



HAL
open science

Automated Classification of Cognitive Decline and Probable Alzheimer's Dementia Across Multiple Speech and Language Domains

Rui He, Kayla Chapin, Jalal Al-Tamimi, Núria Bel, Marta Marquié, Maitee Rosende-Roca, Vanesa Pytel, Juan Pablo Tartari, Montse Alegret, Angela Sanabria, et al.

► **To cite this version:**

Rui He, Kayla Chapin, Jalal Al-Tamimi, Núria Bel, Marta Marquié, et al.. Automated Classification of Cognitive Decline and Probable Alzheimer's Dementia Across Multiple Speech and Language Domains. *American Journal of Speech-Language Pathology*, 2023, 32 (5), pp.2075 - 2086. 10.1044/2023_ajslp-22-00403 . hal-04230592

HAL Id: hal-04230592

<https://hal.science/hal-04230592>

Submitted on 6 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Title page**

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

4 **Running title:** Detect pAD across subsections & language domains

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

9 **Affiliations and addresses of the authors:**

10 ^a Department of Translation & Language Sciences, Universitat Pompeu Fabra, Carrer Roc
11 Boronat, 138, 08018 Barcelona, Spain.

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

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

16 ^d Networking Research Center on Neurodegenerative Diseases (CIBERNED), Instituto de
17 Salud Carlos III, Madrid, Spain.

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

20 **Correspondence:** Rui He, Department of Translation & Language Sciences, Universitat
21 Pompeu Fabra, Carrer Roc Boronat, 138, Barcelona 08018, Spain, rui.he@upf.edu,
22 [@RuiHe76864182](https://twitter.com/RuiHe76864182)

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

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

38

39 **Abstract**

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

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

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

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

61 these domains as well.

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

63

64 **Introduction**

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

84 Automated approaches from natural language processing (NLP) and machine learning
85 have already shown strong capability in predicting AD, and even MCI (de la Fuente Garcia et

86 al., 2020). Paralinguistic measures extracted from speech directly, such as mathematical
87 properties of sound waves, have also shown impressive power in identifying AD (Chen et al.,
88 2021; Haider et al., 2020; Sarawgi et al., 2020). Some studies transcribed audio into texts, either
89 by hand or machine, and extracted features from texts, including lexical, syntactic, and N-gram
90 features (Orimaye et al., 2017; Qiao et al., 2021; C. Thomas et al., 2005). Unlike these
91 traditional feature engineering approaches, transfer learning based on pretrained language
92 models, such as bidirectional encoder representations from transformers (BERT), encodes
93 linguistic information from large corpora into vector representations or word embeddings.
94 These have been proven powerful in language modeling and some studies have suggested to
95 use them for AD detection due to excellent performance in a binary AD vs. control comparison
96 based on the ADReSS dataset (Balagopalan et al., 2021; Jawahar et al., 2019). Roshanzamir et
97 al. (2021) obtained an accuracy of 88.08% with BERT as an encoder with a logistic regression
98 classifier, in a classification of AD vs. controls with English data from the Pitt corpus. Using
99 Hungarian data, Gosztolya et al. (2019) achieved 74%–82% accuracy in classifying diagnostic
100 groups (cognitively healthy, MCI, and AD) based on speech (acoustic) features, and similar
101 results using language features.

102 Despite a number of promising studies, several challenges need to be addressed before
103 speech- and language-based classification can be utilized in clinical applications. First, existing
104 studies have mainly used the dataset from the InterSpeech challenge and its source corpus (Pitt
105 Corpus) (Luz et al., 2020, 2021). These datasets only comprise English data from control and
106 AD groups, so lacks data from disease stages in between, especially prodromal ones. Validation
107 of the technically most advanced classifiers across different languages and the entire AD

108 continuum is necessary. Second, current studies have investigated different linguistic levels
109 with generic linguistic variables in a bottom-top fashion (Chapin et al., 2022), yet have not
110 assessed in detail the differential feature performance across domains. To the best of our
111 knowledge, only some studies have compared features from both speech and text, and they
112 have not engaged in fine-grained comparisons of domains within these modalities, especially
113 for the textual one, which has been represented chiefly by pretrained models (Balagopalan et
114 al., 2020; Cummins et al., 2020; Zhu et al., 2021).

115 Finally, most available studies have elicited connected speech through a picture description,
116 using the Cookie Theft picture (Goodglass et al., 1972). This simple but efficient task has
117 significantly contributed to insights on AD detection through connected speech, but received
118 some criticism including the invocation of stereotypes of family life, elicitation of overly
119 simplified speech, and limited recollection and engagement (Berube et al., 2019; Clarke et al.,
120 2021; Sherratt & Bryan, 2019). Beyond these, the Cookie Theft picture does not challenge the
121 creative and imaginative use of language as speech is generated while looking at the picture
122 and objects are visually available for naming. As a hallmark of language is referencing objects
123 that are not visually present, tasks without visual prompts may show more sensitivity than
124 picture descriptions by challenging language production in one of its core features, namely
125 displaced reference. In particular, scene constructions (SC) have been proposed to stimulate
126 speech for richer information and better discrimination (Irish et al., 2015). SC requires a mental
127 time travel to another location (e.g., a tropical island) and the imaginative representation of
128 directly experienced events at this location. As this information has an episodic and first-
129 personal character, the task taps into a cognitive process widely reported to be impaired in AD

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

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

136

137 **Methods**

138 *Dataset*

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

152 assessment, and other issues. Table 1 shows the demographic and neuropsychological data of
153 the sample (data: median (interquartile range, IQR)).

154

155 *Speech data elicitation and processing*

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

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

172 In particular, we used openSMILE 3.0 to extract spectral, cepstral, voice quality, and part
173 of prosodic features from the CompareE 2016 feature set at the level of functionals (Eyben et

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

188 Syntactic features were manually annotated following the method of Chapin et al. (2022),
189 which targeted specific forms of syntactic complexity involved in referencing objects and
190 events. This feature set, though showing a close relation to neurodegeneration in AD, is not
191 automatable yet. Thus, we added the morpho-lexical features, which have often been used to
192 approximate syntax for measuring language changes in AD, especially in the context of NLP
193 (verb inflection in Fyndanis et al., 2011; ratios of different word classes in Guinn et al., 2014;
194 verb aspect in Manouilidou et al., 2020; verb voice in Nasiri et al., 2022). Morpho-lexical
195 features include the ratios of different word classes and the morphological variants of the

196 content words, which are automatically extracted by the Spanish and Catalan models from
197 Stanza 1.3.0. Lexical and morphological features were found to be important in the progression
198 of AD, in both contexts of group comparison (Kavé & Levy, 2003) and machine learning
199 (Eyigoz et al., 2020). In addition, morphological variants in inflectional languages like Spanish
200 and Catalan provide paths to observe grammar at word level (e.g. aspect and modality).
201 Semantic changes are thought to be prominent in AD, and large computational language
202 models capture distributional aspects of language as a proxy of meaning. A number of studies
203 has found the application of language models on AD detection to be satisfying (Agbavor &
204 Liang, 2022; Balagopalan et al., 2021). The robust optimized version of BERT (RoBERTa)
205 was chosen for encoding the human transcripts. Due to the limitation on token numbers, we
206 truncated the text and only encoded the first 510 tokens, if the text length was longer than that
207 (Liu et al., 2019). We used Catalan BERTa (RoBERTa-base) for Catalan data
208 (<https://huggingface.co/projecte-aina/roberta-base-ca-cased-tc>) and Byte-Pair Encoding
209 RoBERTa for Spanish data (<https://huggingface.co/PlanTL-GOB-ES/roberta-base-bn>)
210 (Armengol-Estapé et al., 2021; Gutiérrez-Fandiño et al., 2022). As RoBERTa models return an
211 embedding for every token in the text, which does not fit with the purpose of assigning a single
212 label to the text, we utilized the pooled output as a semantic representation. This output is the
213 embedding of the initial classification token that arises from the sequence output with
214 contextual information from all tokens of the sequence embedded in it. It is standardly used for
215 classification tasks. The complete list of feature sets is available in SI-C. In addition, we also
216 concatenated all the above-mentioned features into a long array as a comprehensive
217 representation of all features together.

218

219 *Experimental setup*

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

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

240 macro scores). The macro F1 scores served as the major indicator for classifier performance as
241 it takes data distribution into account and compensates for group imbalance.

242

243 *Statistics*

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

257

258 **Results**

259 *Classification performance scores across groups*

260 Table 3 reports the F1 scores of the random forest classifier on the ten classification tasks. To
261 make the table more concise, we report only the averaged F1 scores performance score across

262 groups for classification. The complete group-wise performance matrices, including precision,
263 recall, F1, and accuracy scores, are reported in SM-D. Figure 1 (a) shows the distribution of
264 the F1 scores obtained by the classifier on different group comparisons using different
265 linguistic feature sets, and (b) shows the violin plot of F1 scores for each comparison, with a
266 line of mean F1 scores. The post-hoc comparisons for both groups and linguistic domains are
267 shown in SM-E. As indicated by the RMANOVA, the classifier performed significantly
268 different among different comparisons ($F(10) = 14.423, p < .001, \eta^2 = .673$). Measured with
269 the average F1 scores, compared to the ternary classification (mean = .572, SD = .077), the
270 classifier performed significantly better on all binary classifications ($p < .05$). For binary
271 classifications, the classifier performed best on distinguishing between groups without (HOC
272 and SCD) and with (MCI and pAD) cognitive impairment, specifically SCD from pAD (mean
273 = .878, SD = .034), followed by CON (that is, the combined group without objective cognitive
274 decline) from pAD (mean = .812, SD = .039), CON from the combined ‘pathological’ group
275 with cognitive decline (PATH) (mean = .786, SD = .024), and CON from MCI (mean = .774,
276 SD = .038). Performance on the three comparisons between CON and pathological groups
277 (pAD and MCI) were similar to each other ($p = 1.000$). Next, the classifier distinguished HOC
278 from pAD (mean = .758, SD = .053) and HOC from SCD (mean = .749, SD = .143). It is
279 noteworthy that performance on SCD vs. pAD was significantly better than HOC vs. pAD (p
280 = .005 < .05). The seventh to tenth performance scores were: aMCI vs. naMCI (mean = .747,
281 SD = .101), naMCI vs. pAD (mean = .720, SD = .068), MCI vs. pAD (mean = .695, SD = .061),
282 and finally aMCI vs. pAD (mean = .678, SD = .158). The performance on naMCI vs. pAD was
283 almost the same as that of MCI vs. pAD ($p = 1.000$) and aMCI vs. pAD ($p = 1.000$). The

284 performance on CON vs. MCI was not significantly better than these four comparisons among
285 MCI groups and between them and pAD ($p > .05$).

286

287 *Classification performance scores across linguistic levels and modalities*

288 Figure 1 (c) shows the violin plot of F1 scores for each feature set with a line of mean F1 scores.

289 As indicated by the RMANOVA with Greenhouse-Geisser sphericity correction, the machine
290 learning classifier performed significantly different among different feature sets ($F(2.214) =$

291 $12.423, p < .001, \eta^2 = .554$). Ordered by mean F1 scores, the classifier performed best on the

292 concatenation of all features (mean = .813, SD = .069), followed by the voice quality

293 measurements (mean = .796, SD = .075), the spectral coefficients (mean = .791, SD = .078),

294 the cepstral coefficients (mean = .773, SD = .087), the embeddings from RoBERTa (mean

295 = .738, SD = .111), the prosodic features (mean = .724, SD = .097), the morpho-lexical features

296 (mean = .659, SD = .104), and the syntactic features (mean = .648, SD = .131). Syntactic and

297 morpho-lexical features were similar to each other ($p = 1.000$), and significantly worse than all

298 other feature sets ($p < .05$) except prosodic features ($p = .058, p = .158$, respectively). Prosodic

299 features performed significantly worse only relative to the concatenation of all features (p

300 = .013, $p < .05$). Post-hoc comparisons among other pairs of feature sets were all insignificant

301 ($p > .05$). After grouping linguistic levels into two modalities, speech and text, speech-based

302 features (mean = .771, median = .776, SD = .087) discriminated groups significantly better

303 than text-based ones (mean = .714, median = .736, SD = .123, $z = 2.019, p = .043 < .05$). Figure

304 1 (d) shows the violin plot of F1-scores for each modality with a line of mean F1 scores.

305

306 **Discussion**

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

320 As for performance in different feature domains, accuracies were generally satisfactorily
321 high *across* domains, with the exception of morpho-lexical features and syntactic features. The
322 former finding may be expected, as declarative memory-related cognitive impairment in AD
323 may not result in changes in morphology, associated with procedural memory. More surprising
324 is the finding on syntactic features, in light of the study of Chapin et al. (2022), where a number
325 of hand-selected syntactic measures related to hierarchical syntactic complexity discriminated
326 between controls, MCI and AD groups. One possibility is that significant changes in syntactic
327 impairment occur in MCI, as the F1 scores achieved on distinguishing controls from MCIs and

328 pAD were all around or above .7 (.688-.860), but less than .6 when comparing within groups
329 with or without (objective) cognitive impairment (.460-.589). Decline in syntactic complexity
330 also seems to be a late effect in the pathophysiological process, as compared to speech domains.
331 Thus, in all three textual domains investigated, RoBERTa performed significantly better than
332 syntactic and morpho-lexical features, and comparably to speech domains. Unlike manually
333 designed feature sets including our syntactic one, RoBERTa is an integrated linguistic
334 representation wrapping up what the model learned from the pre-trained corpus and current
335 contexts into the embeddings. This highlights the importance of semantic changes in AD, as
336 RoBERTa originates from distributional semantics-based word embeddings. Nonetheless,
337 current studies have shown that these BERT-based models also capture lexical, syntactic, and
338 conversational information in addition to semantic information (Kumar et al., 2021; Staliūnaitė
339 & Iacobacci, 2020). Syntactic changes at the phrasal level could be important in AD, even for
340 the early stages. Again, morphology and phrasal syntactic complexity could be more a matter
341 of procedural memory, while the syntactic variables in our syntactic feature set could be argued
342 to capture more declarative aspects of language use, such as specific forms of complexity
343 needed to express episodic semantic information.

344 Another unexpected finding was that performance in speech domains was significantly
345 better than in textual domains. This raises the thought-provoking question whether we even
346 need transcripts and textual analysis, given the impressive performances of acoustic features
347 and the high costs of the transcription task. Although fusing speech and text domains gives
348 slight increases (less than .1) in performance scores, it is questionable whether such small
349 increases balance the cost of transcription. The voice quality measures and spectral and cepstral

350 coefficients performed the best and similar to each other. Although previous studies
351 hypothesized voice quality changes as an ex-post effect from the physiological impairment of
352 the fine control and the slowing down of vocal organs due to MCI and AD, the F1 scores of .742
353 on separating HOC and SCD and .814 on separating aMCI and naMCI suggest the roles of
354 cognitive decline and memory loss in the changes of voice quality (Themistocleous et al., 2020).
355 Paralinguistic features have been verified as performative in AD classification for multiple
356 languages (Lindsay et al., 2021) and even across languages (Martinez de Lizarduy et al., 2017),
357 which has been confirmed in our study. Similar to the voice quality measures, some studies
358 also related paralinguistic features to voicing handicaps, so treated these changes as a side
359 effect of AD (Awan et al., 2014). However, we found that the spectral and cepstral coefficients
360 were more discriminative when separating HOC and SCD, followed by separations between
361 individuals with and without cognitive impairment, including the aMCI vs naMCI, and finally
362 between MCI and pAD. A more reasonable interpretation could be that paralinguistic features
363 represent variance from other factors such as affective, apathy and executive functions
364 (Lindsay et al., 2021).

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

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

390 This study has several limitations. First, the dataset is relatively small from a
391 computational perspective, so we can neither train the model with large data nor use state-of-
392 the-art deep learning techniques. Although we managed to validate the result by using cross-
393 validation, this still limits the performance of the classifier. Secondly, it is impossible to capture

394 the full picture in every linguistic domain, though we used the available and already verified
395 features to ensure representativeness. Finally, the performance on ternary classification is not
396 high, likely due to the heterogeneity in groups and the size of the dataset.

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

407

408

409 **Acknowledgments**

410 We acknowledge all individuals participating in this study, including those in the FACEHBI
411 study, for their generosity and their trust in us. We also want to thank our sponsors for making
412 this project possible, as well as all the investigators from Ace Alzheimer Center Barcelona,
413 Hospital Clinic, and Clínica Corachan for their close collaboration and continuous intellectual
414 input. We would like to give special thanks to Mary Lofgren for providing valuable thoughts
415 and assistance in the development of this study. This research was supported by the Generalitat
416 de Catalunya (AGAUR) (grant 2017SGR1265 to WH), the Ministerio de Ciencia, Innovación

417 y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI) (grant PID2019-
418 105241GB-I00/AEI/10.13039/501100011033 to WH), the China Scholarship Council (grant
419 202108390062 to RH), and partially by the French Investissements d’Avenir - Labex EFL
420 program (ANR-10- LABX-0083 to the IdEx Université Paris Cité - ANR-18-IDEX-0001, to
421 JAT). This study was also funded by Grifols SA, Life Molecular Imaging, Laboratorios
422 Echevarne, Araclon Biotech, and Ace Alzheimer Center Barcelona, and was supported by the
423 Spanish Ministry of Health from Instituto de Salud Carlos III (Madrid) (FISS PI10/00945) and
424 by the Agència d’Avaluació de Tecnologia i Recerca Mèdiques. It was also funded by
425 Departament de Salut de la Generalitat de Catalunya (Health Department of the Catalan
426 Government) (390) Ace Alzheimer Center Barcelona is part of the Centro de Investigación
427 Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED, Spain) and is one
428 of the participating centers of the Dementia Genetics Spanish Consortium (DEGESCO). This
429 study has also been funded by Instituto de Salud Carlos III (ISCIII) Acción Estratégica en Salud,
430 integrated in the Spanish National RCDCI Plan and financed by ISCIII-Subdirección General
431 de Evaluación and the Fondo Europeo de Desarrollo Regional (FEDER-Una manera de hacer
432 Europa) grant PI19/00335 awarded to M. Marquié.

433

434 **Data Availability Statement**

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

438 **References**

- 439 Agbavor F., & Liang H. (2022). Predicting dementia from spontaneous speech using large
440 language models. *PLOS Digital Health*, 1(12), e0000168.
441 <https://doi.org/10.1371/journal.pdig.0000168>
- 442 Ahmed, S., Haigh, A.-M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a
443 marker of disease progression in autopsy-proven Alzheimer's disease. *Brain: A Journal*
444 *of Neurology*, 136(Pt 12), 3727–3737. <https://doi.org/10.1093/brain/awt269>
- 445 Armengol-Estapé, J., Carrino, C. P., Rodriguez-Penagos, C., de Gibert Bonet, O., Armentano-
446 Oller, C., Gonzalez-Agirre, A., Melero, M., & Villegas, M. (2021). Are Multilingual
447 Models the Best Choice for Moderately Under-resourced Languages? A
448 Comprehensive Assessment for Catalan. *Findings of the Association for Computational*
449 *Linguistics: ACL-IJCNLP 2021*, 4933–4946.
450 <https://doi.org/10.18653/v1/2021.findings-acl.437>
- 451 Awan, S. N., Roy, N., & Cohen, S. M. (2014). Exploring the relationship between spectral and
452 cepstral measures of voice and the Voice Handicap Index (VHI). *Journal of Voice:*
453 *Official Journal of the Voice Foundation*, 28(4), 430–439.
454 <https://doi.org/10.1016/j.jvoice.2013.12.008>
- 455 Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., & Novikova, J. (2021). Comparing Pre-
456 trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on
457 Speech. *Frontiers in Aging Neuroscience*, 13, 635945.
458 <https://doi.org/10.3389/fnagi.2021.635945>
- 459 Balagopalan, A., Eyre, B., Rudzicz, F., & Novikova, J. (2020). To BERT or Not To BERT:

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

462 Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, S., Wright, A., Tippett, D. C., &
463 Hillis, A. E. (2019). Stealing Cookies in the Twenty-First Century: Measures of Spoken
464 Narrative in Healthy Versus Speakers With Aphasia. *American Journal of Speech-*
465 *Language Pathology*, 28(1 Suppl), 321–329. [https://doi.org/10.1044/2018_AJSLP-17-](https://doi.org/10.1044/2018_AJSLP-17-0131)
466 0131

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

469 Chapin, K., Clarke, N., Garrard, P., & Hinzen, W. (2022). A finer-grained linguistic profile of
470 Alzheimer's disease and Mild Cognitive Impairment. *Journal of Neurolinguistics*, 63,
471 101069. <https://doi.org/10.1016/j.jneuroling.2022.101069>

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

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

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

482 *Interspeech 2020*, 2182–2186. <https://doi.org/10.21437/Interspeech.2020-2635>

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

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

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

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

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

498 Gosztolya, G., Vincze, V., Tóth, L., Pákási, M., Kálmán, J., & Hoffmann, I. (2019).
499 Identifying Mild Cognitive Impairment and mild Alzheimer’s disease based on
500 spontaneous speech using ASR and linguistic features. *Computer Speech & Language*,
501 53, 181–197. <https://doi.org/10.1016/j.csl.2018.07.007>

502 Guinn, C., Singer, B., & Habash, A. (2014). A comparison of syntax, semantics, and pragmatics
503 in spoken language among residents with Alzheimer’s disease in managed-care

504 facilities. *2014 IEEE Symposium on Computational Intelligence in Healthcare and E-*
505 *Health (CICARE)*, 98–103. <https://doi.org/10.1109/CICARE.2014.7007840>

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

510 Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*,
511 366(6461), 55–58. <https://doi.org/10.1126/science.aax0289>

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

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

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

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

- 526 Irish, M., Halena, S., Kamminga, J., Tu, S., Hornberger, M., & Hodges, J. R. (2015). Scene
527 construction impairments in Alzheimer's disease—A unique role for the posterior
528 cingulate cortex. *Cortex; a Journal Devoted to the Study of the Nervous System and*
529 *Behavior*, 73, 10–23. <https://doi.org/10.1016/j.cortex.2015.08.004>
- 530 Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of
531 Language? *Proceedings of the 57th Annual Meeting of the Association for*
532 *Computational Linguistics*, 3651–3657. <https://doi.org/10.18653/v1/P19-1356>
- 533 Jessen, F., Amariglio, R. E., van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois,
534 B., Dufouil, C., Ellis, K. A., van der Flier, W. M., Glodzik, L., van Harten, A. C., de
535 Leon, M. J., McHugh, P., Mielke, M. M., Molinuevo, J. L., Mosconi, L., Osorio, R. S.,
536 Perrotin, A., ... Subjective Cognitive Decline Initiative (SCD-I) Working Group.
537 (2014). A conceptual framework for research on subjective cognitive decline in
538 preclinical Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the*
539 *Alzheimer's Association*, 10(6), 844–852. <https://doi.org/10.1016/j.jalz.2014.01.001>
- 540 Kavé, G., & Levy, Y. (2003). Morphology in picture descriptions provided by persons with
541 Alzheimer's disease. *Journal of Speech, Language, and Hearing Research: JSLHR*,
542 46(2), 341–352.
- 543 Kumar, A., Narayanan Sundararaman, M., & Vepa, J. (2021). What BERT Based Language
544 Model Learns in Spoken Transcripts: An Empirical Study. *Proceedings of the Fourth*
545 *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*,
546 322–336. <https://doi.org/10.18653/v1/2021.blackboxnlp-1.25>
- 547 Lindsay, H., Tröger, J., & König, A. (2021). Language Impairment in Alzheimer's Disease—

548 Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous
549 Speech Through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, 13,
550 642033. <https://doi.org/10.3389/fnagi.2021.642033>

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

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

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

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

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

568 Martinez de Lizarduy, U., Calvo Salomón, P., Gómez Vilda, P., Ecañ Torres, M., & López de
569 Ipiña, K. (2017). ALZUMERIC: A decision support system for diagnosis and

570 monitoring of cognitive impairment. *Loquens*, 4(1), 037.
571 <https://doi.org/10.3989/loquens.2017.037>

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

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

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

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

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

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

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

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

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

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

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

608 Roger, E., Banjac, S., Thiebaut de Schotten, M., & Baciú, M. (2022). Missing links: The
609 functional unification of language and memory ($L \cup M$). *Neuroscience &*
610 *Biobehavioral Reviews*, 133, 104489. <https://doi.org/10.1016/j.neubiorev.2021.12.012>

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

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

615 Sarawgi, U., Zulfikar, W., Soliman, N., & Maes, P. (2020). Multimodal Inductive Transfer
616 Learning for Detection of Alzheimer’s Dementia and its Severity. *Interspeech 2020*,
617 2212–2216. <https://doi.org/10.21437/Interspeech.2020-3137>

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

621 Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang,
622 Y., Coutinho, E., & Evanini, K. (2016). The INTERSPEECH 2016 computational
623 paralinguistics challenge: 17th Annual Conference of the International Speech
624 Communication Association. *Proceedings of the Annual Conference of the
625 International Speech Communication Association, INTERSPEECH, 08-12-September-
626 2016*, 2001–2005. <https://doi.org/10.21437/Interspeech.2016-129>

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

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

634 Thakral, P. P., Madore, K. P., Kalinowski, S. E., & Schacter, D. L. (2020). Modulation of
635 hippocampal brain networks produces changes in episodic simulation and divergent

636 thinking. *Proceedings of the National Academy of Sciences*, 117(23), 12729–12740.
637 <https://doi.org/10.1073/pnas.2003535117>

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

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

644 Thomas, C., Keselj, V., Cercone, N., Rockwood, K., & Asp, E. (2005). Automatic detection
645 and rating of dementia of Alzheimer type through lexical analysis of spontaneous
646 speech. In *IEEE International Conference on Mechatronics and Automation, ICMA*
647 *2005* (Vol. 3, p. 1574 Vol. 3). <https://doi.org/10.1109/ICMA.2005.1626789>

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

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

658 World Health Organization. (2020, December 9). *The top 10 causes of death.*

659 <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>

660 Zhu, Y., Obyat, A., Liang, X., Batsis, J. A., & Roth, R. M. (2021). WavBERT: Exploiting

661 Semantic and Non-Semantic Speech Using Wav2vec and BERT for Dementia

662 Detection. *Interspeech 2021*, 3790–3794. <https://doi.org/10.21437/Interspeech.2021->

663 332

664

665 **Supplementary file description:**

666 The supplementary information is divided into six parts : (A) The recruitment procedure,

667 diagnostic criteria, neuropsychological battery, impact of the epidemic, and criteria for

668 inclusion and exclusion. (B) Speech transcription instruction. (C) Feature list and definitions.

669 (D) Detailed group-wise classifier performance from the random forest. (E) Post-hoc analysis

670 of the RMANOVA tests. (F) Results and statistical comparison from Gradient Boosting.

671

672 **Tables**673 **Table 1:** Participant demographics[†]

	HOC	SCD	naMCI	aMCI	pAD	Test	<i>p</i> 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	/	/

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

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

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

682

683 **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

684 [†]: number of dimensions of the word embeddings

685 **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
-------------	------	------	------	------	------	------	------	------

686 Note: *** F1 score \geq .9 for very good, ** F1 score \geq .85 for good, * F1 score \geq .8 for not bad

687

688 **Figures**

689 **Figure 1.** (a) Scatter plot of F1-scores across different feature sets and classification tasks; (b)
690 Means, standard deviations and distributions of the same F1-scores across classification tasks
691 with the mean line; (c) F1-scores across different feature sets with the mean line; (d) F1-scores
692 between different modalities (speech vs. text) with the mean line.
693

694 **Learning Outcomes**

- 695 • Machine learning based on automatically extracted language features detected cognitive
696 decline from early stages of the AD continuum in a new Spanish-Catalan dataset.
- 697 • Different speech and language domains showed differential discrimination performance
698 between groups, with features extracted directly from speech performing better than those
699 from the text.
- 700 • Before the onset of objective cognitive impairment, speech and language from older adults
701 with Subjective Cognitive Decline (SCD) showed speech and language differences from
702 controls without SCD, indicating potential heterogeneity in these non-clinical groups.
- 703