Final accepted version. For published version, please see: https://doi.org/10.1044/2023_AJSLP-22-00403

Title page

Title: Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains

Running title: Detect pAD across subsections & language domains

Names of the authors: Rui He[a], Kayla Chapin[a], Jalal Al-Tamimi PhD[b], Núria Bel PhD[a], Marta Marquié MD, PhD[c,d], Maitee Rosende-Roca MD[c], Vanesa Pytel MD, PhD[c], Juan Pablo Tartari MD[c], Montse Alegret PhD[c,d], Angela Sanabria PhD[c,d], Agustín Ruiz MD, PhD[c,d], Mercè Boada MD, PhD[c,d], Sergi Valero PhD[c,d], Wolfram Hinzen PhD[a,e]

Affiliations and addresses of the authors:

[a] Department of Translation & Language Sciences, Universitat Pompeu Fabra, Carrer Roc Boronat, 138, 08018 Barcelona, Spain.

[b] Laboratoire de Linguistique Formelle (LLF), CNRS, Université Paris Cité, Bâtiment Olympe de Gouges, 5ème étage. 8, Rue Albert Einstein 75013 Paris.

[c]Ace Alzheimer Center Barcelona, Universitat Internacional de Catalunya, C/Gran Via de Carles III, 85 bis, 08028 Barcelona, Spain.

[d]Networking Research Center on Neurodegenerative Diseases (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain.

[e]Intitut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, Passeig de Lluís Companys, 23, 08010 Barcelona, Spain.

Correspondence: Rui He, Department of Translation & Language Sciences, Universitat Pompeu Fabra, Carrer Roc Boronat, 138, Barcelona 08018, Spain, rui.he@upf.edu, @RuiHe76864182

Conflict of Interest: The authors declare no conflict of interest.

Author Contributions: Rui He: Conceptualization, Methodology, Investigation, Software, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization, Funding acquisition; **Kayla Chapin**: Conceptualization, Methodology, Investigation, Data curation; **Jalal Al-Tamimi**: Conceptualization, Methodology, Software, Validation, Writing - Review & Editing, Funding acquisition; **Núria Bel**: Methodology, Software, Validation, Writing - Review & Editing; **Marta Marquié**: Clinical assessment; **Maitee Rosende-Roca**: Data curation, Clinical assessment; **Vanesa Pytel**: Clinical assessment; **Juan P. Tartari**: Clinical assessment; **Montse Alegret**: Clinical assessment; **Angela Sanabria**: Clinical assessment; **Agustín Ruiz**: Conceptualization, Investigation, Supervision, Funding acquisition; **Mercè Boada**: Conceptualization, Validation, Supervision, Clinical assessment, Funding acquisition; **Sergi Valero**: Conceptualization, Methodology, Validation, Data curation, Supervision, Project administration; **Wolfram Hinzen**: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Abstract

Background: Decline in language has emerged as a new potential biomarker for the early detection of Alzheimer's disease (AD). It remains unclear how sensitive language measures are across different tasks, language domains, and languages, and to what extent changes can be reliably detected in early stages such as Subjective Cognitive Decline (SCD) and Mild Cognitive Impairment (MCI).

Methods: Using a scene construction task for speech elicitation in a new Spanish/Catalan speaking cohort (n = 119), we automatically extracted features across seven domains, three acoustic (spectral, cepstral, voice quality), one prosodic, and three from text (morpho-lexical, semantic syntactic). They were forwarded to a random forest classifier to evaluate the discriminability of participants with probable AD dementia (pAD), amnestic and non-amnestic MCI, SCD, and cognitively healthy controls. Repeated measure ANOVAs and paired-sample Wilcoxon sign-ranked test were used to assess whether and how performance differs significantly across groups and linguistic domains.

Results: The performance scores of the machine learning classifier were generally satisfactorily high, with the highest scores over .9. Model performance was significantly different for linguistic domains ($p < .001$), and speech vs. text ($p = .043$), with speech features outperforming textual features, and voice quality performing best. High diagnostic classification accuracies were seen even within both cognitively healthy (controls vs. SCD) and MCI (amnestic and non-amnestic) groups.

Conclusions: Speech-based machine learning is powerful in detecting cognitive decline and pAD across a range of different feature domains, though important differences exist between

these domains as well.

Keywords: Machine Learning; probable Alzheimer's Disease; Connected Speech

Introduction

Although Alzheimer's disease (AD) is one of the leading causes of death in older adults, there are no drugs in clinical practice that can cure or prevent the disease (Mossello & Ballini, 2012; World Health Organization, 2020). In this regard and the context of global aging, efficient and accessible approaches to predicting AD risk at earlier stages have been widely sought, including in Mild Cognitive Impairment (MCI) and even preclinically, in individuals with Subjective Cognitive Decline (SCD) (Rabin et al., 2017). Traditional methods for AD detection suffer from a number of limitations, including invasiveness and high cost (e.g., lumbar puncture, neuroimaging markers), or low specificity (e.g., Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR)). In this context, language has emerged as a new and potentially promising biomarker for detecting AD at very early stages, and developing with disease progression (Ahmed et al., 2013; Uretsky et al., 2021). It is noteworthy that people with SCD, by definition, know that they are complaining, and that such awareness may impact on speech parameters, beyond potential organic factors relating to their elevated risk of AD. Automation of speech and language analysis is rapidly advancing, and theoretical models support the integration of language and memory networks in the brain, pointing to shared underlying neural mechanisms (Hagoort, 2019; Roger et al., 2022). The integration of these methods with other signals from behavior and/or with biological markers such as blood may prove essential for advancing on inexpensive, widely available, and robust markers of early disease progression in AD.

Automated approaches from natural language processing (NLP) and machine learning have already shown strong capability in predicting AD, and even MCI (de la Fuente Garcia et al., 2020). Paralinguistic measures extracted from speech directly, such as mathematical properties of sound waves, have also shown impressive power in identifying AD (Chen et al., 2021; Haider et al., 2020; Sarawgi et al., 2020). Some studies transcribed audio into texts, either by hand or machine, and extracted features from texts, including lexical, syntactic, and N-gram features (Orimaye et al., 2017; Qiao et al., 2021; C. Thomas et al., 2005). Unlike these traditional feature engineering approaches, transfer learning based on pretrained language models, such as bidirectional encoder representations from transformers (BERT), encodes linguistic information from large corpora into vector representations or word embeddings. These have been proven powerful in language modeling and some studies have suggested to use them for AD detection due to excellent performance in a binary AD vs. control comparison based on the ADReSS dataset (Balagopalan et al., 2021; Jawahar et al., 2019). Roshanzamir et al. (2021) obtained an accuracy of 88.08% with BERT as an encoder with a logistic regression classifier, in a classification of AD vs. controls with English data from the Pitt corpus. Using Hungarian data, Gosztolya et al. (2019) achieved 74%–82% accuracy in classifying diagnostic groups (cognitively healthy, MCI, and AD) based on speech (acoustic) features, and similar results using language features.

Despite a number of promising studies, several challenges need to be addressed before speech- and language-based classification can be utilized in clinical applications. First, existing studies have mainly used the dataset from the InterSpeech challenge and its source corpus (Pitt Corpus) (Luz et al., 2020, 2021). These datasets only comprise English data from control and AD groups, so lacks data from disease stages in between, especially prodromal ones. Validation of the technically most advanced classifiers across different languages and the entire AD continuum is necessary. Second, current studies have investigated different linguistic levels with generic linguistic variables in a bottom-top fashion (Chapin et al., 2022), yet have not assessed in detail the differential feature performance across domains. To the best of our knowledge, only some studies have compared features from both speech and text, and they have not engaged in fine-grained comparisons of domains within these modalities, especially for the textual one, which has been represented chiefly by pretrained models (Balagopalan et al., 2020; Cummins et al., 2020; Zhu et al., 2021).

Finally, most available studies have elicited connected speech through a picture description, using the Cookie Theft picture (Goodglass et al., 1972). This simple but efficient task has significantly contributed to insights on AD detection through connected speech, but received some criticism including the invocation of stereotypes of family life, elicitation of overly simplified speech, and limited recollection and engagement (Berube et al., 2019; Clarke et al., 2021; Sherratt & Bryan, 2019). Beyond these, the Cookie Theft picture does not challenge the creative and imaginative use of language as speech is generated while looking at the picture and objects are visually available for naming. As a hallmark of language is referencing objects that are not visually present, tasks without visual prompts may show more sensitivity than picture descriptions by challenging language production in one of its core features, namely displaced reference. In particular, scene constructions (SC) have been proposed to stimulate speech for richer information and better discrimination (Irish et al., 2015). SC requires a mental time travel to another location (e.g., a tropical island) and the imaginative representation of directly experienced events at this location. As this information has an episodic and first-personal character, the task taps into a cognitive process widely reported to be impaired in AD

(Hassabis & Maguire, 2007; Schacter et al., 2017; Thakral et al., 2020).

In the present study we aimed (i) to test for the generalizability of previous results of automatic classification of AD using machine learning from English to Spanish/Catalan in a new dataset; (ii) to assess the performance of the classifier across a comprehensive number of language-related featural domains, and (iii) to include different MCI and preclinical groups at risk of AD in the classification.

Methods

Dataset

We recruited 119 participants at the Memory Clinic from Ace Alzheimer Center Barcelona. All of them were native Spanish and/or Catalan speakers. The referral center ethics committee (Hospital Clínic i Provincial de Barcelona) approved the patient recruitment. Collection protocols were under ethical standards according to the World Medical Association Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. Participants were diagnosed as cognitively healthy older controls (HOC), SCD, non-amnestic MCI (naMCI), amnestic MCI (aMCI), and probable AD dementia (pAD). Briefly, SCD refers to the self-perception of cognitive problems, including memory loss, without impairment on the standardized cognitive test (Jessen et al., 2014), while MCI implies that one or more cognitive domains are impaired on the standardized cognitive test but activities of daily living are preserved, according to Petersen's criteria (Petersen, 2004). Most SCD participants were part of the FACEHBI study (Rodriguez-Gomez et al., 2017). Supplementary Information-A (SI-A, similar below) specifies details about recruitment, diagnostic criteria, neuropsychological assessment, and other issues. Table 1 shows the demographic and neuropsychological data of the sample (data: median (interquartile range, IQR)).

Speech data elicitation and processing

Speech data were elicited through a scene construction task adapted from previous studies (Hassabis et al., 2007; Irish et al., 2015). Participants were instructed to construct a scene they imagined to witness and describe it with as much detail as possible, choosing one prompt from three options: "You are lying on the beach in a tropical bay."; "You are in a house that has been abandoned for many years."; and "You are in a circus tent". Most participants chose the first prompt. Speech samples were cut off after three minutes and transcribed into texts manually by a single researcher. The instructions for interview transcripts are shown in SI-B.

We extracted linguistic features from multiple domains, four directly from the audios and three from the transcripts (see Table 2). An overarching aim in feature selection was comprehensiveness, in the sense that we wanted to comparatively assess all major levels of organization of language/speech, including automatically extractable (i) acoustic spectral coefficients, (ii) acoustic cepstral coefficients, (iii) voice quality, (iv) prosodic and (v) morpho-lexical features (the latter at the interface of the lexicon and morpho-syntax), (vi) manually extractable syntactic features, and finally (vii) semantic ones (insofar as we can approximate the latter with current NLP-technologies). In all of the cases (i) to (vii), there is some evidence from previous literature to assume that the features might be discriminative.

In particular, we used openSMILE 3.0 to extract spectral, cepstral, voice quality, and part of prosodic features from the CompareE 2016 feature set at the level of functionals (Eyben et al., 2010; Schuller et al., 2016). Spectral and cepstral coefficients reveal the mathematical properties of the sound waves, while voice quality features involve both phonatory and resonatory characteristics. These three features process speech as sound waves and have been widely used in previous studies, demonstrating powerful detection capacities and high robustness (Haider et al., 2020; Nasrolahzadeh et al., 2016; Thomas et al., 2020). Prosodic features reflect suprasegmental patterns in how a speaker combines individual sounds into integrated sequences with stress and intonation. Speech prosody has proved to be a sensitive measure to cognitive decline for AD even at early onset (Lofgren & Hinzen, 2022; Pistono et al., 2016) and can contribute to accurate automated classification (Themistocleous et al., 2018). As the prosodic profile in CompareE 2016 is not complete, we used Prosogram 3.0.1 to extract more prosodic information, such as pitch variation and pitch stylization (Mertens, 2004). These acoustic-prosodic measures are extracted in a fully automated form, not requiring human transcription and annotation, and they are time-efficient (less than one minute to get all features for each audio) manner, giving them special practical significance.

Syntactic features were manually annotated following the method of Chapin et al. (2022), which targeted specific forms of syntactic complexity involved in referencing objects and events. This feature set, though showing a close relation to neurodegeneration in AD, is not automatable yet. Thus, we added the morpho-lexical features, which have often been used to approximate syntax for measuring language changes in AD, especially in the context of NLP (verb inflection in Fyndanis et al., 2011; ratios of different word classes in Guinn et al., 2014; verb aspect in Manouilidou et al., 2020; verb voice in Nasiri et al., 2022). Morpho-lexical features include the ratios of different word classes and the morphological variants of the content words, which are automatically extracted by the Spanish and Catalan models from Stanza 1.3.0. Lexical and morphological features were found to be important in the progression of AD, in both contexts of group comparison (Kavé & Levy, 2003) and machine learning (Eyigoz et al., 2020). In addition, morphological variants in inflectional languages like Spanish and Catalan provide paths to observe grammar at word level (e.g. aspect and modality). Semantic changes are thought to be prominent in AD, and large computational language models capture distributional aspects of language as a proxy of meaning. A number of studies has found the application of language models on AD detection to be satisfying (Agbavor & Liang, 2022; Balagopalan et al., 2021). The robust optimized version of BERT (RoBERTa) was chosen for encoding the human transcripts. Due to the limitation on token numbers, we truncated the text and only encoded the first 510 tokens, if the text length was longer than that (Liu et al., 2019). We used Catalan BERTa (RoBERTa-base) for Catalan data (https://huggingface.co/projecte-aina/roberta-base-ca-cased-tc) and Byte-Pair Encoding RoBERTa for Spanish data (https://huggingface.co/PlanTL-GOB-ES/roberta-base-bn) (Armengol-Estapé et al., 2021; Gutiérrez-Fandiño et al., 2022). As RoBERTa models return an embedding for every token in the text, which does not fit with the purpose of assigning a single label to the text, we utilized the pooled output as a semantic representation. This output is the embedding of the initial classification token that arises from the sequence output with contextual information from all tokens of the sequence embedded in it. It is standardly used for classification tasks. The complete list of feature sets is available in SI-C. In addition, we also concatenated all the above-mentioned features into a long array as a comprehensive representation of all features together.

Experimental setup

We carried out several classifications involving different comparison groups, moving from broader divisions to more fine-grained comparisons on the AD continuum: ten binary and one ternary classification tasks. The binary comparisons were, firstly, the combined preclinical group (HOC+SCD = CON) vs. the clinical groups (MCI and pAD separately), and the general pathology (PATH) group comprising MCI and pAD together. Next, we performed three further binary classifications: HOC from pAD, SCD from pAD, and HOC from SCD; and four involving the MCI group: MCI from pAD, aMCI from pAD, naMCI from pAD, and aMCI from naMCI. The ternary classification attempted to discriminate CON, MCI, and pAD. Classification experiments were completed with the scikit-learn 1.0.2.

The random forests algorithm served as the classifier. This is an ensemble learning method for constructing multiple decision trees to vote for the final label. Its robustness to overfitting and noises motivated our choice (Breiman, 2001). Instead of forwarding all features to the classifier, we selected only the most informative variables to reduce computational load, lower the risk of overfitting, and remove noises from the feature set. In each experiment, we thus computed the ANOVA F-value between features in the feature set and ordered the features based on these F values. The number of selected features was automatically determined based on the classifier's performance, with a maximum of 1500 features. The classifier was evaluated using ten-fold cross-validation with precision, recall, F1, and accuracy scores averaged across the ten folds. Considering the imbalance between the number of participants in each comparison group, we averaged the performance scores across comparison groups (i.e., the macro scores). The macro F1 scores served as the major indicator for classifier performance as it takes data distribution into account and compensates for group imbalance.

Statistics

To test the power of the classifier in distinguishing between different stages of AD, a repeated measure ANOVA (RMANOVA) across different group comparisons was carried out. In addition, we conducted a RMANOVA across the different speech and language domains, to investigate how different feature sets influence the performance of the machine learning classifier. For each RMANOVA, we tested the assumption of sphericity and corrected with Greenhouse-Geisser method when the assumption was violated. Post-hoc tests were conducted in case of significant difference with Holm adjusted p-values. All statistical results are rounded to three digits, SD stands for standard deviation. Furthermore, we categorized feature sets into two overall modalities with the equal number of features, one 'speech' modality including the four features directly extracted from audios, and one 'text'-modality including the three feature sets from transcripts and the concatenation of all features. We carried out a paired-sample Wilcoxon signed-rank test to check if the speech modality and textual modality are different from each other. Analyses were run with JASP 0.16.3.0.

Results

Classification performance scores across groups

Table 3 reports the F1 scores of the random forest classifier on the ten classification tasks. To make the table more concise, we report only the averaged F1 scores performance score across groups for classification. The complete group-wise performance matrices, including precision, recall, F1, and accuracy scores, are reported in SM-D. Figure 1 (a) shows the distribution of the F1 scores obtained by the classifier on different group comparisons using different linguistic feature sets, and (b) shows the violin plot of F1 scores for each comparison, with a line of mean F1 scores. The post-hoc comparisons for both groups and linguistic domains are shown in SM-E. As indicated by the RMANOVA, the classifier performed significantly different among different comparisons ($F(10) = 14.423$, $p < .001$, $\eta^2 = .673$). Measured with the average F1 scores, compared to the ternary classification (mean = .572, SD = .077), the classifier performed significantly better on all binary classifications ($p < .05$). For binary classifications, the classifier performed best on distinguishing between groups without (HOC and SCD) and with (MCI and pAD) cognitive impairment, specifically SCD from pAD (mean = .878, SD = .034), followed by CON (that is, the combined group without objective cognitive decline) from pAD (mean = .812, SD = .039), CON from the combined 'pathological' group with cognitive decline (PATH) (mean = .786, SD = .024), and CON from MCI (mean = .774, SD = .038). Performance on the three comparisons between CON and pathological groups (pAD and MCI) were similar to each other ($p = 1.000$). Next, the classifier distinguished HOC from pAD (mean = .758, SD = .053) and HOC from SCD (mean = .749, SD = .143). It is noteworthy that performance on SCD vs. pAD was significantly better than HOC vs. pAD ($p = .005 < .05$). The seventh to tenth performance scores were: aMCI vs. naMCI (mean = .747, SD = .101), naMCI vs. pAD (mean = .720, SD = .068), MCI vs. pAD (mean = .695, SD = .061), and finally aMCI vs. pAD (mean = .678, SD = .158). The performance on naMCI vs. pAD was almost the same as that of MCI vs. pAD ($p = 1.000$) and aMCI vs. pAD ($p = 1.000$). The performance on CON vs. MCI was not significantly better than these four comparisons among MCI groups and between them and pAD ($p > .05$).

Classification performance scores across linguistic levels and modalities

Figure 1 (c) shows the violin plot of F1 scores for each feature set with a line of mean F1 scores. As indicated by the RMANOVA with Greenhouse-Geisser sphericity correction, the machine learning classifier performed significantly different among different feature sets ($F(2.214) = 12.423$, $p < .001$, $\eta^2 = .554$). Ordered by mean F1 scores, the classifier performed best on the concatenation of all features (mean = .813, SD = .069), followed by the voice quality measurements (mean = .796, SD = .075), the spectral coefficients (mean = .791, SD = .078), the cepstral coefficients (mean = .773, SD = .087), the embeddings from RoBERTa (mean = .738, SD = .111), the prosodic features (mean = .724, SD = .097), the morpho-lexical features (mean = .659, SD = .104), and the syntactic features (mean = .648, SD = .131). Syntactic and morpho-lexical features were similar to each other ($p = 1.000$), and significantly worse than all other feature sets ($p < .05$) except prosodic features ($p = .058$, $p = .158$, respectively). Prosodic features performed significantly worse only relative to the concatenation of all features ($p = .013$, $p < .05$). Post-hoc comparisons among other pairs of feature sets were all insignificant ($p > .05$). After grouping linguistic levels into two modalities, speech and text, speech-based features (mean = .771, median = .776, SD = .087) discriminated groups significantly better than text-based ones (mean = .714, median = .736, SD = .123, $z = 2.019$, $p = .043 < .05$). Figure 1 (d) shows the violin plot of F1-scores for each modality with a line of mean F1 scores.

Discussion

This study aimed to (i) test for the generalizability of previous results of automatic classification of AD from English to Spanish/Catalan in a new dataset; (ii) assess the performance of the classifier across different language-related featural domains, and (iii) include different MCI as well as preclinical groups at risk of AD in the classification. Our results confirm that similar performance as in previous studies based on Hungarian and English data can be generalized on our new Spanish/Catalan data (Gosztolya et al., 2019; Haulcy & Glass, 2021). For all binary comparisons we made, most of the F1 scores are around or higher than .7. For some specific separations we even achieved more impressive results, such as the F1 score of .912 in separating HOC from pAD based on spectral coefficients. For ternary classification, though the performance is worse, between .447-.662 by F1 score, they were higher than the chance level of .330, with the highest scores doubling the latter. These results suggest that speech analysis can be a potentially powerful and generalizable approach for automated pAD detection and risk for it.

As for performance in different feature domains, accuracies were generally satisfactorily high *across* domains, with the exception of morpho-lexical features and syntactic features. The former finding may be expected, as declarative memory-related cognitive impairment in AD may not result in changes in morphology, associated with procedural memory. More surprising is the finding on syntactic features, in light of the study of Chapin et al. (2022), where a number of hand-selected syntactic measures related to hierarchical syntactic complexity discriminated between controls, MCI and AD groups. One possibility is that significant changes in syntactic impairment occur in MCI, as the F1 scores achieved on distinguishing controls from MCIs and pAD were all around or above .7 (.688-.860), but less than .6 when comparing within groups with or without (objective) cognitive impairment (.460-.589). Decline in syntactic complexity also seems to be a late effect in the pathophysiological process, as compared to speech domains. Thus, in all three textual domains investigated, RoBERTa performed significantly better than syntactic and morpho-lexical features, and comparably to speech domains. Unlike manually designed feature sets including our syntactic one, RoBERTa is an integrated linguistic representation wrapping up what the model learned from the pre-trained corpus and current contexts into the embeddings. This highlights the importance of semantic changes in AD, as RoBERTa originates from distributional semantics-based word embeddings. Nonetheless, current studies have shown that these BERT-based models also capture lexical, syntactic, and conversational information in addition to semantic information (Kumar et al., 2021; Staliūnaitė & Iacobacci, 2020). Syntactic changes at the phrasal level could be important in AD, even for the early stages. Again, morphology and phrasal syntactic complexity could be more a matter of procedural memory, while the syntactic variables in our syntactic feature set could be argued to capture more declarative aspects of language use, such as specific forms of complexity needed to express episodic semantic information.

Another unexpected finding was that performance in speech domains was significantly better than in textual domains. This raises the thought-provoking question whether we even need transcripts and textual analysis, given the impressive performances of acoustic features and the high costs of the transcription task. Although fusing speech and text domains gives slight increases (less than .1) in performance scores, it is questionable whether such small increases balance the cost of transcription. The voice quality measures and spectral and cepstral coefficients performed the best and similar to each other. Although previous studies hypothesized voice quality changes as an ex-post effect from the physiological impairment of the fine control and the slowing down of vocal organs due to MCI and AD, the F1 scores of .742 on separating HOC and SCD and .814 on separating aMCI and naMCI suggest the roles of cognitive decline and memory loss in the changes of voice quality (Themistocleous et al., 2020). Paralinguistic features have been verified as performative in AD classification for multiple languages (Lindsay et al., 2021) and even across languages (Martinez de Lizarduy et al., 2017), which has been confirmed in our study. Similar to the voice quality measures, some studies also related paralinguistic features to voicing handicaps, so treated these changes as a side effect of AD (Awan et al., 2014). However, we found that the spectral and cepstral coefficients were more discriminative when separating HOC and SCD, followed by separations between individuals with and without cognitive impairment, including the aMCI vs naMCI, and finally between MCI and pAD. A more reasonable interpretation could be that paralinguistic features represent variance from other factors such as affective, apathy and executive functions (Lindsay et al., 2021).

As for prosody, when ranking the performance of prosody and syntax on different classification tasks from highest to lowest, similar results were found between them, with the comparisons between groups with and without cognitive impairment at the top and comparisons within these two general groups at the bottom. As prosody and syntax respectively represent the unification of sounds and words, we may conclude that organization abilities in these two domains decline greatly from cognitively healthy to cognitive impairment, but slowly progress within these two general phases.

Our final aim was to test the extendibility of previous classification results to further groups on the AD spectrum, specifically SCD and different MCI groups. Remarkably, very high accuracies were obtained when classifying cognitively healthy individuals with and without cognitive complaints (HOC and SCD), and classifying each from pAD, specifically when using the purely speech-based feature domains – spectral and cepstral coefficients. To our knowledge, this is the first report of an automatized differentiation between these preclinical groups, which is particularly noteworthy insofar as it does not depend on transcription. Future work following people with SCD over time could investigate this issue, by comparing the speech of converting vs. non-converting SCDs after an interval. Equally striking in our results is the differentiability of the amnestic and non-amnestic MCI groups, again based on spectral and cepstral features. Despite the high differentiability, although aMCI and naMCI are not that similar between themselves, they are similarly different from AD, unlike SCD and HOC. Furthermore, similar patterns were observed in groups, linguistic domains, and linguistic modalities, when we applied the gradient boosting algorithm as the classifier, another ensemble learning algorithm where the decision trees are not independent but will correct each other. Results from this algorithm and statistical comparisons can be found in SI-F. These similar patterns from a different algorithm indicate robustness in our classification results.

This study has several limitations. First, the dataset is relatively small from a computational perspective, so we can neither train the model with large data nor use state-of-the-art deep learning techniques. Although we managed to validate the result by using cross-validation, this still limits the performance of the classifier. Secondly, it is impossible to capture the full picture in every linguistic domain, though we used the available and already verified features to ensure representativeness. Finally, the performance on ternary classification is not high, likely due to the heterogeneity in groups and the size of the dataset.

In conclusion, our study shows that using machine learning based on speech can be a potentially powerful tool for detecting cognitive impairment and its gradation based on pAD pathology and that different linguistic domains play significantly different roles in this procedure. With clinical applications in mind, we underline both the high performance of speech-based measures as compared to text-based ones, and the discriminability even of objectively unimpaired healthy older adults with and without SCD, and of groups with amnestic and non-amnestic MCI. Combined with other behavioral markers or biological ones such as blood or retinal appearances, speech analysis may well prove to provide essential help for establishing early and robust diagnostic markers of AD, which are inexpensive and widely available.

Acknowledgments

We acknowledge all individuals participating in this study, including those in the FACEHBI study, for their generosity and their trust in us. We also want to thank our sponsors for making this project possible, as well as all the investigators from Ace Alzheimer Center Barcelona, Hospital Clinic, and Clínica Corachan for their close collaboration and continuous intellectual input. We would like to give special thanks to Mary Lofgren for providing valuable thoughts and assistance in the development of this study. This research was supported by the Generalitat de Catalunya (AGAUR) (grant 2017SGR1265 to WH), the Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI) (grant PID2019-105241GB-I00/AEI/10.13039/501100011033 to WH), the China Scholarship Council (grant 202108390062 to RH), and partially by the French Investissements d'Avenir - Labex EFL program (ANR-10- LABX-0083 to the IdEx Université Paris Cité - ANR-18-IDEX-0001, to JAT). This study was also funded by Grifols SA, Life Molecular Imaging, Laboratorios Echevarne, Araclon Biotech, and Ace Alzheimer Center Barcelona, and was supported by the Spanish Ministry of Health from Instituto de Salud Carlos III (Madrid) (FISS PI10/00945) and by the Agència d'Avaluació de Tecnologia i Recerca Mèdiques. It was also funded by Departament de Salut de la Generalitat de Catalunya (Health Department of the Catalan Government) (390) Ace Alzheimer Center Barcelona is part of the Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED, Spain) and is one of the participating centers of the Dementia Genetics Spanish Consortium (DEGESCO). This study has also been funded by Instituto de Salud Carlos III (ISCIII) Acción Estratégica en Salud, integrated in the Spanish National RCDCI Plan and fnanced 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. Marquié.

Data Availability Statement

The datasets generated and analyzed during the current study are not publicly available due to ethical requirements. Codes for processing the data and generating the figures are available from the corresponding author upon request.

References

Agbavor F., & Liang H. (2022). Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, *1*(12), e0000168. https://doi.org/10.1371/journal.pdig.0000168

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain: A Journal of Neurology*, *136*(Pt 12), 3727–3737. https://doi.org/10.1093/brain/awt269

Armengol-Estapé, J., Carrino, C. P., Rodriguez-Penagos, C., de Gibert Bonet, O., Armentano-Oller, C., Gonzalez-Agirre, A., Melero, M., & Villegas, M. (2021). Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4933–4946. https://doi.org/10.18653/v1/2021.findings-acl.437

Awan, S. N., Roy, N., & Cohen, S. M. (2014). Exploring the relationship between spectral and cepstral measures of voice and the Voice Handicap Index (VHI). *Journal of Voice: Official Journal of the Voice Foundation*, *28*(4), 430–439. https://doi.org/10.1016/j.jvoice.2013.12.008

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., & Novikova, J. (2021). Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. *Frontiers in Aging Neuroscience*, *13*, 635945. https://doi.org/10.3389/fnagi.2021.635945

Balagopalan, A., Eyre, B., Rudzicz, F., & Novikova, J. (2020). To BERT or Not To BERT:

Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection. *ArXiv:2008.01551 [Cs]*. http://arxiv.org/abs/2008.01551

Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, S., Wright, A., Tippett, D. C., & Hillis, A. E. (2019). Stealing Cookies in the Twenty-First Century: Measures of Spoken Narrative in Healthy Versus Speakers With Aphasia. *American Journal of Speech-Language Pathology*, *28*(1 Suppl), 321–329. https://doi.org/10.1044/2018_AJSLP-17-0131

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chapin, K., Clarke, N., Garrard, P., & Hinzen, W. (2022). A finer-grained linguistic profile of Alzheimer's disease and Mild Cognitive Impairment. *Journal of Neurolinguistics*, *63*, 101069. https://doi.org/10.1016/j.jneuroling.2022.101069

Chen, J., Ye, J., Tang, F., & Zhou, J. (2021). Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only. *Interspeech 2021*, 3830–3834. https://doi.org/10.21437/Interspeech.2021-2002

Clarke, N., Barrick, T. R., & Garrard, P. (2021). A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning. *Frontiers in Computer Science*, *3*. https://www.frontiersin.org/article/10.3389/fcomp.2021.634360

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., Blackburn, D., Schuller, B. W., Magimai-Doss, M., Strik, H., & Härmä, A. (2020). A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition.

Interspeech 2020, 2182–2186. https://doi.org/10.21437/Interspeech.2020-2635

de la Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, *78*(4), 1547–1574. https://doi.org/10.3233/JAD-200888

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., & Naylor, M. (2020). Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, *28*. https://doi.org/10.1016/j.eclinm.2020.100583

Fyndanis, V., Manouilidou, C., Koufou, E., & Tsapakis, E. M. (2011). Grammatical Disorders in Alzheimer's Disease: Evidence from Verb Inflection in Greek. *Procedia - Social and Behavioral Sciences*, *23*, 221–222.

Goodglass, H., Kaplan, E., & Barresi, B. (1972). *The assessment of aphasia and related disorders*. Lea & Febiger.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, *53*, 181–197. https://doi.org/10.1016/j.csl.2018.07.007

Guinn, C., Singer, B., & Habash, A. (2014). A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. *2014 IEEE Symposium on Computational Intelligence in Healthcare and E-Health (CICARE)*, 98–103. https://doi.org/10.1109/CICARE.2014.7007840

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Armentano-Oller, C., Rodriguez-Penagos, C., Gonzalez-Agirre, A., & Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, *68*(0), Article 0.

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55–58. https://doi.org/10.1126/science.aax0289

Haider, F., de la Fuente, S., & Luz, S. (2020). An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 272–281. https://doi.org/10.1109/JSTSP.2019.2955022

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(52), 14365–14374. https://doi.org/10.1523/JNEUROSCI.4549-07.2007

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, *11*(7), 299–306. https://doi.org/10.1016/j.tics.2007.05.001

Haulcy, R., & Glass, J. (2021). Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.624137

Irish, M., Halena, S., Kamminga, J., Tu, S., Hornberger, M., & Hodges, J. R. (2015). Scene construction impairments in Alzheimer's disease—A unique role for the posterior cingulate cortex. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *73*, 10–23. https://doi.org/10.1016/j.cortex.2015.08.004

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jessen, F., Amariglio, R. E., van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K. A., van der Flier, W. M., Glodzik, L., van Harten, A. C., de Leon, M. J., McHugh, P., Mielke, M. M., Molinuevo, J. L., Mosconi, L., Osorio, R. S., Perrotin, A., … Subjective Cognitive Decline Initiative (SCD-I) Working Group. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *10*(6), 844–852. https://doi.org/10.1016/j.jalz.2014.01.001

Kavé, G., & Levy, Y. (2003). Morphology in picture descriptions provided by persons with Alzheimer's disease. *Journal of Speech, Language, and Hearing Research: JSLHR*, *46*(2), 341–352.

Kumar, A., Narayanan Sundararaman, M., & Vepa, J. (2021). What BERT Based Language Model Learns in Spoken Transcripts: An Empirical Study. *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 322–336. https://doi.org/10.18653/v1/2021.blackboxnlp-1.25

Lindsay, H., Tröger, J., & König, A. (2021). Language Impairment in Alzheimer's Disease—

Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, *13*, 642033. https://doi.org/10.3389/fnagi.2021.642033

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lofgren, M., & Hinzen, W. (2022). Breaking the flow of thought: Increase of empty pauses in the connected speech of people with mild and moderate Alzheimer's disease. *Journal of Communication Disorders*, *97*, 106214. https://doi.org/10.1016/j.jcomdis.2022.106214

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2021). *Detecting cognitive decline using speech only: The ADReSSo Challenge* (arXiv:2104.09356). arXiv. https://doi.org/10.48550/arXiv.2104.09356

Luz, S., Haider, F., Fuente, S. de la, Fromm, D., & MacWhinney, B. (2020). Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. *Interspeech 2020*, 2172–2176. https://doi.org/10.21437/Interspeech.2020-2571

Manouilidou, C., Roumpea, G., Nousia, A., Stavrakaki, S., & Nasios, G. (2020). Revisiting Aspect in Mild Cognitive Impairment and Alzheimer's Disease: Evidence From Greek. *Frontiers in Communication*, *5*. https://www.frontiersin.org/articles/10.3389/fcomm.2020.434106

Martinez de Lizarduy, U., Calvo Salomón, P., Gómez Vilda, P., Ecay Torres, M., & López de Ipiña, K. (2017). ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens*, *4*(1), 037. https://doi.org/10.3989/loquens.2017.037

Mertens, P. (2004, January). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. *Proceedings of Speech Prosody 2004*.

Mossello, E., & Ballini, E. (2012). Management of patients with Alzheimer's disease: Pharmacological treatment and quality of life. *Therapeutic Advances in Chronic Disease*, *3*(4), 183–193. https://doi.org/10.1177/2040622312452387

Nasiri M., Moayedfar S., Purmohammad M., & Ghasisin L. (2022). Investigating sentence processing and working memory in patients with mild Alzheimer and elderly people. *PLOS ONE*, *17*(11), e0266552. https://doi.org/10.1371/journal.pone.0266552

Nasrolahzadeh, M., Mohammadpoory, Z., & Haddadnia, J. (2016). A novel method for early diagnosis of Alzheimer's disease based on higher-order spectral estimation of spontaneous speech signals. *Cognitive Neurodynamics*, *10*(6), 495–503. https://doi.org/10.1007/s11571-016-9406-0

Orimaye, S. O., Wong, J. S.-M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, *18*(1), 34. https://doi.org/10.1186/s12859-016-1456-0

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, *256*(3), 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., Köpke, B., Puel, M., & Pariente, J. (2016). Pauses During Autobiographical Discourse Reflect Episodic Memory Processes in Early Alzheimer's Disease. *Journal of Alzheimer's*

Disease, *50*(3), 687. https://doi.org/10.3233/JAD-150408

Qiao, Y., Xuefeng, W., Daniel, W., & Elma, K. (2021). Alzheimer’s Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models. *InterSpeech 2021*, 3805--3809. https://doi.org/10.21437/Interspeech.2021-1415

Rabin, L. A., Smart, C. M., & Amariglio, R. E. (2017). Subjective Cognitive Decline in Preclinical Alzheimer’s Disease. *Annual Review of Clinical Psychology*, *13*, 369–396. https://doi.org/10.1146/annurev-clinpsy-032816-045136

Rodriguez-Gomez, O., Sanabria, A., Perez-Cordon, A., Sanchez-Ruiz, D., Abdelnour, C., Valero, S., Hernandez, I., Rosende-Roca, M., Mauleon, A., Vargas, L., Alegret, M., Espinosa, A., Ortega, G., Guitart, M., Gailhajanet, A., Sotolongo-Grau, O., Moreno-Grau, S., Ruiz, S., Tarragona, M., … Boada, M. (2017). FACEHBI: A Prospective Study of Risk Factors, Biomarkers and Cognition in a Cohort of Individuals with Subjective Cognitive Decline. Study Rationale and Research Protocols. *The Journal of Prevention of Alzheimer’s Disease*, *4*(2), 100–108. https://doi.org/10.14283/jpad.2016.122

Roger, E., Banjac, S., Thiebaut de Schotten, M., & Baciu, M. (2022). Missing links: The functional unification of language and memory (L ∪ M). *Neuroscience & Biobehavioral Reviews*, *133*, 104489. https://doi.org/10.1016/j.neubiorev.2021.12.012

Roshanzamir, A., Aghajan, H., & Soleymani Baghshah, M. (2021). Transformer-based deep neural network language models for Alzheimer’s disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, *21*(1), 92.

https://doi.org/10.1186/s12911-021-01456-3

Sarawgi, U., Zulfikar, W., Soliman, N., & Maes, P. (2020). Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity. *Interspeech 2020*, 2212–2216. https://doi.org/10.21437/Interspeech.2020-3137

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences*, *17*, 41–50. https://doi.org/10.1016/j.cobeha.2017.06.002

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., & Evanini, K. (2016). The INTERSPEECH 2016 computational paralinguistics challenge: 17th Annual Conference of the International Speech Communication Association. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-September-2016*, 2001–2005. https://doi.org/10.21437/Interspeech.2016-129

Sherratt, S., & Bryan, K. (2019). Textual cohesion in oral narrative and procedural discourse: The effects of ageing and cognitive skills. *International Journal of Language & Communication Disorders*, *54*(1), 95–109. https://doi.org/10.1111/1460-6984.12434

Staliūnaitė, I., & Iacobacci, I. (2020). Compositional and Lexical Semantics in RoBERTa, BERT and DistilBERT: A Case Study on CoQA. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7046–7056. https://doi.org/10.18653/v1/2020.emnlp-main.573

Thakral, P. P., Madore, K. P., Kalinowski, S. E., & Schacter, D. L. (2020). Modulation of hippocampal brain networks produces changes in episodic simulation and divergent thinking. *Proceedings of the National Academy of Sciences*, *117*(23), 12729–12740. https://doi.org/10.1073/pnas.2003535117

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2018). Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. *Frontiers in Neurology*, *9*, 975. https://doi.org/10.3389/fneur.2018.00975

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PLoS ONE*, *15*(7), e0236009. https://doi.org/10.1371/journal.pone.0236009

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., & Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference on Mechatronics and Automation, ICMA 2005* (Vol. 3, p. 1574 Vol. 3). https://doi.org/10.1109/ICMA.2005.1626789

Thomas, J. A., Burkhardt, H. A., Chaudhry, S., Ngo, A. D., Sharma, S., Zhang, L., Au, R., & Hosseini Ghomi, R. (2020). Assessing the Utility of Language and Voice Biomarkers to Predict Cognitive Impairment in the Framingham Heart Study Cognitive Aging Cohort Data. *Journal of Alzheimer's Disease: JAD*, *76*(3), 905–922. https://doi.org/10.3233/JAD-190783

Uretsky, M., Gibbons, L. E., Mukherjee, S., Trittschuh, E. H., Fardo, D. W., Boyle, P. A., Keene, C. D., Saykin, A. J., Crane, P. K., Schneider, J. A., & Mez, J. (2021). Longitudinal cognitive performance of Alzheimer's disease neuropathological subtypes. *Alzheimer's & Dementia (New York, N. Y.)*, *7*(1), e12201. https://doi.org/10.1002/trc2.12201

World Health Organization. (2020, December 9). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Zhu, Y., Obyat, A., Liang, X., Batsis, J. A., & Roth, R. M. (2021). WavBERT: Exploiting Semantic and Non-Semantic Speech Using Wav2vec and BERT for Dementia Detection. *Interspeech 2021*, 3790–3794. https://doi.org/10.21437/Interspeech.2021-332

Supplementary file description:

The supplementary information is divided into six parts : (A) The recruitment procedure, diagnostic criteria, neuropsychological battery, impact of the epidemic, and criteria for inclusion and exclusion. (B) Speech transcription instruction. (C) Feature list and definitions. (D) Detailed group-wise classifier performance from the random forest. (F) Post-hoc analysis of the RMANOVA tests. (F) Results and statistical comparison from Gradient Boosting.

Tables

Table 1: Participant demographics†

	HOC	SCD	naMCI	aMCI	pAD	Test	*p* value
Number	18	31	23	16	31	/	/
Age	66 (8)	69 (10)	75 (8)	78 (7)	81 (9)	KW test	< .001***
Age Range	58 - 89	56 - 85	52 - 88	66 - 89	60 - 93	/	/
Sex	61.1	25.8	39.1	37.5	45.2	χ^2	.177
Education	7 (4)	7 (3)	5 (3)	5 (3)	5 (2)	KW test	.000***
Language	33.3	22.6	4.4	12.5	12.9	χ^2 (Fisher)	.126
MMSE score	29.0 (1.0)	29.0 (2.0)	27.0 (3.0)	28.0 (2.1)	23.0 (4.5)	KW test	< .001***
CDR score	0	0	.5	.5	1 or 2	/	/

Note: HOC: healthy older control; SCD: subjective cognitive decline; naMCI: non-amnestic mild cognitive impairment; aMCI: amnestic mild cognitive impairment; pAD: probable Alzheimer's disease dementia; KW test: Kruskal-Wallis test; MMSE: Mini-Mental State Examination; CDR: Clinical Dementia Rating.

*$p < .05$, **$p < .01$, and *** $p < .001$.

†: Age in years. Gender is represented by the proportion of females. Education in education level. Language is represented by the proportion of people answering questions in Catalan. Fisher's exact test is applied for language as several expected frequency less than 5.

Table 2: Linguistic features extracted

Levels	Description	Num	Tools
Spectral	Auditory spectrum and the relative spectral transform	2800	openSMILE
Cepstral	Mel-Frequency cepstral coefficients 0–14.	1400	openSMILE
Voice quality	Jitter, shimmer, loudness, and log harmony-noise-ratio	2012	openSMILE
Prosodic	Include frequency, speech rate, pitch variation, pitch stylization etc.	199	openSMILE& Prosogram
Morpho-lexical	Ratios of different word classes and the morphological variants, e.g. masculine nouns	154	Stanza
Syntactic	Ratios of syntactic features selected from Chapin et al.(Chapin et al., 2022), e.g. verb modality	21	Manual
Semantic	Pooled output of RoBERTa models	768[†]	RoBERTa

†: number of dimensions of the word embeddings

Table 3: Classifier performance across different comparisons and feature sets

Comparison	spectral	cepstral	prosodic	voice quality	syntactic	morpho-lexical	RoBERTa	all
CON/PATH	.789	.776	.805*	.804*	.802*	.738	.768	.806*
CON/pAD	.825*	.833*	.819*	.853**	.771	.735	.826*	.834*
CON/MCI	.801*	.769	.798	.799	.743	.700	.771	.814*
HOC/pAD	.765	.775	.757	.800*	.688	.676	.832*	.771
HOC/SCD	.912***	.908***	.702	.742	.534	.649	.649	.900***
HOC/pAD	.765	.775	.757	.800*	.688	.676	.832*	.771
SCD/pAD	.878**	.880**	.854**	.903***	.860**	.819*	.928***	.898**
MCI/pAD	.732	.724	.682	.753	.589	.635	.675	.768
aMCI/pAD	.719	.724	.539	.866**	.460	.508	.737	.869**
naMCI/ pAD	.734	.764	.718	.801*	.581	.690	.696	.776
aMCI/ naMCI	.881**	.776	.696	.814*	.581	.656	.730	.847*

CON/MCI/pAD	.662	.579	.591	.618	.516	.447	.505	.660

Note: *** F1 score > = .9 for very good, ** F1 score > = .85 for good, * F1 score > = .8 for not bad

Figures

Figure 1. (a) Scatter plot of F1-scores across different feature sets and classification tasks; (b) Means, standard deviations and distributions of the same F1-scores across classification tasks with the mean line; (c) F1-scores across different feature sets with the mean line; (d) F1-scores between different modalities (speech vs. text) with the mean line.

Learning Outcomes

- Machine learning based on automatically extracted language features detected cognitive decline from early stages of the AD continuum in a new Spanish-Catalan dataset.
- Different speech and language domains showed differential discrimination performance between groups, with features extracted directly from speech performing better than those from the text.
- Before the onset of objective cognitive impairment, speech and language from older adults with Subjective Cognitive Decline (SCD) showed speech and language differences from controls without SCD, indicating potential heterogeneity in these non-clinical groups.