Johnson KA. Preclinical Alzheimer disease-the challenges ahead. Nat Rev Neurol. 2013;9(1):54–8.
12. Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci. 2015;7:195.
13. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: a survey. Sci China Technol Sci. 2020;63(10):1872–97.
14. Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. Appl Acoust. 2020;158:107020.
15. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput. 2015;7(2):190–202.
16. Schuller B, Steidl S, Batliner A, Hirschberg J, Burgoon JK, Baird A, et al. The interspeech 2016 computational paralinguistics challenge: deception, sincerity & native language. In: 17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), ISCA; 2016; Vols 1-5. vol. 8; p. 2001–5. https://doi.org/10.21437/Interspeech.2016-129.
17. Lindsay H, Tröger J, König A. Language impairment in Alzheimer's disease-robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. Front Aging Neurosci. 2021;13:642033.
18. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. Alzheimers Res Ther. 2021;13:1–15.
19. Mahajan P, Baths V. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. Front Aging Neurosci. 2021;13:623607.
20. Chen J, Ye J, Tang F, Zhou J. Automatic detection of alzheimer's disease using spontaneous speech only. Interspeech, 2021;3830–4. https://doi.org/10.21437/interspeech.2021-2002.
21. Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. Alzheimers Dement. 2023;19(3):946–55.
22. He R, Chapin K, Al-Tamimi J, Bel N, Marquié M, Rosende-Roca M, et al. Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains. Am J Speech Lang Pathol. 2023;32(5):2075–86.
23. Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimers Dement Transl Res Clin Interv. 2017;3(2):219–28.
24. Themistocleous C, Eckerström M, Kokkinakis D. Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. Front Neurol. 2018;9:975.
25. Gosztolya G, Balogh R, Imre N, Egas-Lopez JV, Hoffmann I, Vincze V, et al. Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech. Comput Speech Lang. 2021;69:101215.

26. Hajjar I, Okafor M, Choi JD, Moore E, Abrol A, Calhoun VD, et al. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. Alzheimers Dement Diagn Assess Dis Monit. 2023;15(1):e12393.
27. García-Gutiérrez F, Marquié M, Muñoz N, Alegret M, Cano A, De Rojas I, et al. Harnessing acoustic speech parameters to decipher amyloid status in individuals with mild cognitive impairment. Front Neurosci. 2023;17.
28. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alzheimers Res Ther. 2022;14(1):1–16.
29. Liu Z, Proctor L, Collier PN, Zhao X. Automatic diagnosis and prediction of cognitive decline associated with Alzheimer's dementia through spontaneous speech. In: 2021 ieee international conference on signal and image processing applications (icsipa). IEEE; 2021. p. 39–43. https://doi.org/10.1109/ICSIPA52582.2021.9576784.
30. Haulcy R, Glass J. Classifying Alzheimer's disease using audio and text-based representations of speech. Front Psychol. 2021;11:624137.
31. Meghanani A, Anoop C, Ramakrishnan A. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In: 2021 IEEE spoken language technology workshop (SLT). IEEE; 2021. p. 670–7. https://doi.org/10.1109/SLT48900.2021.9383491.
32. Luz S, Haide F, Fuente S. d. l., Fromm D, MacWhinney B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. Proc Interspeech. 2020;2172–6. https://doi.org/10.21437/Interspeech.2021-2002.
33. Abdelnour C, Rodríguez-Gómez O, Alegret M, Valero S, Moreno-Grau S, Sanabria Á, et al. Impact of recruitment methods in subjective cognitive decline. J Alzheimers Dis. 2017;57(2):625–32.
34. Folstein MF, Robins LN, Helzer JE. The mini-mental state examination. Arch Gen Psychiatr. 1983;40(7):812.
35. Del Ser T, Sánchez-Sánchez F, de Yébenes MJG, Otero A, Munoz DG. Validation of the seven-minute screen neurocognitive battery for the diagnosis of dementia in a Spanish population-based sample. Dement Geriatr Cogn Disord. 2006;22(5–6):454–64.
36. Boada M, Tárraga L, Modinos G, López O, Cummings J. Neuropsychiatric inventory-nursing home version (NPI-NH): Spanish validation. Neurologia (Barcelona, Spain). 2005;20(10):665–73.
37. Hachinski VC, Lassen NA, Marshall J. Multi-infarct dementia: a cause of mental deterioration in the elderly. Lancet. 1974;304(7874):207–9.
38. Blessed G, Tomlinson BE, Roth M. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. Brit J Psychiatry. 1968;114(512):797–811.
39. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993;43(11):2412–2412.
40. Boada M, Tárraga L, Hernández I, Valero S, Alegret M, Ruiz A, et al. Design of a comprehensive Alzheimer's disease clinic and research center in Spain to meet critical patient and family needs. Alzheimers Dement. 2014;10(3):409–15.
41. Jessen F, Amariglio RE, Van Boxtel M, Breteler M, Ceccaldi M, Chételat G, et al. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. Alzheimers Dement. 2014;10(6):844–52.
42. Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med. 2004;256(3):183–94.
43. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. 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. 2011;7(3):263–9.
44. Cummings L. Describing the cookie theft picture: sources of breakdown in Alzheimer's dementia. Pragmat Soc. 2019;10(2):153–76.
45. Défossez A, Synnaeve G, Adi Y. Real Time Speech Enhancement in the Waveform Domain. Proc Interspeech. 2020;3291–5. https://doi.org/10.21437/Interspeech.2020-2409.
46. Eyben F, Wöllmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. Assoc Comput Machinery. 2010;4:1459–62. https://doi.org/10.1145/1873951.1874246.
47. Hair JF, Hult GTM, Ringle CM, Sarstedt M, Danks NP, Ray S. An Introduction to Structural Equation Modeling. In: Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R. Cham: Springer; 2021; p. 1–29.
48. Dowling NM, Hermann B, La Rue A, Sager MA. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. Neuropsychology. 2010;24(6):742.
49. Park LQ, Gross AL, McLaren DG, Pa J, Johnson JK, Mitchell M, et al. Confirmatory factor analysis of the ADNI neuropsychological battery. Brain Imaging Behav. 2012;6:528–39.
50. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Softw. 2012;48:1–36.
51. Wechsler D. WMS-III: Wechsler memory scale administration and scoring manual. Psychological Corp; 1997.
52. Wechsler D. Technical manual. Giunti, OS Organizzazioni Speciali; 2002.
53. Pillon B, Dubois B, Bonnet A, Esteguy M, Guimaraes J, Vigouret J, et al. Cognitive slowing in Parkinson's disease fails to respond to levodopa treatment: the 15-objects test. Neurology. 1989;39(6):762.
54. Sala SD, Laiacona M, Trivelli C, Spinnler H. Poppelreuter-Ghent's overlapping figures test: its sensitivity to age, and its clinical use. Arch Clin Neuropsychol. 1995;10(6):511–34.
55. Golden CJ. In reply to Adams's "In search of Luria's battery: A false start.". J Consult Clin Psychol. 1980;48(4):511–6.
56. Artiola L, Hermosillo D, Heaton R, Pardee R. Manual de normas y procedimientos para la batería neuropsicológica en español. Tucson: mPress. 1999.
57. Goodglass H, Kaplan E. The assessment of aphasia and related disorders. Lea & Febiger; 1972.
58. Erzigkeit H. The SKT-a short cognitive performance test as an instrument for the assessment of clinical efficacy of cognition enhancers. In: Diagnosis and treatment of senile dementia. Springer; 1989. p. 164–174.
59. Kaplan E, Goodglass H, Weintraub S, et al. Boston naming test. Encycl Clin Neuropsychol. 2001. https://doi.org/10.1007/978-0-387-79948-3_869.
60. Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends. 2020;1(2):56–70.
61. Xue B, Zhang M, Browne WN, Yao X. A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput. 2015;20(4):606–26.
62. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. Assoc Comput Machinery. 2019. p. 2623–31. https://doi.org/10.1145/3292500.3330701.
63. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst. 2011;24:2546–54.
64. Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863–905.
65. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. 2013. arXiv preprint arXiv:13090238.
66. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. Assoc Comput Machinery. 2016. p. 785–94.
67. Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology. 2000;14(1):71–91.
68. Jarrold W, Peintner B, Wilkins D, Vergryi D, Richey C, Gorno-Tempini ML, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Assoc Comput Linguist. 2014. p. 27–37. https://doi.org/10.3115/v1/W14-3204.
69. Meilán JJG, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? Dement Geriatr Cogn Disord. 2014;37(5–6):327–34.
70. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis. 2016;49(2):407–22.
71. Liu L, Zhao S, Chen H, Wang A. A new machine learning method for identifying Alzheimer's disease. Simul Model Pract Theory. 2020;99:102023.
72. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion. 2022;81:84–90.
73. Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: a review. Expert Syst Appl. 2020;150:113213.

Publisher's Note