Artificial Intelligence, speech and language processing approaches to monitoring Alzheimer's Disease: a systematic review

Sofia **de la Fuente Garcia**[a,*], Craig **Ritchie**[b] and Saturnino **Luz**[a]

[a] *Usher Institute, Edinburgh Medical School, The University of Edinburgh, Scotland, UK*
E-mails: sofia.delafuente@ed.ac.uk, s.luz@ed.ac.uk
[b] *Centre for Dementia Prevention, The University of Edinburgh, Scotland, UK*
E-mail: craig.ritchie@ed.ac.uk

Abstract.
Background: Language is a valuable source of clinical information in Alzheimer's Disease, as it declines concurrently with neurodegeneration. Consequently, speech and language data have been extensively studied in connection with its diagnosis.
Objective: firstly, to summarise the existing findings on the use of artificial intelligence, speech and language processing to predict cognitive decline in the context of Alzheimer's Disease. Secondly, to detail current research procedures, highlight their limitations and suggest strategies to address them.
Method: Systematic review of original research between 2000 and 2019, registered in PROSPERO (reference CRD42018116606). An interdisciplinary search covered six databases on engineering (ACM and IEEE), psychology (PsycINFO), medicine (PubMed and Embase) and Web of Science. Bibliographies of relevant papers were screened until December 2019.
Results: from 3,654 search results 51 articles were selected against the eligibility criteria. Four tables summarise their findings: *study details*, (aim, population, interventions, comparisons, methods and outcomes), *data details* (size, type, modalities, annotation, balance, availability and language of study), *methodology* (pre-processing, feature generation, machine learning, evaluation and results) and *clinical applicability* (research implications, clinical potential, risk of bias and strengths/limitations).
Conclusion: promising results are reported across nearly all 51 studies, but very few have been implemented in clinical research or practice. The main limitations of the field are poor standardisation, limited comparability of results, and a degree of disconnect between study aims and clinical applications. Active attempts to close these gaps will support translation of future research into clinical practice.

Keywords: screening, Alzheimer's Disease, dementia, cognitive decline, computational linguistics, speech processing, machine learning, artificial intelligence

Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that involves decline of cognitive and functional abilities as the illness progresses [1]. It is the most common aetiology of dementia. Given its prevalence, it has effects beyond just patients and carers as it also has a severe societal and economic impact worldwide [2]. Although memory loss is often considered the signature symptom of AD, language impairment may also appear in its early stages [3]. Consequently, and due to the ubiquitous nature of speech and language, multiple studies rely on these modalities as sources of clinical information for AD, from foundational qualitative research [e.g. 4, 5] to more recent work on computational speech technology [e.g. 6–8]. The potential for using speech as a biomarker

*Corresponding author. E-mail: sofia.delafuente@ed.ac.uk. Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh EH16 4UX

Table 1
Feature taxonomy, adapted from Voleti et al. [19].

Category	Subcategory	Feature type	Feature name, abbreviation, reference.
Text-based (NLP)	Lexical features	Bag of words, vocabulary analysis	*BoW*, *Vocab*.
		Linguistic Inquiry and Word Count	*LIWC* [20]
		Lexical diversity	Type-Token Ratio (*TTR*), Moving Average TTR (*MATTR*), Simpson's Diversity Index (*SDI*) Brunét's Index (*BI*), Honoré's Statistic (*HS*).
		Lexical Density	Content density (*CD*), Idea Density (*ID*), *P*-Density (*PD*).
		Part-of-Speech tagging	*PoS*.
	Syntactical features	Constituency-based parse tree scores	*Yngve* [21], *Frazier* [22].
		Dependency-based parse tree scores	
		Speech graph	Speech Graph Attributes (*SGA*).
	Semantic features	Matrix decomposition methods	Latent Semantic Analysis (*LSA*), Principal Component Analysys (PCA).
	(Word and sentence	Neural word/sentence embeddings	*word2vec* [23]
	embeddings)	Topic modelling	*Latent Dirichlet Allocation* [24].
		Psycholinguistics	Reliance on familiar words (*PsyLing*).
	Pragmatics	Sentiment analysis	*Sent*.
		Use of language *UoL*	Pronouns, paraphrasing, filler words (*FW)*.
		Coherence	*Coh*.
Acoustic	Prosodic features	Temporal	Pause rate (*PR*), Phonation rate (*PhR*), Speech rate (*SR*), Articulation rate (*AR*). Vocalisation events.
		Fundamental Frequency	F_0 and trajectory.
		Loudness and energy	*loud*, *E*.
		Emotional content	*emo*.
	Spectral features	Formant trajectories	F_1, F_2, F_3.
		Mel Frequency Cepstral Coefficients	*MFCCs* [25].
	Vocal quality	Jitter, Shimmer, harmonic-to-noise ratio	*jitt*, *shimm*, *HNR*.
	ASR-related	Filled pauses, repetitions, dysfluencies, hesitations. fractal dimension, entropy.	*FP, rep, dys, hes, FD, entr*.
		Dialogue features (i.e. Turn-Taking)	*TT*:avg turn length, inter-turn silences.

dertaken with this dataset uses the Cookie Theft picture description and reading tasks subsets, all recorded in Swedish.

For dialogue data, the *Carolina Conversations Collection (CCC)* is the only available database. It consists of conversations between healthcare professionals and patients suffering from a chronic disease, including AD. For dementia research, participants are assigned to an AD group or a non-AD group, if their chronic condition is unrelated to dementia (i.e. diabetes, heart disease). Conversations are prompted by questions about their health condition and experience in healthcare. It is collected and distributed by the Medical University of South Carolina [54].

In addition, some of the reviewed articles refer to the *IVA dataset*, which consists of structured interviews undertaken and recorded simultaneously by an Intelligent Virtual Agent (a computer "avatar") [55]. However, the potential availability of this dataset is unknown.

2. SPICMO (PICOS) table

The first table is based on the PICOS design, widely used in the clinical field. Its columns are:

- **Population**: total number of participants followed by number of participants per group (always starting with the less impaired group). Average demographic figures (i.e. age, education, MMSE) follow the same order.
- **Interventions**: assessments that participants underwent as part of the study. These are usually either cognitive or full clinical assessments, recorded speech tasks and written tasks.
- **Comparison groups**: different stages of cognitive impairment of the study participants conform the groups to be compared. These are Healthy Controls (HC), Subjective Cognitive Impairment (SCI), Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD) and Cognitive Impairment (CI), when unspecified. This terminology is not standardised across publications, but we have standardised it for the purpose of this review. Hence, for instance, normal controls (NC), healthy elderly (HE), Subjective Memory Complaints (SMC) or Dementia (e.g. Alzheimer's Type Dementia) are hereby equated to HC, SCI and AD, respectively.
- **Outcomes of interest**: detection, prediction or discrimination performance of the method used in an article. This mostly includes classification metrics, such as overall accuracy, sensitivity and specificity.
- **Study aim/design**: most frequently, automatic detection of a target group when compared to a healthy one, or automatic discrimination between different stages of target groups. It also includes the main design, i.e. text vs. speech, narrative vs. monologue.

We have extended this by adding a column on methods, an essential part of this review:

- **Methodology**: brief overview of the approach for feature generation (i.e. acoustic analysis or natural language processing), as well as the approach for feature set reduction (when reported). These are feature selection (i.e. filtering or wrapping) and feature extraction (i.e. combination or transformation of original features, e.g. PCA, LSA, ADR). This section also mentions the machine learning task used in the paper (i.e. machine learning task).

Lastly, we considered it to be more intuitive for this review to have information about **S**tudy aim at the beginning, and therefore, the conventional order of the columns has been shifted, yielding SPICMO (study aim, population, intervention, comparisons, methodology and outcomes) as a result.

Table 5: SPICMO (PICOS) table

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Beltrami et al. [91]	Automatic detection of MCI based on acoustic and linguistic fts from narrative speech data.	39 pps: 20 HC, 19 MCI. Aged range: 50-75 years old; educ: high school/university. Italy.	Cognitive ast. **Speech task:** narratives picture description, working day, last dream.	HC and MCI. Based on cognitive ast (MMSE, MoCA, GPCog, CDT, VF, CANTAB-PAL)	Acoustic and linguistic analysis of speech for ft extraction, statistics for ft selection, ML for group classification.	Detection performance: 76.9% accuracy (picture task) ($F_1 = 78.1\%$).
Ben Ammar and Ben Ayed [95]	Automatic detection of AD based on linguistic fts from narrative speech data.	484 samples: 242 HC, AD; unreported pps from Pitt[1]. Age > 44; educ > 7; MMSE > 10. USA.	Clinical ast. **Speech task:** narrative picture description (Cookie Theft).	HC and MCI. Based on cognitive ast (i.e. MMSE).	Audio enhancement, linguistic analysis for ft extraction, ML for ft selection and group classification.	Detection performance: 79% accuracy.
Bertola et al. [57]	Discrimination between HC, MCI and AD using graph analysis on word sequences obtained from SVF data.	100 pps: 25 HC, aMCI, a+mdMCI, AD. Median age: 76, 76 (MCI), 79; educ: 4, 4, 4; MMSE: 27, 25, 20. Brazil.	Clinical ast (i.e. medical, cognitive). **Speech task:** recorded SVF answers (animals).	HC, amnesic single/multiple domain (a/+mdMCI), and AD. Based on cognitive ast (Katz, Lawton, MMSE).	Graph analysis for ft extraction from word sequences, statistics for ft selection (incl SVF scores), ML for group classification.	Discrimination performance: $AUC = 0.68$ HC-MCI, $AUC = 0.73$ MCI-AD, $AUC = 0.88$ HC-AD (graph attributes only).
Chien et al. [82]	Automatic detection of AD based on acoustic fts from narrative speech and SVF data.	60 pps: 30 HC, 30 AD from Mandarin_Lu[2]. 150 speech samples, demographics unreported. China, Taiwan.	**Speech task:** recorded SVF answers (fruits, locations) and narrative picture description.	HCand AD. Unreported criteria.	Manual acoustic analysis for ft extraction, ML for group classification.	Detection performance: $AUC = 0.95$.
Clark et al. [67]	Automatic prediction of MCI conversion to AD from automatic SVF scores combined with neuroimaging data.	107 pps: 83 MCI-non (46F), 24 MCI-con (15F). Avg age: 68.7, 73.8; educ: 16, 16; MMSE: 27.9, 25.1. USA.	Cognitive ast, brain MRI. **Speech task:** recorded SVF (vegetables, animals) and OVF (letters F, A, S).	MCI-non and MCI-con (upon conversion to AD). Based on Petersen criteria (incl. MMSE, CDR).	Automatic scoring of verbal fluency answers (electronically transcribed) and neuroimaging scores for ft extraction, ML for group classification.	Conversion prediction performance (at 4-year follow-up): $AUC = 0.872$ with automatic scores.
D'Arcy et al. [93]	Detection of probable CI based on manually and automatically extracted temporal speech fts.	87 pps: 50 HC 37 probable CI. Age range: 62-92, 62%F, MMSE over 24 (excl severe CI). Ireland.	Cognitive ast . **Speech task**: narrative picture descriptions, Scribe[3], SVF (animals), word list, Heidi passage.	HC (MMSE > 27) and probable CI (MMSE $\leqslant 27$).	Manual and automatic temporal analysis of speech for ft extraction (syntactic, acoustic), ML for group classification.	Detection performance: 76.74% accuracy (manual approach far superior than automatic).Vowel 17% longer in probable CI.
Dos Santos et al. [90]	Automatic detection of MCI based on speech fts from transcribed narratives in English and Portuguese.	86 Pitt[4] pps. USA. 40 Cinderella[5] pps. USA. 43 ABCD[6] pps. Brazil.	Cognitive ast. **Speech task:** picture description (Pitt), story retelling (Cinderella), recall (ABCD).	HC and MCI. Based on cognitive ast (Pitt), Petersen criteria (Cinderella), clinical diagnosis (ABCD).	Topological network and linguistic analysis for ft extraction, ML for group classification.	Detection performance: 65% Pitt accuracy, 65% Cinderella, 75% ABCD.
Duong et al. [69]	Description of discourse patterns and heterogeneity in AD based on transcribed narrative speech.	99 pps: 53 HC (40F), 46 AD (39F). Avg age: 73.8, 74.3; educ: 10.2, 8.3. Canada.	Cognitive ast. **Speech task:** P1 and P7 narrative picture descriptions from PENO[7].	HC (NE: normal elderly) and AD. Based on NINCDS-ADRDA criteria [39].	Discourse analysis of the transcribed narratives, cluster analysis to group pps with similar discourse patterns.	Clusters inconclusive for prototypical AD discourse (heterogeneity). 4 different discourse patterns for P1 and 5 for P7.
Egas López et al. [62]	Discrimination between HC, MCI and AD using the i-vector approach on acoustic fts from narrative speech.	75 pps: 25 HC, MCI, AD. Avg age: 73.96, 72.4, 70.72; educ: 10.76, 10.84, 12.08; MMSE: 29.24, 27.16, 23.92.	Cognitive ast. **Speech task:** narrative of previous day, immediate and delayed recall of 2 short films.	HC, MCI and AD. Based on cognitive ast (i.e. MMSE, CDT, ADAS-Cog)	Acoustic analysis of speech recordings ft extraction, i-vector approach for dimensionality reduction, ML for group classification.	Discrimination performance: 56% accuracy and $F_1 = 78.4\%$, all tasks. $F_1 = 79.2\%$ immediate recall only.

[1] This paper reports the number of speech samples they used, but not to how many pps they belonged to.
[2] Mandarin_Lu corpus is hosted within DementiaBank
[3] 5 pictures from an in-house task (Picture Taboo). Scribe consists of sentences designed to cover English language phones.
[4] 86 pps: 43 HC (20F), MCI (16F). Avg age: 64.1, 69.3; educ over 7, MMSE over 10.
[5] 40pps: 20 HC (16F), MCI (14F). Avg age: 74.8, 73.3; educ: 11.4, 10.8.
[6] 43 pps: 20 HC, 23 MCI. Avg age: 61, 72; educ: 16, 13.3.
[7] PENO is a cognitive battery in French [32]. "Bank robbery" and "Car accident" are language subtests from this battery.

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Espinoza-Cuadros et al. [83]	Automatic detection of MCI based on speech fts from interviews and narratives.	19 pps: 11 HC (6F), 8 MCI (2F). Avg age: 78.9, 80.3; educ: 8, 5	**Speech task:** structured interview recorded from MEC[8] and a reading short passage[9].	HC and MCI. Unreported criteria.	Acoustic analysis of speech recordings for ft extraction, statistics for ft selection, ML for group classification.	Detection performance: 78.9% accuracy with seven prosodic features from the passage reading task.
Fraser et al. [7]	Automatic detection of MCI based on multi- modal fts from language tasks (audio, text, eye-tracking).	55 pps: 29 HC (21F), 26 MCI (14F). Avg age: 67.8, 70.6; educ: 13.3, 14.3; MMSE: 29.6, 28.2.	Cognitive ast. **Speech task:** picture description (Cookie Theft), read short text[10] aloud and in silence.	HC and MCI. Based on Petersen criteria. Gothemburg MCI Study.	Multi-modal approach for ft extraction, cascaded ML for group classification (ft, mode, task and session).	Detection performance: 83% accuracy (AUC$=0.88$), with multi-modal fts. 84% accuracy (AUC$=0.90$) incl cognitive scores.
Fraser et al. [87]	Automatic detection of MCI based on topic modelling and information content in Swedish and English.	In-domain[11]: 67 pps, Gothenburg ds, Sweden. Out-of-domain[12]: 96 pps, Karolinska ds Sweden; 78 pps, Pitt, USA.	Cognitive ast. **Speech task:** picture description (Cookie Theft) spoken or written (Karolinska ds).	HC and MCI. Based on cognitive ast and clinical diagnosis.	Linguistic analysis for ft extraction, multilingual topic models for dimensionality reduction and ft seletion, ML for group classification.	Detection performance: 63% accuracy in English; 72% accuracy in Swedish. Based on information content.
Fraser et al. [9]	Automatic detection of AD based on linguistic (mostly) and acoustic fts from narrative speech.	473 samples[13]: 233 HC (151F), 240 AD (158F). Avg age: 65.2, 71.8; educ: 14.1, 12.5, MMSE: 29.1, 18.5.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC and AD. Based on cognitive ast (i.e. MMSE).	Acoustic and linguistic analysis of speech for ft extraction, factor analysis for dimensionality reduction, ML for group classification.	Detection performance: 81.92% accuracy. Factors identified (4): semantic, acoustic, syntactic and information.
Gonzalez-Moreira et al. [89]	Automatic detection of mild dementia based on acoustic fts from narrative speech in Spanish.	20 pps: 10 HC (1F), 10 CI[14] (4F). Avg age 78.9, 80.3; educ 7.8, 4.	Cognitive ast. **Speech task:** read short text aloud ("The Grandfather Passage".	HC and CI. Based on cognitive ast (i.e. MEC scores).	Acoustic analysis for ft extraction of recorded speech, ML for group classification.	Detection performance: 85% accuracy based on four prosodic fts (articulation rate, mean syllables duration, F0 sd and mean).
Gosztolya et al. [63]	Automatic detection of AD and MCI, as well as discrimination between HC, MCI and AD based on acoustic (ASR) and linguistic speech fts.	75 pps: 25 HC, MCI, AD. Avg age: 70.72, 72.4, 73.96; educ: 12.08, 10.84, 10.7; MMSE: 29.24, 27.16, 23.92. 225 recordings.	Cognitive ast. **Speech task:** narrative of previous day, immediate and delayed recall of 2 short films.	HC, MCI and AD groups. Based on cognitive ast (i.e. MMSE, CDT, ADAS-Cog).	Acoustic (ASR) and linguistic analysis for ft extraction, ML for group classification.	Detection performance: 86% HC-AD, 80% HC-MCI, 81.3% HC-CI accuracy. Discrimination performance: 66.7% (morphologic and acoustic).
Guinn et al. [98]	Automatic detection of AD based on linguistic fts from dialogue transcripts.	56 pps from CCC[15]: 28 nonAD, 28 AD. Multiple transcripts per pp: 204 nonAD, 77 AD.	**Speech task:** conversational interview about pp's chronic condition and their experience in healthcare.	HC (nonAD: patients with chronic conditions unrelated to AD) and AD. Based on clinical diagnosis.	Linguistic analysis for ft extraction of dialogue transcripts (syntax, semantics, pragmatics), ML for gorup classification.	Detection performance: $Prec_{AD} = 80.8\%$, $recall_{AD} = 0.75\%$; $Prec_{nonAD} = 79.3\%$, $recall_{nonAD} = 82.1\%$.

[8] Mini-Examen Cognoscitivo (MEC) is the Spanish adaptation of the MMSE[34].

[9] In particular, a Spanish version of "The Grandfather Passage"Darley et al. [48].

[10] Short texts obtained from the International Reading Speed Texts (IReST).

[11] Data including MCI pps: 67 Gothenburg. 36 HC (23F), 31 MCI (16F). Avg age: 67.9, 70.1; educ: 13.1, 14.1; MMSE: 29.6, 28.2.

[12] Data not including MCI pps: 96 Karolinska (52F) and 78 Pitt (48): all HC. Avg age: 57.2, 63.9; educ: 13, 13.9; MMSE: N/A, 29.1.

[13] Pitt: repeated samples from 264 participants (97 HC, 167 AD).

[14] This category is named MD (mild dementia) in study.

[15] Carolina Conversations Collection (CCC) [54] is conversational corpus consisting of conversations about health and healthcare gathered longitudinally with people with different chronic conditions, including AD.

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Guo et al. [92]	Automatic detection of AD based on linguistic fts from narrative speech.	268 pps[16]: 99 HC (58F), 169 AD (114F). Avg age: 61.3, 71; educ: 13.3, 11.8; MMSE: 27.9, 18.7.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC and AD. Based on cognitive ast (i.e. MMSE).	Linguistic analysis for ft extraction (phonetics, semantics, syntax, pragmatics) including perplexity, ML for group classification.	Detection performance: 85.4% accuracy including perplexity fts derived from language models.
Haider et al. [11]	Automatic detection of AD based on standardised acoustic ft sets extracted from narrative speech.	164 pps (Pitt): 82 HC (46F), AD (46F). Aged 50-80, mostly 65-75 (BWGB); educ over 7 years: MMSE over 10.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HCand AD. Based on cognitive ast (i.e. MMSE).	Acoustic analysis for ft extraction from the recorded narratives, ADR[17] for ft selection and representation, ML for group classification.	Detection performance: 71.34% accuracy with one ft set; 78-80% accuaracy with "hard fusion" of all ft sets.
Kato et al. [64]	Discrimination between HC, MCI and AD with a two-phase system based on speech fts cerebral blood flow.	48 pps: 20 HC (13F), 19 MCI (13F), 9 AD (4F). Age range: 64-92 years old. Other demographics unreported.	Cognitive ast. **Speech task:** topics hometown and childhood, HDS-R, memory tasks. Simultaneous fNIRS[18].	HC, MCI and AD. Based on cognitive ast. (i.e. CDR = 0, 0.5 or 1).	Multivariate statistics to generate cognitive rating (SPCIR) based on prosody, ML for group classification in two phases: fNIRS and prosody.	Discrimination performance: 85.4% overall accuracy. 32% MCI participants misclassified into HC group.
Khodabakhsh and Demiroğlu [84]	Automatic detection of AD based on acoustic fts from conversational speech.	54 pps: 27 HC (15F), 27 AD (10F). Age range: 60-80 years old. Other demographics unreported.	**Speech task:** pps were asked casual questions to elicit 10 min of spontaneous conversation.	HC and AD. Unreported criteria.	Voice activity detection (VAD) and acoustic analysis for ft extraction from recordings, ML for group classification.	Detection performance: 79.2% with best pair of fts (log of voicing ratio + avg absolute delta pitch.
Konig et al. [102]	Discrimination between HC, MCI and AD based on automatically extracted speech fts across different tasks.	Dem@care: 64 pps. 15 HC (9F), 23 HC (12F), 26 AD (13F). Avg age 72,73,80; educ[19]: uni, col, hs; MMSE: 29,26,19.	Cognitive ast. **Speech task:** countdown, picture description, repetition, SVF (animals).	HC, MCI and AD. Based on subjective memory complain (HC), Petersen criteria (MCI), NINCDS-ADRDA (AD)	Acoustic analysis for ft extraction from speech recordings, statistics for ft selection, ML for group classification.	Detection performance: $EER_{HC-MCI} = 21\%$, $EER_{HC-AD} = 13\%$, $EER_{MCI-AD} = 20\%$. Equal ss-sp: 79%, 87%, 80%
Lopez-de Ipiña et al. [80]	Discrimination between HC and stages of AD based on acoustic fts, incl. emotional response (pilot study).	10 pps (AZTITXIKI[20]): 5 HC/AD (2F). AD: 1ES, 2SS, 2AS. Age label: middle (HC), elderly (HC and AD).	**Speech task:** telling pleasant stories, recounting pleasant feelings, conversational interaction.	HCand ES, IS, AS, which stand for early, intermediate and advanced AD. Unreported criteria.	Acoustic analysis and emotional response analysis (ERA) for ft extraction, ML for group classification.	Discrimination performance: 93.79% accuracy with speech and emotional fts. Unclear whether this result is on 4 groups or 2.
Lopez-de Ipiña et al. [79]	Discrimination between HC and stages of AD based on acoustic fts, incl. emotional response.	40 pps (AZTIAHORE[21]): 20 HC (10F), 20 AD (12F). AD: 4ES, 10SS, 6AS. Age range: 20-98 HC, 68-98 AD. Others unreported.	**Speech task:** telling pleasant stories, recounting pleasant feelings, conversational interaction.	HC[22] and ES, IS, AS, which stand for early, intermediate and advanced AD. Unreported criteria.	Acoustic analysis and fractal dimension for ft extraction, incl. emotional response, ML for group classification.	Discrimination performance: accuracy reported per class, avg. 96.89%. Overall performance unclear.
Lundholm Fors et al. [59]	Discrimination between HC, SCI and MCI based on syntactic fts extracted from narrative speech.	Gothemburg[23]: 90pps. 36 HC (23F), 23 SCI (14F), 31 MCI (16F). Avg age: 67.9, 66.3, 70.1; educ: 13.2, 16.1, 14.1; MMSE: 29.6, 29.5, 28.2.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC, SCI and MCI. Based on clinical diagnosis.	Linguistic analysis for syntactic ft extraction, statistical analysis for feature selection, and ML for group classification.	Discrmination performance (binary detection): HC-MCI: $F_1 = 0.68$, HC-SCI: $F_1 = 0.54$, SCI-MCI: $F_1 = 0.66$

[16] subset from Pitt with multiple samples per participant. This study used 498 samples: 242 HC, 256 AD.
[17] Active Data Representation: novel method presented in this paper.
[18] Functional near-infrared spectroscopy: measures cortical brain activity by monitoring changes of oxy/deoxygenated hemoglobin concentration.
[19] Mode, that is, most frequent educational category.
[20] Subset of AZTIAHORE, which is, in turn, a subset of AZTIAHO.
[21] which is, in turn, a subset of AZTIAHO
[22] Group annotated as CR (control group) in the paper, equated to HC for the purpose of this review.
[23] Gothenburgh MCI data set Wallin et al. [53]

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Luz [6]	Automatic detection of AD based on acoustic fts extracted directly from voice recordings of narrative speech.	398 recordings (Pitt): 184 HC, 214 AD. Other pp information unreported.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC and AD. Based on cognitive ast (i.e. MMSE).	Acoustic analysis to extract paralingusitic fts directly from speech recordings, no ft selection, ML for group classification.	Detection performance: 68% accuracy (baseline classification with simple algorithms for voice activity detection and speech rate).
Luz et al. [10]	Automatic detection of AD based on dialogical, content-free fts extracted from transcripts.	38 pps (CCC, ibid.) 21 nonAD (12F), 17 AD (15F). Age over 65. Other demographics unreported.	**Speech task:** conversational interview about pp's chronic condition and their experience in healthcare.	HC (nonAD: patients with chronic conditions unrelated to AD) and AD. Based on clinical diagnosis.	Dialogue analysis for ft extraction from conversational transcripts, Markov chains for data representation, ML for group classification.	Detection performance: 86.5% accuracy with vocalisations and speech rate.
Martinez de Lizarduy et al. [60]	Automatic detection of MCI and AD based on acoustic fts with a novel decision support system (ALZUMERIC).	SVF ds: 62 HC (36F), 38 MCI (21F). Avg age: 56.73, 57.15. PD ds[24]: 12 HC, 6 AD. SS ds (AZTIAHORE): 20 HC (9F), 20 AD (12F).	**Speech tasks**: SVF (animals), picture description (PD), spontaneous speech (SS, see AZTIAHORE).	SVF: HCand MCI. PD: HC and AD. SS: HC and AD. Unreported criteria.	ALZUMERIC system: acoustic analysis for ft extraction from voice samples, automatic ft selection, ML for group classification.	Detection performance: HC-MCI (SVF): 80% accuracy. HC-AD: 94% (PD) and 95% accuracy.
Meilan et al. [100]	Automatic detection of AD based on temporal and acoustic speech fts.	66 pps: 36 HC (80%F) 30 AD (68%FF). Avg age: 74.06, 78.66; educ: 7.30, 6.27; MMSE: 27.97, 18.07.	Cognitive ast. **Speech task:** reading familiar sentences on screen.	HC and AD. Based on NINCDS-ADRDA and cognitive ast (i.e. GDS, MMSE).	Acoustic and temporal speech analysis for ft extraction, ML for group classification.	Detection performance: 83.3% accuracy with speech fts such as voice breaks.
Mirheidari et al. [86]	Discrimination between ND and FMD[25], based on doctor-patient conversational fts.	30 pps: 15 FMD (9F), 15 ND[26] (8F). Avg age: 57.8, 63.73; MMSE: 28.87, 18.79.	Cognitive ast. **Speech task:** neurology consultation (conversation).	FMD and ND. Based on Schmidtke et al. [40] (FMD), Petersen and NINCDS-ADRDA (ND)	Automatic conversation analysis for ft extraction, ML for ft selection and group classification.	Discrimination performance: 97% classification accuracy between FMD and ND with top-10 fts.
Mirheidari et al. [85]	Discrimination between different conditions based on linguistic fts from conversational speech.	IMDB[27]: 50000 entries. Pitt: 473 narratives. Hallam: 45 conversations. IVA: 18 conversations. Seizure: 241 conversations.	**Written task:** movie feedback (IMDB). **Speech task:** picture description (Pitt), neurology consultation.	Pitt: HC, AD. Hallam: FMD, ND, DPD[28]. IVA: FMD, ND, MCI. Unreported criteria.	ASR and linguistic analysis (vector representation) for ft extraction and selection, ML for group classification.	Discrimination performance: 65.8% (Hallam), 70% (IVA). Best binary: 93.7% FMD-DPD (Hallam), 100% FMD-ND (IVA).
Mirheidari et al. [8]	Automatic detection of ND based on speech fts, comparing neurologist-led with virtual-agent-led interactions (IVA).	HUM, 30pps: 15 FMD (9F), 15ND[29] (8F). Age: 57.8, 63.73; MMSE: 28.87, 18.79. IVA, 12 pps: 6 FMD (1F), 6 ND (3F). Avg age: 55.67, 65.83, ACE-R: 83.67, 59.57.	Cognitive ast. **Speech task:** neurology consultation (HUM) or avatar interaction (IVA).	FMD and ND. Based on Schmidtke et al. [40] (FMD), Petersen and NINCDS-ADRDA (ND)	Acoustic, linguistic and conversational analysis of recordings, comparing IVA-patient with neurologist-patient interactions. ML for group classification.	Detection performance: Neurologist-patient: 90.0% accuracy. IVA-patient: 90.9% accuracy.
Mirheidari et al. [12]	Discrimination between HC, FMD, MCI and ND based on acoustic and linguistic fts from IVA-led speech.	61 pps: 14 HC (8F), 10 FMD (6F), 18 MCI (12F), 19 ND (7F). Avg age: 69.4, 56.4, 62.2, 69.8.	Cognitive ast. **Speech task:** SVF (animals) and 10 question conversations. Both led by an IVA.	HC, FMD, MCI and ND. Based Schmidtke, (FMD), Petersen (MCI) and NINCDS-ADRDA (ND) criteria.	Acoustic, linguistic and conversational analysis of SVF answers and IVA-patient interactions. ML for ft selection and group classification.	4-way discrimination performance: 62% accuracy and ROC-AUC: 81.5% with top 22 fts (48% all).
Mirzaei et al. [49]	Discriminate between HC, MCI and AD based on acoustic fts from narrative speech.	48 pps: 16 HC, 16 MCI, 16 AD. Avg age: 72.7, 77.6, 77.9; MMSE: 28.6, 28.3, 22.4.	Cognitive ast **Speech task:** reading familiar sentences on screen.	HC, MCI, AD. Based on cognitive ast ((MMSE over 20 for inclusion).	Acoustic analysis for ft extraction, ML for ft selection (wrapper) and group classification.	Discrimination performance: 62% accuracy (three-way classification).

[24] Both are subsets from the Gipuzkoa-Alzheimer Project: http://www.cita-alzheimer.org/projects/gipuzkoa-alzheimer-project-basque-cohort

[25] ND: neurodegenerative disorder (e.g. AD). FMD: functional memory disorder)

[26] Heterogeneous ND group: 8 AD, 3 AD+vD, 2 MCI and 2 FTD (frontotemporal dementia).

[27] DS details. IMBD: text entries on movies feedback. Pitt (previously described). Hallam [86], IVA [55] and Seizure [125]

[28] DPD: depressive pseudo-dementia.

[29] Heterogeneous ND group: 8 AD, 3 AD+vD, 2 MCI and 2 FTD (frontotemporal dementia).

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Nasrolahzadeh et al. [65]	Discriminate between HC and three stages of AD based on higher-order spectral analysis of speech data.	60 pps[30]: 30 HC (15F), 6 FS (3F), 15 SS (6F), 9 TS (5F). Avg age: 75.6, 73.3, 70.6, 77.4; MMSE: 28.39, 27.5, 26.8, 23.8.	Clinical ast (i.e. cognitive, medical). **Speech task:** prompted to talk about personal stories and feelings.	HC and AD, subdivided in FS, SS and TS. AD subgroups diagnosed with NINCDS-ADRDA criteria.	Acoustic spectral analysis for nonlinear feature extraction, ML for for ft selection and group classification.	Discrimination performance: 97.71% accuracy with 4-way classifier based on higher-order spectral fts.
Orimaye et al. [75]	Automatic detection of AD based on linguistic fts extracted from narrative speech.	198 pps (Pitt): 99 HC and 99 AD. Avg age: 65.26, 70.45. Other demographics unreported.	**Speech task:** narrative picture description (Cookie Theft).	HC and ADUnreported criteria.	Linguistic analysis for ft extraction, statistical analysis for ft selection, ML for group classification.	Detection performance: $AUC = 0.93$ with 1000 top combined fts (syntacic, lexical, n-grams).
Prud'Hommeaux and Roark [73]	Detection of MCI based on automatic alignment scores between transcribed recall tasks and source narratives.	124 pps: 52 HC, 72 MCI. Demographics unreported.	Cognitive ast. **Speech task:** immediate and delayed recall of the Anna Thomson story (LM-WMS-III[31]).	HC (non-MCI) and MCI. Based on cognitive ast (i.e. CDR=0.5 for MCI diagnosis).	Linguistic analysis and text alignment for automatic scoring of recall tasks, ML for group classification.	Scoring performance: $F_1 = 0.791$ (alignment based). Detection performance: $AUC = 0.795\%$).
Prud'hommeaux and Roark [70]	Detection of MCI based on automatic alignment scores between transcribed recall tasks and source narratives.	235 pps: 163 HC, 72 MCI. Avg age: 87.3, 88.7. Avg educ: 15.1, 14.9. Gender unreported.	Cognitive ast. **Speech task:** immediate and delayed recall of the Anna Thomson story (LM-WMS-III).	HC (non-MCI) and MCI. Based on cognitive ast (i.e. CDR=0.5 for MCI diagnosis).	Linguistic analysis and graph-based text alignment for automatic scoring of recall tasks, ML for group classification.	Scoring performance: $F_1 = 0.891$. Detection performance: $AUC = 0.748$. Pitt[32]: $AUC = 0.704$
Rentoumi et al. [88]	Automatic detection of AD based on linguistic fts extracted from written narrative data.	60 pps: 30 HC (14F), 30 AD (17F). Avg age: 68.03, 66.48; educ: 13.93, 12; MMSE: 28.26, 22.68.	Cognitive ast. **Written task:** narrative picture description (Cookie Theft).	HC (NC) and AD. Based on cognitive ast. ($MMSE_{AD} = 10 - 25$).	Computational linguistic analysis for text ft extraction (morphosyntactic, lexical). ML for group classification.	Detection performance: 80% accuracy. (88.5% with synthetically enlarged ds).
Roark et al. [78]	Automatic detection of MCI based on scores and speech fts from recorded cognitive tests.	74 pps: 37 HC, 37 MCI. Avg age: 88.8, 89.8; educ: 15.1, 14.5; MMSE: 28.2, 26.4. Gender unreported.	Cognitive ast. **Speech task:** immediate and delayed recall of the Anna Thomson story.	HC and MCI. Based on cognitive ast (i.e. CDR=0.5 for MCI diagnosis).	Linguistic and acoustic analysis for ft extraction, statistical analysis for ft selection, ML for group classification.	Detection performance: $AUC = 0.861$ (test scores and automatically derived speech and language fts).
Rochford et al. [76]	Automatic detection of CI based on pause distribution fts from narrative speech.	187 pps: 150 HC, 37 CI. Avg age (all): 72.44; MMSE: 27.68, 114 females. educ unreported.	Cognitive ast. Speech task: ready aloud a passage from a children's story.	HC and CI. Based on cognitive ast. ($MMSE_{HC} >= 27$, $MMSE_{CI} < 27$).	Linguistic and acoustic analysis for ft extraction, statistical analysis for ft selection, ML for group classification.	Detection performance: 68.66% acc ($AUC = 0.74$).
Sadeghian et al. [43]	Automatic detection of AD based on acoustic and linguistic fts from narrative speech, and with MMSE scores.	72 pps: 46 HC, 26 AD. Avg age: 71.43, 78.48; educ: 13.28, 13.81; MMSE: 28.70, 20.92.	Cognitive ast. **Speech task:** narrative picture description (new picture[33]).	HCand AD. Based on medical diagnosis.	Customised ASR for speech transcription, acoustic and linguistic analysis ft extraction, ML for ft selection and group classification.	Detection performance (acc): 88.3%: demogr.+acoustic. 91.7%: ASR linguistic. 94.4%: MMSE+manual ling
Satt et al. [61]	Discrimination between HC, MCI and AD based on acoustic fts (content-free) from voice recordings.	Dem@care: 89 pps. 19 HC (15F), 43 MCI (31F), 27 AD (24F). Avg age: 67, 73, 72.	**Speech task:** narrative picture description, sentence repetition (15), syllable repetition ("pa-ta-ka").	HC, MCI and AD. Based on medical diagnosis.	Speech segmentation (VAD), acoustic analysis for ft extraction, statistical analyis for ft selection, ML for group classification.	Discrimination performance: $EER_{HC-MCI+AD} = 18\%$ $EER_{HC-MCI} = 17\%$ $EER_{HC-AD} = 15.5\%$
Shinkawa et al. [71]	Automatic detection of MCI based on single modality and multimodal behavioural data (gait and speech).	34 pps: 19 HC (12F), 15 MCI (8F). Avg age: 71.63, 74.87; MMSE: 28.42, 25.33.	Clinical ast. **Speech task:** narrative picture description (Cookie Theft). Gait task: 5-meter walk.	HC and MCI. Based on Petersen criteria.	ASR for speech transcription, gait and linguistic analysis for ft extraction, statistic analysis and ML for ft selection and group classification.	Detection performance: Multimodal: 82.4% acc. Single modality: 76.6% acc each ($F_{1speech} = 0.733$, $F_{1gait} = 0.667$).

[30] 30 AD participants distributed in three levels: First Stage (FS), Second Stage (SS) and Third Stage (TS).)

[31] Logical Memory Test of the Weschler Memory Scale III [45]

[32] This is an attempt from to authors to apply their model on unseen data. Only results based on their graph-based method are reported here.

[33] https://acoustics.org/wp-content/uploads/2015/10/Sadeghian-Figure1b.jpg

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Tanaka et al. [94]	Automatic detection of CI based on audiovisual fts from dialogues with a computer avatar.	29 pps: 15 HC (4F), 14 CI[34] (4F). Avg age: 74.1, 76.3; educ: 10.5, 14.1; MMSE: 27.5, 21.4.	Cognitive ast. **Speech task:** 10-15 min interaction with an avatar (dialogue system).	HC and CI. Based on medical diagnosis.	Acoustic, linguistic and image analysis for ft extraction, statistics for ft selection, ML for group classification.	Detection performance: 83% unweighted acc. $AUC = 0.93$. $AUC_{ADonly} = 0.89$.
Thomas et al. [66]	Discrimination of different stages of CI based on linguistic fts from interviews.	95 pps (ACADIE[35]): 85 high, 73 low; 35 normal, 50 mild, 53 moderate, 20 severe.	Cognitive ast. **Speech task:** two interviews about donepezil, 12 weeks apart.	Two or four groups. Based on MMSE scores (0-15, 16-20, 21-24 and 25-30).	Linguistic analysis for ft extraction (lexical, n-gram), statistics for ft selection, ML for group classification.	Discrimination performance: 95% acc severe vs. normal. 69.6% moderate vs. mild. 50% 4-way classification.
Tóth et al. [99]	Automatic detection of MCI based on acoustic fts from narrative speech	84 pps: 36 HC (23F), 48 MCI (32F). Avg age: 64.13, 73.08; educ: 12.47, 11.82; MMSE: 29.17, 26.97;	Cognitive ast. **Speech task:** previous day, immediate/delayed recall of 2 short films.	HC and MCI. Based on cognitive ast (i.e. MMSE, CDT, ADAS-Cog)	Customised ASR and acoustic analysis for ft extraction, statistics for ft selection, ML for gorup classification.	Detection performance: 75% acc automatic procedure ($F_1 = 0.788$, $AUC = o.676$) .
Tröger et al. [77]	Automatic detection of AD based on acoustic fts from narrative speech.	Dem@Care: 115 pps. 47 HC (40F), 68 AD (38F). Avg age: 72.4, 78.9.	Cognitive ast. **Speech task:** two life events, previous day, picture description.	HC and AD[36] Based on medical diagnosis.	Acoustic signal processing for ft extraction, univariate ft selection, ML for group classification.	Detection performance: 89% acc relying solely on vocal fts (ASR and content-free).
Tröger et al. [96]	Discrimination between SCI, MCI and AD with a simulated telephone-based SVF test (feasibility study).	166 pps: 40 SCI (32F), 47 MCI (24F), 79 AD (40F). Avg age: 72.65, 76.59, 79; educ: 11.35, 10.81, 9.47; MMSE: 28.27, 26.02, 18.81;	Clinical ast. **Speech task:** recorded SVF answers (animals).	SCI, MCI and AD. Based on subjective reports (SCI), Petersen (MCI) and NINCDS-ADRDA (AD)	ASR, acoustic and linguistic analysis for ft extraction, ML for group classification.	ASR performance: VFER[37] $= 33.4\%$. $AUC = 0.855$
Weiner et al. [74]	Automatic detection of AD based on acoustic fts from conversational speech in German.	ISLE: 74 pps[38]. 98 samples: 80HC, 13 AACD. 5AD. Age range: 70-74 years old.	Clinical ast. **Speech task:** semi-standardized biographic interviews.	HC, AACD (ageing associated cognitive decline) and AD. Based on medical diagnosis.	Acoustic analysis for ft extraction (focus on pause patterns), ML for group classification.	Detection performance: 85.7% acc. $UAR = 0.66$ $F_{1_{HC}} = 0.92$,$F_{1_{AD}} = 0.80$,$F_{1_{AACD}} = 0.80$.
Weiner and Schultz [68]	Automatic prediction of the development of CI from conversational speech in German.	ISLE: 51 pp. 35 HC, 16 CI[39] (developed within three visits). Age range: 61-77 (1st-3rd visit).	Clinical ast. **Speech task:** semi-standardized biographic interviews.	*No change* (HC) and *Change* (CI). Based on whether they remained healthy or not.	VAD and acoustic analysis for ft extraction (focus on pause patterns), ML for group classification.	Prediction performance: 80.4% acc (overall). $Rc_{no-change} = 0.91$, $Rc_{change} = 0.56$
Yu et al. [97]	Automatic detection of CI based on speech fts collected through remote assessments.	ADCS[40]: 167 pps. Pre-processed 180 samples: 160 HC, 20 CI.	Clinical ast. **Speech task:** SVF and EBi/EBo[41] delivered by telephone system.	HC and CI. Based on longitudinal medical diagnosis.	Acoustic analysis for ft extraction (articulatory, phonemic), ML for ft selection and group classification.	Detection performance: Speech+scores: AUC= 0.77 Speech only: AUC= 0.74 Scores only: AUC= 0.54

[34] Heterogeneous CI group: 9 AD, 1 NPH (normal pressure hydrocephalus), 1 AD+NPH, 1 DBL (dementia with Lewy bodies)
[35] ACADIE study [126]. Pps are divided by MMSE in either 2 (high, low) or four groups (normal, mild, moderate, severe)
[36] Heterogeneous group: diagnosed with either AD or a form of mixed dementia (including AD).
[37] VFER: Verbal Fluency Error Rate
[38] subset from ILSE: Interdisciplinary Longitudinal Study on Adult Development and Aging [127].
[39] Heterogeneous group: AACD, MCD (mild cognitive disorder), AD, VAD (vascular dementia)
[40] ADCS: Alzheimer's Disease Cooperative Study. 4-year longitudinal data collection for home-based assessment.
[41] East Boston Immediate/Delayed: summarise a story immediately/delayed after listening to it).

2.1. Data details table

This table accounts for details of the datasets, as well as specific subsets, used in the reviewed studies. It is structured as follows:

- **Data set size**: number of participants or samples, including details on number of words, or number of hours recorded, when available.
- **Data type**, with two distinctions: a) writings, audio recordings and/or transcripts (abbreviated as per Table 3); b) monologues or dialogues. Monologues, in turn, are divided into spontaneous, narratives and answers to cognitive tests (most frequently fluency task), whilst dialogues are subdivided into three groups: structured, semi-structured and conversational. When available, information about transcription (i.e. software used, manual vs. automatic) is included.
- **Other modalities**: such as video, cognitive scores or motor measurements, when applicable ("NA" is written otherwise).
- **Data annotation**: group labels available in the data, corresponding with what was described in the comparison groups column of the SPICMO table. It includes groups' *n*, i.e. group size, as well as groups' *m*, i.e. number of speech/test samples per group, as sometimes these two figures differ (e.g. in longitudinal studies).
- **Data balance**: whether the dataset or subset used in the study is balanced in terms of age, gender and education. It accounts for dataset balance, within class balance and between class balance when applicable (see "keys to table interpretability", above, for acronyms). If a feature is not reported in the table, this is because it was not reported in the article.
- **Data availability**: whether the data used in the study is available to the wider research community.
- **Language**: language in which the dataset was collected, including country of origin, since many languages are spoken in more than one country.

Names of particular datasets are underlined (e.g. Pitt) The second table aims to provide the community with benchmark information about current databases and their availability, in order to highlight recurrent gaps that future research projects should target when designing their data collection procedures.

Table 6: Detailed Data information

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Beltrami et al. [91]	39 pps.	Narrative monologues. Rec. and manual tr. (Transcriber[42]).	Cognitive scores: MMSE, MoCA, GPCog, CDT, VF.	HC (CON) / MCI: $n = 20/19$.	CB. Text reports balanced demogr., but no figures.	Unreported.	Italian (Italy).
Ben Ammar and Ben Ayed [95]	Pitt: $m = 484$. No. of pps unreported.	Narrative monologues. Rec. and manual tr. (CHAT[43]).	Cognitive scores: MMSE.	HC / AD (Dementia): $m = 242/242$.	CB. Demogr. unreported.	Pitt avail. (DementiaBank). Enhanced unreported	English (US).
Bertola et al. [57]	100 pps.	Monologues: fluency task. Rec. (unclear).	Cognitive scores: MMSE, Katz, Lawton, SVF.	HC (NC), aMCI, a+mdMCI[44], AD: $n = 25$	no-CB/CB (ibid.). no-WCGB: HC. BCB: A, G, E.	Unreported.	Portuguese (Brazil).
Chien et al. [82]	60 pps: 30 HC *ad-hoc*, 30 AD (Mandarin_Lu). 3 tasks each: $m = 150$.	Narrative monologues and fluency task. Rec.	Cognitive scores: SVF.	HC (CH) / AD: $n = 30/30$; $m = 75/75$.	CB. Demogr. unreported.	Mandarin_Lu avail. (DementiaBank). HC unreported.	Chinese, Taiwanese (Taiwan).
Clark et al. [67]	158 pps.	Monologues: fluency task. Manual tr. (text file).	Cognitive scores: CDR, MMSE, SVF. NI meas.: MRI.	HC (CN)/MCI-non[45]/MCI-con: $n = 51/83/24$.	No-CB, no-GB. no-WCGB. BCB: no-A, no-G, E.	Unreported.	English (US).
D'Arcy et al. [93]	87 pps. 5 tasks each: $m = 435$.	Narrative monologues and fluency task. Rec. and manual tr.	Cognitive scores: MMSE, NART, Memory, SVF.	HC (MMSE $>$ 27) / CI (MMSE $\leqslant$ 27): $n = 50/37$.	No-CB, no-GB. WCGB: unreported. BCB: unreported.	Unreported.	English (Ireland).
Dos Santos et al. [90]	3 ds (HC, MCI): Pitt: 86 tr. S/N: 9.58, 10.97; W/S: 9.18, 10.33. Cs[46]: 40 tr. S/N: 30.80, 29.90; W/S: 12.17,13.03 ABCD: 85 tr. (46, 39); S/N: 5.23, 4.95; W/S: 11, 12.04.	Pitt: narrative monologues. Cs: narrative monologues. ABCD:narrative monologues (cognitive test). All ds: Rec. and manual tr.	Pitt: MMSE. Cs: NA. ABCD: NA.	HC / MCI: Pitt: $n = 43/43$ Cs: $n = 20/20$ ABCD: $n = 20/23$.	Pitt: CB, no-GB. No-WCGB. No-BCB(A,G). Cs: CB, no-GB. No-WCGB. No-BCB(A,G,E). ABCD: no-CB, GB. No-WCGB. No-BCB(A,G,E).	All ds: available as used in study upon request to authors.	Pitt: English (US). Cs: Portuguese (Brazil). ABCD: Portuguese (Brazil).
Duong et al. [69]	99 pps. 2 tasks each: $m = 198$.	Narrative monologues. Rec. and manual tr. (verbatim).	Cognitive scores: PENO[47], WMS, language, visual.	HC (NE) / AD: $n = 53/46$; $m = 106/92$.	no-CBn, no-GBn. no-WCGBn. BCBn: A, no-G, no-E.	Unreported.	French (Canada).
Egas López et al. [62]	Dementia ds: 75 pps. 3 tasks each: $m = 225$. BEA ds: $m = 44$.	Dementia: Narrative monologues. Rec. and ASR tr. (Kaldi[48]).	Cognitive scores: MMSE, ADASCog, CDT (Dementia).	Dementia: HC, MCI, AD, $n = 25$. BEA: unreported.	Dementia: CB, GB. WCGB unknown. BCB: A, G, E.	Dementia: unreported. BEA: unreported, but avail. online[49].	Dementia&BEA: Hungarian (Hungary).
Espinoza-Cuadros et al. [83]	19 pps.	Narrative monologues. Structured dialogues (test). Rec. and tr.	Cognitive scores: MEC (Spanish MMSE), HDS-R.	HC (non-MCI): $n = 11$; MCI: $n = 8$.	No-CB, no-GB. WCGB (HC only). BCB: A, G, E (unclear).	Unreported.	Spanish (Cuba).
Fraser et al. [7]	55 pps. 3 tasks each: $m = 165$.	Narrative monologues. Rec. and tr.	Eye-tracking. Comprehension questions.	HC / MCI: $n = 29/26$; $m = 87/78$.	no-CBn, no-GBn. no-WCGBn. BCBn: no-A, G, E.	Restricted upon request to authors.	Swedish (Sweden).
Fraser et al. [87]	Gothenburg, Got: 67pps Karolinska, Kar: 96 pps Pitt: 116 pps.	Narrative monologues. Rec. and tr. (Got, Pitt). Written (Kar).	Cognitive scores: MMSE.	Got / Kar / Pitt: HC: $n = 36/96/97$; MCI: n =31/NA/19	CB, GB: Pitt only. WCGB: Pitt only. BCB: A,G (Pitt), E(all)	Got & Kar: unreported. Pitt: avail. (DementiaBank)	Got & Kar: Swedish. Pitt: Eng (US)

[42] http://trans.sourceforge.net

[43] CHAT protocol: Codes for the Human Analysis of Transcripts [128].

[44] aMCI: amnesic single-domain; a+mdMCI: amnesic multiple-domain. Class-balance depends on whether they are considered 1 or 2 groups.

[45] MCI-non (non converters) and MCI-con (converters) refer to whether MCI pps converted to AD or not over a 4-year follow-up.

[46] Cs: retellings of Cinderella Story [129].

[47] PENO is a cognitive battery in French (Joanette et al., 1995). Two pictures, "Bank robbery" and "Car accident" were described in this study.

[48] Kaldi speech recognition toolkit [104].

[49] Available under an Academic-Non Commercial use licence: http://www.nytud.hu/adatb/bea/index.html

Table 6: Detailed Data information (ctd.)

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Fraser et al. [9]	Pitt: 264 pps. Several visits: $m = 473$. W/S: 100.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD: $n = 97/176$; $m = 233/240$.	no-CB*n*,CB*m*,no-GB*m*. no-WCGB*m*. BCB*m*: no-A, G, no-E.	Avail. (DementiaBank).	English (US)
Gonzalez-Moreira et al. [89]	20 pps.	Narrative monologues. Rec.	Cognitive scores: MEC (Spanish MMSE).	HC / CI (MD): $n = 10/10$.	CB, no-GB. no-WCGB. BCB: no-A, no-G, no-E.	Unreported.	Spanish (Cuba).
Gosztolya et al. [63]	Dementia ds: 75 pps. 3 tasks each: $m = 225$.	Narrative monologues. Rec. and phonetic ASR tr.	Cognitive scores: MMSE, ADASCog, CDT.	HC / MCI / AD: $n = 25/25/25$; $m = 75/75/75$	CB*n*, GB*n*. WCGB unreported. BCB*n*: A, G, E.	Unreported.	Hungarian (Hungary).
Guinn et al. [98]	CCC: 56 pp. Several visits: $m = 281$.	Conversational dialogues. Rec. and tr. (Ten Have[50]). Pps with 1+ tr. merged.	Video (not all pps).	HC (non-AD) / AD: $n = 28/28$; $m = 204/77$;	CB*n*, no-CB*m*. Demogr. unreported.	Unreported, but avail. on request (ibid.).	English (US).
Guo et al. [92]	Pitt: 268 pps. Several visits: $m = 498$.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD: $n = 99/169$; $m = 242/256$.	no-CB*n*, CB*m*, no-GB*n*. no-WCGB*n*. BCB*n*: no-A,no-G,no-E.	Avail. (DementiaBank).	English (US)
Haider et al. [11]	Pitt: 164 pps. Speech segments: $m = 4076$.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD: $n = 82/82$; $m = 2033/2043$.	CB*n*,*m*, no-GB*n*. WCGB*n*. BCB*n*: A, G.	Avail. (DementiaBank).	English (US)
Kato et al. [64]	48 pps.	Narrative monologues. Rec.	Cognitive scores: CDR, HDS-R. NI meas.: fNIRS[51]	HC (NC)/MCI/AD: $n = 20/19/9$.	no-CB, no-GB. WCGB: AD only. BCB: no-A, no-G.	Unreported.	Japanese.
Khodabakhsh and Demiroğlu [84]	54 pps. 10 min conversation each.	Semi-structured dialogues. Rec.	NA	HC / AD (Patient): $n = 27/27$.	CB, no-GB. no-WCGB. BCB: no-G. Demogr. unreported.	Unreported.	Turkish.
Konig et al. [102]	64 pps. 4 tasks each.	Monologues: countdown, repetition, picture description, fluency task.	Cognitive scores: MMSE, VF, IADL.	HC[52] / MCI / AD: $n = 15/23/26$	no-CB, GB. WCGB: MCI, AD. BCB: no-A, no-G, no-E.	Unreported.	French (France).
Lopez-de Ipiña et al. [80]	AZTITXIKI: 10 pps. (subset of AZTIAHORE[53]).	Narrative monologues. Conversational dialogues. Rec.	Video.	HC (CR) / AD_{ES} / AD_{IS} / AD_{AS}: $n = 5/1/1/2$	no-CB, no-GB. no-WCGB. BCB: no-A, G.	Unreported.	Multilingual (ibid.).
Lopez-de Ipiña et al. [79]	AZTIAHORE (ibid.): 40 pps.	Narrative monologues. Conversational dialogues. Rec.	Video.	HC (CR) / AD_{ES} / AD_{IS} / AD_{AS}: $n = 10/4/10/6$	no-CB, no-GB. WCGB: HC only. BCB: A, no-G.	Unreported.	Multilingual (ibid.).
Lundholm Fors et al. [59]	Gothemburg: 90 pps.	Narrative monologues. Rec. and tr.	Cognitive scores: MMSE.	HC / SCI / MCI: $n = 36/23/31$.	no-CB, no-GB. WCGB: MCI only. BCB: A, no-G, no-E.	Unreported.	Swedish (Sweden).
Luz [6]	Pitt: Unreported No. pps. Several visits: $m = 398$.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD (ATD): $m = 184/214$.	no-CB.*m*. Unreported CB*n*. Demogr. unreported.	Unreported, but avail, (DementiaBank).	English (US).
Luz et al. [10]	CCC: 38 pps. 17 non-AD and 21 AD.	Conversational dialogues. Rec. and tr. (Ten Have - ibid.)	Video (not all pps).	HC (non-AD) / AD: $n = 17/21$	no-CB, no-GB. no-WCGB. BCB: no-G. A, E unreported.	CCC avail. (ibid.) Study identifiers avail. on request to authors.	English (US).

[50] CCC was transcribed using the Ten Have method [130] and is available upon request through carolinaconversations.musc.edu/

[51] fNIRS: Functional near-infrared spectroscopy. It measures hemodynamic responses in the brain as a proxy to measure neuron behaviour.

[52] The HC group in this study is conformed by pps who did actually have memory concerns but did not meet any diagnostic criteria (i.e. SCI).

[53] In turn, a subset of AZTIAHO: 50HC, 9hours (80% after pre-processing) and 20AD, 60min (50%). AD group is conformed by three AD stages, namely, ES (early), IS (intermediate) and AS (advanced). Multilingual: English, French, Spanish, Catalan, Basque, Chinese, Arabian and Portuguese.

Table 6: Detailed Data information (ctd.)

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Martinez de Lizarduy et al. [60]	AN: 100 pps. PD[54]: 18 pps. SS: 40 pps (AZTIAHORE subset).	AN: fluency task. PD: narrative monologues. SS: spontaneous monologues. Rec.	Video.	AN / PD / SS: HC: $n = 62/12/20$; MCI:n =38/NA/NA AD: =NA/6/20;	CB: SS only. No-GB. no-WCGB$_{AN}$. CBC$_{AN}$: A, no-G. PD & SS: unreported.	Unreported.	AN: unreported PD: unreported SS: multilingual (ibid.).
Meilan et al. [100]	66 pps.	Narrative monologues. Rec.	Cognitive scoreS: MMSE.	HC (control) / AD: $n = 36/30$	no-CB, no-GB. BCB: A, no-G, E. No-WCGB.	Unreported.	Spanish (Spain).
Mirheidari et al. [86]	30 pps: 15 ND, 15 FMD.	Semi-structured dialogues. Rec and manual (verbatim) and ASR tr.	Cognitive scores: MMSE.	FMD / ND: $n = 15/15$	CB, no-GB. WCGB: ND only. BCB: A, G (unclear).	Unreported.	English (UK).
Mirheidari et al. [85]	Pps/files/utt/h/MLU(s): Pitt: 255/473/473/8/61.1 Hallam: 117/45/8970/12/4.8 IVA: 40/18/785/3.25/14.9 Seizure: 597/241/28000/50/6.3	Pitt: narrative monologues. Rec. and tr. Hallam, IVA, Seizure: semi-structured dialogues. Rec and manual (verbatim) and ASR tr.	Cognitive scores: MMSE.	Pitt: HC, AC. Hallam: FMD, ND, DPD. IVA: FMD, MCI, ND. Seizure: different seizure diagoses.	n unreported. Demogr. unreported.	Pitt: avail (DementiaBank). Hallam: unreported. IVA: unreported. Seizure: unreported.	Pitt: English (US). Hallam: English (UK). IVA: English (UK). Seizure: English (UK).
Mirheidari et al. [8]	HUM: 30 pps. IVA: 12 pps.	HUM: structured dialogues. IVA: structured dialogues (with avatar). Rec and CA annotations.	Video (IVA only). Cognitive scores: MMSE, ACE-R.	FMD / ND: HUM: $n = 15/15$. IVA: $n = 6/6$.	HUM & IVA: CB, no-GB. WCGB: ND only. BCB: no-A, no-G.	Unreported.	English (UK).
Mirheidari et al. [12]	61 pps. 4.3h, 1944 utt, 85 spk (incl. chaperons), 8s MLU.	Monologues: fluency task. Structured dialogues (IVA). Rec. and ASR tr. (Kaldi).	Video. Cognitive scores: MMSE, ACE-R.	HC / FMD / MCI / ND[55]: $n = 14/10/18/19$.	No-CB, no-GB. no-WCGB. BCB: no-A, no-G.	Unreported.	English (UK).
Mirzaei et al. [49]	48 pps. Avg samples length: 17.47 s.	Narrative monologues. Rec.	Cognitive scores: MMSE.	HC / MCI / AD: $n = 16/16/16$	CB. G & E unreported. BCB: A (MCI-AD only)	Unreported.	French (France).
Nasrolahzadeh et al. [65]	60 pps. 16h after pre-processing[56]. Segments (60s):$m = 960$	Spontaneous monologues. Rec.	Cognitive scores: MMSE, CDR.	HC/AD$_{FS}$/SS/TS: $n = 30/6/15/6$ $m = 720/70/110/60$	no-CB, no-GB. WCGB: HC only. BCB: no-A, no-G.	Unreported.	Persian (Iran).
Orimaye et al. [75]	Pitt: 198 pps. MLU: 4.03s HC, 2.65s AD.	Narrative monologues. Rec. and manual tr (CHAT)	Cognitive scores: MMSE.	HC / AD: $n = 99/99$	CB. CBC: no-A.	Study data avail on GitHub[57].	English (US).
Prud'Hommeaux and Roark [73]	124 pps. 2 tasks each.	Narrative monologues. Rec. and manual tr.	Cognitive scores: CDR, WMS-III.	HC / MCI: $n = 52/72$	no-CB. Demogr. unreported.	Unreported.	English (US).
Prud'hommeaux and Roark [70]	235 pps. 2 tasks each.	Narrative monologues. Rec. and manual tr.	Cognitive scores: CDR, WMS-III.	HC / MCI: $n = 163/72$	no-CB. BCB: A, E. Gender unreported.	Unreported.	English (US).
Rentoumi et al. [88]	60 pps.	Narrative monologues. Written.	Cognitive scores: MMSE.	HC (NC) / AD: $n = 30/30$	CB. GB. no-WCGB. BCB: A, G (unclear), E.	Unreported.	Greek (Greece).
Roark et al. [78]	74 pps. 2 tasks each.	Narrative monologues. Rec. and manual tr.	Cognitive scores: CDR, MMSE, WMS	HC / MCI: $n = 37/37$	CB. Demogr. unreported.	Unreported.	English (US).
Rochford et al. [76]	187 pps	Narrative monologues. Rec and manual tr.	Cognitive scores: MMSE.	HC / CI: $n = 150/37$	no-CB, no-GB. Class demogr. unreported.	Unreported.	English (Ireland).
Sadeghian et al. [43]	72 pps. Avg sample length: 75.1s (sd 61.0).	Narrative monologues. Rec and tr. (manual+ASR).	Cognitive scores: MMSE.	HC (NL) / AD: $n = 46/26$	no-CB. BCB: no-A, E.	Unreported.	English (US).

[54] AN and PD are the "animal naming" and "picture description" subsets from the Gipuzkoa-Alzheimer Project (PGA): http://www.cita-alzheimer.org/projects/gipuzkoa-alzheimer-project-basque-cohort

[55] HC: healthy control. FMD: functional memory disorder. MCI: mild cognitive impairment. ND: neurodegenerative disorder (i.e. AD).

[56] 32h recorded, 15 from HC and 17 from AD stages. After pre-processing 12h remain from HC, 4h from AD stages.

[57] https://github.com/soori1/ADresearch.

Table 6: Detailed Data information (ctd.)

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Satt et al. [61]	89 pps.	Narrative monologues. Sentence/syllable repetition. Rec.	NA	HC / MCI / AD: $n = 19/43/27$	no-CB, no-GB. No-WCGB. BCB: A (MCI-AD only), no-G.	Unreported.	Greek (Greece).
Shinkawa et al. [71]	34 pps.	Monologue narratives. Wizard[58] of Oz method. Rec.	Gait ast (positional 3D). MMSE scores.	HC / MCI: $n = 19/15$	no-CB, no-GB. WCGB: MCI only. BCB: A, no-G.	Unreported.	Japanese (Japan).
Tanaka et al. [94]	29 pps. Avg interaction: $m = 10 - 15min$.	Structured dialogues (avatar). Rec. and manual tr.	Eye-tracking. Video.	HC / AD: $n = 15/14$ $m = 7 - 3/8 - 22$	CB, no-GB. no-WCGB. BCB: A, G, no-E.	Unreported.	Japanese (Japan).
Thomas et al. [66]	ACADIE: 95 pps. $m = 158$	Conversational dialogues. Rec. and manual tr.	Cognitive scores: MMSE.	HC/ Mild/ Moderate/ Severe: $m = 35/50/53/20$	Unreported.	Unreported.	English (Canada).
Tóth et al. [99]	Dementia: 84 pps. 3 tasks each: $m = 252$. Dementia: unreported.	Narrative monologues. Rec. and phonetic ASR tr.	Cognitive scores: MMSE, ADASCog, CDT.	HC (NC) / MCI: $n = 36/48$	no-CB, no-GB. no-WCGB. BCB: no-A, G, E.	Dementia: unreported. BEA: unreported, but avail. online[59].	Hungarian (Hungary).
Tröger et al. [77]	115 pps. Avg sample length: 140s.	Narrative monologues and countdown task. Rec and ASR tr.	NA	HC / AD: $n = 47/68$	no-CB, no-GB. no-WCGB. CBC: no-A, no-G.	Unreported.	French (France).
Tröger et al. [96]	166 pps.	Monologues: fluency task.	Cognitive scores: MMSE, CDR.	SCI (SMC) / MCI / AD: $n = 40/47/79$	No-CB, no-GB. WCGB: MCI and AD. BCB: no-A, no-G, no-E.	Unreported.	French (France).
Weiner et al. [74]	ISLE: 74 pps. $m = 98$ (treated as n). 230h.	Semi-structured dialogues. Rec. and manual tr.	NA.	HC/ AACD[60] / AD: $m = 80/13/5$	no-CB. Demogr. unreported.	Unreported.	German (Germany).
Weiner and Schultz [68]	ISLE: 23 pps. 112h $m = 51$ (treated as n).	Semi-structured dialogues. Rec. and manual tr.	NA	No-change/Change: $m = 35/16$[61]	no-CB. Demogr. unreported.	Unreported.	German (Germany).
Yu et al. [97]	167 pps. $m = 180$ (treated as n).	Narrative monologues and fluency task. Rec.	Cognitive scores: WMS-III,SVF,Trail	HC / CI[62]. $m = 160/20$	No-CB, unclear G. BCB: A, G, B	Unreported.	English (US).

[58] Wizard of Oz: experiment method by which human-computer interaction is examined. In this case the experimenter pretended to be the computer.
[59] Available under an Academic-Non Commercial use licence: http://www.nytud.hu/adatb/bea/index.html
[60] AACD: ageing-associated cognitive decline.
[61] HC who changed to AACD (ageing-associated cognitive decline), MCD (mild cognitive disorder), AD or VAD (vascular dementia)
[62] CI: cognitive impairment. Heterogeneous group including dementia, amnestic MCI single domain, amnestic MCI multiple domain). Recordings collected quarterly or annually (50-50%).

2.2. Methodology table

This table summarises the features and methods employed in the reviewed studies. It is structured as follows:

- **Pre-processing**: where available, this column describes the procedures undertaken on text and audio data as preparation steps for subsequent analysis. before. For text, this includes transcription (manual or ASR), tokenisation, removal of unanalysable events and *stopwords*, and so on. For audio, this includes background noise removal, normalisation, speaker diarisation.
- **Feature generation**: whether the features were generated from raw data through text analysis and/or through acoustic analysis, followed by more specific subcategories as per the taxonomy described in Table 1. When reported, this column also includes the paper's approach to reduce the extracted feature set, essentially either selection or extraction. On the one hand, *'filtering' selection* uses extrinsic criteria, such as information gain or, commonly, *p*-values (i.e. whether the differences between the experimental groups, e.g. AD and HC, for a particular feature are statistically significant or not); whereas *'wrapping' selection* uses a cross-validation model that searches through the power set of features. On the other hand, *feature extraction* entails creating a new reduced feature set by combining or transforming the original one with method such as PCA, LSA, clustering or ADR.
- **ML task/method**: supervised vs. unsupervised learning. Task: clustering, classification, regression. Method: clustering algorithm, classifiers and regression method as per Table 4. This column also includes information on the number of classes that the classifier outputs.
- **Evaluation technique**: describes four points, when available. First, the baseline against which the study results are compared (i.e. random guess, neuropsychological scores, different feature sets). Second, the performance metrics reported by the authors (i.e. *acc, F1, pc, rc, ss, sp, AUC, EER*, see Table 4). This will include information about different ASR precision measures, such as WER, where applicable. Third, the cross-validation technique used. Fourth, whether a test set held out, unused for model training, and its size.
- **Results**: numerical results of the selected performance metrics for the baseline and for the fitted model/s. When multiple metrics are reported, only summary metrics such as *EER, acc, F1* and *AUC* are included in this column.

Table 7: Methodology

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Beltrami et al. [91]	Processing unit: utt **Text:** manual tr. Paralinguistic annotation. **Audio:** VAD (ssvad[63]) and Kaldi[64]-ASR forced alignment	Filtering (selection): *p-values*. **Text-based:** lex diversity, PoS, lex density, syntactical (dependency). **Acoustic:** prosodic (temporal, F_0, energy), spectral.	Supervised learning. **Classification**: binary (HC-MCI) with *k*-NN ($k = 3$), LR and NN.	B/L: unreported. Metrics: *acc, pc, rc, F*1. CV: unreported. Hold-out set: 80/20%.	LR and NN performed best on "Picture" task: $acc = 76.9\%$, $pc = 0.727$, $rc = 0.842$ and $F1 = 0.781$.
Ben Ammar and Ben Ayed [95]	**Text:** ASR tr. **Audio:** removal of background noise and non-analysable[65] events.	Filtering: IG; Wrapping: *k*-NN, SVM. **Text-based:** lex diversity, lex density, syntactical (constituency), pragmatics (*UoL*).	Supervised learning. **Classification**: binary (HC-AD) with NN, SVM and DT.	B/L: no ft set reduction. Metrics: *acc*. CV: unreported. Hold-out set: unreported.	Best performance: $acc = 79\%$ SVM. Best ft set: *k*-NN ($acc = 69\%$ NN, 71% DT).
Bertola et al. [57]	**Text:** SVF word sequence $\rightarrow$ speech graph.	Filtering (selection): corr w/ cognitive ast. **Text-based:** syntactical (*SGA*).	Supervised learning. **Classification**: binary and 3-way with NB.	B/L: unreported. Metrics: *ss, sp, AUC*. CV: unreported. Hold-out set: unreported.	HC-MCI-AD, HC-MCI, MCI-AD: $AUC = 0.6 - 0.8$ HC-AD: $AUC > 0.8$ MCI subgroups: $AUC < 0.6$
Chien et al. [82]	Processing unit: syl **Text:** ASR tr, tokenization, pause annotation.	Filtering (selection): suitability, trainability, generalizability. DR: manual Feature Sequence. **Text-based:** syllable tokens, ASR-related (*FP, rep, dys*).	Supervised learning. **Classification**: binary (HC-AD) with bidirectional LSTM (RNN).	B/L: unreported. Metrics: *AUC*. CV: unreported. Hold-out set: 85/15%. Random shuffling.	$AUC = 0.954$.
Clark et al. [67]	Processing unit: word **Text:** fluency test manually transcribed for automatic scoring.	Wrapping (selection): RF (importance). **Text-based:** lexical (*BoW,n-grams*), syntactical (*SGA*), semantic (matrix decomposition: ICA), pragmatics (*coh*), fluency scores.	Supervised learning. **Classification**: binary (MCI: non-con) with ensemble RF, SVM, NB and MLP. Combined w/ LASSO	B/L: unreported. Metrics: *AUC*. CV: LOO. Hold-out set: unreported. Bootstrap.	$AUC = 0.872$ incl fluency scores. MRI enhances *sp* but not *ss*.
D'Arcy et al. [93]	**Text:** manual tr. **Audio:** removal of begin/end pauses> 250ms and visually inspected disturbances.	Ft set reduction: unreported. **Acoustic:** prosodic (temporal), ASR-related (pauses patterns)	Supervised learning. **Classification**: binary (MMSE: low-high) with LDA.	B/L: unreported. Metrics: *acc*. CV: unreported. Hold-out set: unclear.	$acc = 76\%$ LDA. Avg vowel duration +17% in low MMSE group.
Dos Santos et al. [90]	**Text:** manual tr., utt segmentation, tokenization, removal of *stopwords*, punctuation, dysfluencies.	Wrapping (selection): majority vote in BoW, CN and CNE[66]. **Text-based:** lexical (*BoW*), syntactical adjacency network (*SGA*) enriched w/ semantic word embeddings.	Supervised learning. **Classification**: binary (HC-MCI) w/ GNB, *k*-NN, RF, SVM (linear and RBF). Multi-view and ensemble.	B/L: unreported. Metrics: *acc*. CV: 5-fold. Hold-out set: unclear.	Pitt: $acc = 65\%$ ensemble. Cinderella: $acc = 65\%$ SVM-RBF, CNE fts. ABCD: $acc = 75\%$ SVM-linear, BoW fts.
Duong et al. [69]	**Text:** manual tr (verbatim), discourse processing (multilayered cognitive model).	Ft set reduction: unreported. **Text-based:** lex diversity, lex density, syntactical (dependency, complexity), pragmatics (*UoL*).	Unsupervised learning. **Clustering**: Euclidean distance on discourse fts. **Factor analysis:** PCA.	B/L: unreported. Metrics: cluster *acc*. CV: N/A. Hold-out set: N/A.	Cluster composition: AD cluster: $acc = 61\%$ (sequence pic), $acc = 41\%$ (single pic)
Egas López et al. [62]	**Audio:** 25 ms signals, 10 ms time-shift. UBM[67] trained on BEA ds.	Extraction: i-vector[68] model fitted w/ UCM and *MFCCs*. **Acoustic:** spectral fts (*20 MFCCs*).	Supervised learning. **Classification**: binary (HC-MCI+AD), 3-way (HC-MCI-AD) w/ SVM.	B/L: unreported. Metrics: *acc, F1*. CV: 5-fold. Hold-out set: unreported.	$F1 = 0.792$, immediate recall task (binary). $acc = 56\%$, all utt (3-way).
Espinoza-Cuadros et al. [83]	Unreported.	Filtering (selection): *p-values*. **Acoustic:** prosodic (temporal: *SR, PR, PhR, AR*).	Supervised learning. **Classification**: binary (HC-MCI) w/ RF.	B/L: no ft set reduction. Metrics: *acc*. CV: LOO. Hold-out set: unreported.	$acc = 78.9\%$, RF (20 trees). Same *acc* w/ all fts and significant fts.

[63] VAD proposed by [131]
[64] http://kaldi.sourceforge.net/about.html
[65] Non-analysable events in this context refers to breaks, overlapping speech, coughing, laughter, short hard noises and the like.
[66] These are different feature spaces (BoW: Bag of Words; CN: Complex Networks; CNE: Complex Networks Enriched with word embeddings.
[67] UBM: Universal Background Model, trained to represent speaker-independent distribution of features [132]
[68] Dimensionality reduction method of the GMM supervector (Gaussian Mixture Model). It assumes each utt is produced by a different speaker

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Fraser et al. [7]	**Text:** manual tr. **Audio:** unreported. + Eye-movement + comprehension.	Ft set reduction: unreported. **Text-based:** lex diversity, lex density, *PoS*, syntactical (dependency). **Acoustic:** prosodic (temporal), ASR-rel. (*FP, dys*).	Supervised learning. **Classification**: binary (HC-MCI) w/ LR and RBF-SVM (Platt's[69]). Cascade: mode,task,session	B/L: train w/ cognitive scores. Metrics: *AUC, acc, ss, sp*. CV: LPO. Hold-out set: unreported.	B/L: $AUC = 0.75$, $acc = 65\%$. Best: $AUC = 0.88$, $acc = 83\%$, task level (both LR and SVM).
Fraser et al. [87]	**Text:** manual tr., removal of dysfluencies, laughter, PoS, lemmatization, extrat Ns and Vs.	Ft set reduction: unreported. **Text-based:** lex density, *n*-gram embeddings (*FastText*), topic modelling (cosine distance, topic frequency, words per topic).	Supervised learning. **Classification**: binary (HC-MCI; HC-AD) w/ linear SVM.	B/L: train w/o topic model fts. Metrics: *acc, ss, sp*. CV: LOO. Hold-out set: unreported.	Multilingual topic model: $acc = 63\%$ English (MCI); $acc = 72\%$ Swedish (MCI). $acc = 82\%$ English (AD).
Fraser et al. [9]	**Text:** word-level tr. and utt segmentation. Remove false starts and *FPs* (other *dys* remain). **Audio:** MP3 to mono WAV.	Filtering (selection): Pearson's corr. **Text-based:** *BoW*, lex diversity/density, *PoS*, syntactical (constituency), semantic (*PsyLing*), pragmatics (*UoL*). **Acoustic:** spectral (*MFCCs*).	Supervised learning. **Classification**: binary (HC-AD) w/ multilinear LR. + Factor analysis.	B/L: unreported. Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	$acc = 81.92\%$ w/ 35 top fts (drops w/ 50+). Four factors: semantic, acoustic, syntactic, information content.
Gonzalez-Moreira et al. [89]	**Audio:** bandpass filter, subband selection, temporal weight, subband corr, Gaussian filter, energy thresdold, F_0 detection.	Filtering (selection): *p-values*. **Acoustic:** automatic syllable nuclei detection to extract prosodic fts (temporal, F_0 and functionals in semitones).	Supervised learning. **Classification**: binary (HC-CI) w/ SVM.	B/L: unreported. Metrics: *acc, ss, sp*. CV: LOO. Hold-out set: unreported.	$acc = 85\%$, $ss = 81.8\%$ and $sp = 88.8\%$, w/ prosodic temporal fts and F_0.
Gosztolya et al. [63]	**Text:** phone-based ASR[70] tr., phonectic segmentation, time-aligned phoneme sequences.	Ft set reduction: unreported. **Text-based:** *PoS*, lex density, syntactical, semantic (topic words). **Acoustic:** phone based prosodic (temporal) and ASR-related (*FP, rep, hes*).	Supervised learning. **Classification**: binary (HC-MCI+AD) and 3-way (HC-MCI-AD) w/ SVM (SMO).	B/L: w/ demogr scores. Metrics: *acc, pc, rc, sp, F1, UAR*. CV: 5-fold. Hold-out set: unreported.	Binary: $UAR = 0.83$, $acc = 82.7\%$, $F1 = 86.3$ (B/L $acc = 68\%$). 3-way (only *acc*): $acc = 69.3$ (B/L 40%).
Guinn et al. [98]	**Text:** manual tr., subjects w/ multiple tr. conglomerated into one.	Filtering (selection): *p-values*. **Text-based:** *PoS*, lex diversity (*TTR, BI, HS*), syntactical (constituency) pragmatics (*UoL*).	Supervised learning. **Classification**: binary (HC-AD) w/ DT and NB.	B/L: unreported. Metrics: *pc, rc* (HC/AD). CV: LOO. Hold-out set: unreported.	$NB_{pc} = 79.3/80.8\%$, $NB_{rc} = 82.1/0.75\%$; $DT_{pc} = 67.9/67.9\%$, $NB_{rc} = 66.7/66.7\%$.
Guo et al. [92]	**Text:** manual tr., removal of annotation codes. Merge "Possible" and "Probable" AD into one AD group.	Filtering (selection): *AUC* (β). **Text-based:** *PoS*, lex diversity (perplexity), lex density, syntactical (constituency), pragmatics (*UoL*). **Acoustic:** prosodic (temporal, F_0), spectral (*MFCCs*), ASR-related (*FP*).	Supervised learning. **Classification**: binary (HC-AD) w/ LR, SVM, DT, RF, *k*-NN.	B/L: all 49 initial fts. Metrics: *acc*. CV: nested LOO. Hold-out set: unreported.	B/L: $acc = 74.8 - 80.7\%$ $acc = 76.8\%$ w/ unigram perplexity; $acc = 85.4\%$ w/ unigram perplexity + initial fts.
Haider et al. [11]	Create one AD group, matched for age and gender. **Audio:** VAD segmentation (energy thresghold= 65), 10s per segment, volume normalisation.	Filtering (selection): standard ft sets[71]. **Acoustic:** prosodic, spectral, vocal quality. Comprehensive ft sets: *emobase*, *ComParE*, *eGeMAPS*, *MRCG* functionals.	Supervised learning. **Classification**: binary (HC-AD) w/ DT, *k*-NN, LDA, RF and SVM.	B/L: random guess. Metrics: *acc, UAR*, confusion matrices. CV: LOO. Hold-out set: unreported.	B/L: $acc = 50.12\%$ $acc = 78.7\%$ w/ DT, hard fusion of ft sets and ADR[72].
Kato et al. [64]	**Audio:** phrase level segmentation, 23ms frames, Hamming window (1024 points). Voice extracted w/ short-time Fourier transform (every 11ms).	Extraction: PCA (+ stepwise regr). **Acoustic:** prosodic (F_0 and trajectories, energy), spectral (formant trajectories, *MFCCs*). + fNIRS[73] measures.	Supervised learning. **Classification**: binary, two-phased (first: HC-CI, second: MCI-AD) w/ NB. Empirical fts cut-off: 26/28.	B/L: unreported. Metrics: *acc, predictive value*. CV: LOO. Hold-out set: unreported.	$acc = 85.4$ w/ 26 cut-off, $acc = 83.3$ w/ 28 cut-off (this improves MCI classification from $acc = 94.7\%$ to $acc = 68.4\%$).

[69] Because SVM does not output probabilities directly.
[70] trained on BEA Hungarian Spoken Language Database Gósy [52].
[71] Standard feature sets available for openSMILE: https://www.audeering.com/opensmile/
[72] ADR: active data representation, novel method presented in this paper.
[73] fNIRS (functional near-infrared spectroscopy) measures cortical activity.

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Khodabakhsh and Demiroğlu [84]	**Audio:** VAD based on the distribution of the short-time frame energy (speech-silence). Automatic Turkish phoneme recogniser.	Ft set reduction: unreported. **Acoustic:** prosodic (temporal, F_0, energy), spectral (formants).	Supervised learning. **Classification**: binary (HC-AD) w/ LDA, SVM and DT.	B/L: unreported. Metrics: *acc, TP, FA*, confusion matrices. CV: LOO. Hold-out set: unreported.	Best performance w/ SVM: $acc = 83\%$, $TP = 88.9\%$, $FA = 23.1\%$
Konig et al. [102]	**Audio:** VAD segmentation based on energy envelop and pitch contour (periodicity). Praat software.	Filtering (selection): *p*-values. **Acoustic:** prosodic (temporal, energy).	Supervised learning. **Classification**: binary (pairwise: HC, MCI, AD) w/ SVM.	B/L: unreported. Metrics: *EER*[74] *or where missclassification rates are equal.*. CV: random subsampling. Hold-out set: unreported.	$EER_{HC-MCI} = 21\%$ (*equal sp-ss* = 0.79). $EER_{HC-AD} = 13\%$ (0.87). $EER_{MCI-AD} = 20\%$ (0.80)
Lopez-de Ipiña et al. [80]	**Audio:** removal of background noise and non-analysable events, VAD segmentation.	Filtering: ft type; Wrapping: CV. **Acoustic:** prosodic (temporal, F_0, energy, *emo*), spectral (formants), vocal quality (*jitt, shimm, HNR*).	Supervised learning. **Classification**: binary (HC-AD) w/ polynomial SVM, MLP, *k*-NN, DT, NB.	B/L: no *emo* fts. Metrics: *acc, CER* (graph). CV: 10-fold. Hold-out set: unreported.	B/L: CER$= 17 - 25\%$ Performance: CER$= 2 - 20\%$ Best: $acc = 93.79\%$ w/ SVM and all *emo* fts.
Lopez-de Ipiña et al. [79]	**Audio:** removal of background noise and non-analysable events, VAD segmentation.	Selection: ft type and CV. **Acoustic:** ibid previous study. + ASR-related: Higuchi Fractal dimension (*FD*).	Supervised learning. **Classification**: binary (HC-AD) w/ MLP and *k*-NN.	B/L: no *FD* fts. Metrics: *acc, CER* (graph). CV: 10-fold. Hold-out set: unreported.	B/L: $CER \approx 14\%$. Best: $CER = 3.11\%$ ($acc = 96.89\%$) w/ MLP and comprehensive ft set.
Lundholm Fors et al. [59]	**Text:** manual tr. and dysfluency annotation.	Ft set reduction: unreported. **Text-based:** syntactical (constituency and dependency).	Supervised learning. **Classification**: binary (pairwise: HC, SCI, MCI) w/ RF.	B/L: unreported. Metrics: *F1*. CV: LOO. Hold-out set: unreported.	$F1_{HC-SCI} = 0.54$, $F1_{HC-MCI} = 0.68$, $F1_{SCI-MCI} = 0.66$.
Luz [6]	**Audio:** VAD segmentation based on amplitude (empirical threshold at -25dB). Syllable nuclei detection	Ft set reduction: N/A. **Acoustic:** prosodic (temporal: vocalisation events and speech rate).	Supervised learning. **Classification**: binary (HC-AD) w/ NB.	B/L: comparable paper. Metrics: *acc, F1, AUC*. CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 58.5\%$ Performance: $acc = 68\%$ ($AUC = 0.734\%$, $F1_{HC} = 0.70\%$, $F1_{AD} = 0.64\%$).
Luz et al. [10]	**Audio:** vocalisation graph generation[75] (VG). Syllable nuclei detection, speech rate normalisation.	Filtering (selection): with and w/o speech rate. **Acoustic:** prosodic (temporal: vocalisation events and speech rate), dialogue turn-taking patterns.	Supervised learning. **Classification**: binary (HC-AD) w/ additive LR. VGO (vocalisation), VGS (vocalisation + speech).	B/L: random guess. Metrics: *acc, pc, rc F1, AUC*. CV: LOO, 10-fold. Hold-out set: unreported.	B/L: $acc \approx 50\%$ VGO: $acc = 81.1\%$, $AUC = 0.798$. VGS: $acc = 86.6\%$, $AUC = 0.894$.
Martinez de Lizarduy et al. [60]	Matched: age and emotion. **Audio:** VAD segmentation in speech signal and dysfluencies (60s instances).	Filtering: *p*-values; Wrapping: CV. **Acoustic:** prosodic (temporal, energy, *loud*), spectral (formants, *MFCCs*), vocal quality (*jitt, shimm, HNR, NHR*). + ASR-related: Higuchi *FD*, entropy.	Supervised learning. **Classification**: binary (SVF: HC-MCI, PD: HC-AD, SS: HC-AD) w/ *k*-NN, SVM, MLP, CNN.	B/L: unreported. Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	SVF: $acc = 80\%$, PD: $acc = 94\%$, SS: $acc = 95\%$, w/ CNN.
Meilan et al. [100]	**Audio:** unreported.	Ft set reduction: unreported. **Acoustic:** prosodic (temporal, F_0, *loud*, energy), vocal quality (*jitt, shimm, HNR, NHR*).	Supervised learning. **Classification**: binary (HC-AD) w/ stepwise LDA.	B/L: unreported. Metrics: *acc*. CV: resubstitution. Hold-out set: unreported.	no-CV: $acc = 84.8\%$ (misclassified: 4 HC, 6 AD). CV: $acc = 83.3\%$ (misclassified: 4 HC, 7 AD).
Mirheidari et al. [86]	**Text:** ASR tr., diarization, conversion to XML, turn start time equated to previous turn end time.	Wrapping (selection): RFE. **Text-based:** *BoW*, lex diversity, semantics (*FW*, topic modelling). **Acoustic**: ASR (dialogue: *TT, dys*).	Supervised learning. **Classification:** binary (FMD-ND) w/ linear SVM, RF, AdaBoost, MLP, SGD.	B/L: no ft set reduction. Metrics: *acc*. CV: LOO. Hold-out set: unreported.	B/L: $acc = 93\%$ Top-10 fts: $acc = 97\%$ w/ SVM, AdaBoost and SGD.
Mirheidari et al. [85]	**Text:** ASR tr., diarization.	Ft set reduction: unreported. **Text-based:** *BoW*, neural word embeddings (*GloVe*: vector average/variance and cosine distance).	Supervised learning. **Classification:** binary and 3-way (FMD, DPD, MCI) w/ LR and CNN-LSTM	B/L: manual approach. Metrics: *acc, WER* (ASR). CV: 10-fold. Hold-out set: unreported.	Binary / 3-way. B/L:*acc*=50-81.25/66.5-70% LR: *acc*=62−100/65.8−70% CNN_LSTM: *acc*=62.3%

[74]EER: Equal Error Rate, the point at which false alarm rate equals misdetection rate. Also the point were specificity=sensitivity (specificity-sensitivity = 1- EER/100)

[75]Markov diagrams encoding conditional transition probabilities between vocalisation events and steady-state probabilities. Vocalisation events: patient/interviewer talk, joint talk, silence (pause and switching pause).

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Mirheidari et al. [8]	**Text:** manual and ASR tr., diarization. **Audio:** unreported.	Wrapping (selection): RFE. **Text-based:** *BoW*, lex diversity. **Acoustic**: prosodic (temporal, $F-0$), vocal quality (*jitt, shimm, HNR, NHR*), ASR (dialogue: *TT, dys*).	Supervised learning. **Classification:** binary (FMD-ND) w/ linear SVM.	B/L: no ft set reduction. Metrics: *acc WER/DER*. CV: LOO. Hold-out set: unreported.	B/L: $acc = 90.0\%$ (manual tr). Top-10 fts: $acc = 100\%$ (manual tr), $acc = 90\%$ (ASR).
Mirheidari et al. [12]	**Text:** manual and ASR tr., diarization.	Wrapping (selection): RFE. **Text-based:** *BoW*, lex diversity, *PCA*. **Acoustic**: prosodic (temporal, $F-0$), vocal quality (*jitt, shimm, HNR, NHR*), ASR (dialogue: *TT, dys*).	Supervised learning. **Classification:** 4-way and binary (HC, FMD, MCI, ND) w/ LR.	B/L: no ft set reduction. Metrics: *acc, AUC, WER/DER*. CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 48-85\%$. Top-22 fts: $acc = 62-94\%$ (lowest for 4-way). $AUC_{4-way} = 0.815$
Mirzaei et al. [49]	**Audio:** band-pass filter (30-100 Hz), speech segmentation (10ms instances).	Wrapping (selection): two-stage. **Acoustic**: prosodic (temporal, F_0), vocal quality (*jitt, shimm, HNR*), spectral (*MFCCs, FBEs*).	Supervised learning. **Classification:** binary (pairwise: HC, MCI, AD) w/ *k*-NN, linear SVM, DT.	B/L: no ft set reduction. Metrics: *acc*. CV: 8-fold. Hold-out set: unreported.	B/L: $acc = 32-36\%$. Selected fts: $acc = 59-62\%$ DT (60%) selects 3 fts only.
Nasrolahzadeh et al. [65]	**Audio:** removal of background noise and non-analysable events. Segmentation (60s instances).	Filtering (selection): IG. **Acoustic**: ASR (*entr*), spectral. Higher order spectral analysis (*HOS*): bispectrum estimation FFT and AR.	Supervised learning. **Classification:** 4-way (HC-FS-SS-TS) w/ *k*-NN, RBF-SVM, NB, DT.	B/L: comparable paper. Metrics: *acc, ss, sp* (class). CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 81-97.96\%$. FFT fts: $acc = 95.42\%$ DT. AR fts: $acc = 97.71\%$ *k*-NN
Orimaye et al. [75]	Pp selection (last visit). **Text:** manual tr.	Filtering (selection): *p*-values. **Text-based:** *BoW* (*n*-grams), syntatical (constituency, dependency), semantic (*FW*), pragmatics (*UoL: rep, dys*).	Supervised learning. **Classification:** binary (HC, AD) w/ SVM (SMO).	B/L: previous work. Metrics: *AUC*. CV: LPO. Hold-out set: unreported.	B/L: $AUC = 0.75$. Top-1000 fts: $AUC = 0.93$
Prud'Hommeaux and Roark [73]	**Text:** manual word-level tr., tokenisation, downcase. Removal of partial words, punctuation, fillers.	Ft set reduction: unreported.	Supervised learning. **Classification:** binary (HC-MCI) w/ SVM.	B/L: manual scores. Metrics: *AUC, pc, rc, F1*. CV: LPO. Hold-out set: alignment	B/L: $AUC = 0.822$ Training: $AUC = 0.795$ Weighting: $AUC = 0.784$ Intersection: $AUC = 0.767$
Prud'hommeaux and Roark [70]	**Text:** manual utt level tr., downcase. Removal of partial words, punctuation, fillers.	Ft set reduction: unreported. **Text-based:** automatic task scoring alignment based (retelling and phrase level) and graph based.	Supervised learning. **Classification:** binary (HC-MCI) w/ RBF-SVM.	B/L: manual scores, MMSE. Metrics: *AUC, pc, rc, F1*. CV: LPO. Hold-out set: alignment.	B/L: $AUC = 0.733-0.751$ Alignment: $AUC = 0.751$ Graph: $AUC = 0.748$ Pitt: $AUC = 0.832/0.823$
Rentoumi et al. [88]	**Text:** written data. Experiment A: $n = 60$ Experiment B: $n = 200$[76].	Ft set reduction: unreported. **Text-based:** lex diversty (*TTR, BI*), *PoS (word type freq)*, syntactical complexity (constituency).	Supervised learning. **Classification:** binary (HC-AD) w/ SVM (SMO) and NB.	B/L: ZeroR Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 0.50$ $NB_A = 78\%$, $NB_B = 85\%$; $SVM_A = 80\%$, $SVM_B = 88.5\%$.
Roark et al. [78]	**Text:** manual utt tr., manual syntactic annotation (Penn Tree-bank), automatic parsing (Charniak parser), manual and forced time-alignment.	Ft set reduction: unreported. **Text-based:** lex density, *PoS*, syntactical (constituency, dependency). **Acoustic:** prosodic (temporal), spectral (*MFCCs*).	Supervised learning. **Classification:** binary (HC-MCI) w/ SVMlight.	B/L: unreported Metrics: *AUC, corr*. CV: LPO. Hold-out set: unreported.	$corr = 0.87-0.96$ (manual-automatic fts) $AUC = 0.861$
Rochford et al. [76]	**Audio:** removal of background noise (high-pass filter) and breath. Full-wave signal rectification. Step segmentation.	Filtering (selection): *p*-values. **Acoustic:** distribution fts and prosodic temporal fts (conventional static and individual dynamic thresholds).	Supervised learning. **Classification:** binary (HC-CI) w/ LDA.	B/L: unreported Metrics: *acc, ss, sp, AUC*. CV: *k*-fold. Hold-out set: unreported.	Distribution: *acc*=68.66% (*AUC*=0.74) Static= 65.39% (0.69) Dynamic= 61.97% (0.58)
Sadeghian et al. [43]	**Text:** manual and ASR[77] tr. **Audio:** removal of begin/end pause and click. Signal normalisation. VAD for segmentation.	Wrapping (selection): best first greedy. **Text-based:** *LIWC*, *PoS*, lex diversity, lex density, syntactical (constituency). **Acoustic:** prosodic (temporal,F_0,*emo*).	Supervised learning. **Classification:** binary (HC-AD) w/ MLP.	B/L: MMSE scores Metrics: *acc, WER* (ASR). CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 70.8\%$. Manual: $acc = 93.1\%$. ASR: $acc = 91.7\%$ Audio+demogr: *acc*=83.3%

[76] Synthetic samples created with SMOTE [133]

[77] Developed custom ASR with limited domain vocabulary and no requirement for real-time ASR. RNN GRU (Gated Recurrent Units) used for automatic punctuation.

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Satt et al. [61]	**Audio:** manual segmentation (silences above 60ms are pauses).	Filtering (selection): *p*-values. **Acoustic:** prosodic (temporal, energy).	Supervised learning. **Classification:** binary (HC-AD, HC-MCI, HC-*both*) w/ SVM.	B/L: unreported Metrics: *EER*. CV: 4-fold. Hold-out set: unreported.	$EER_{HC-AD} = 15.5\%$. $EER_{HC-MCI} = 17\%$. $EER_{HC-both} = 18\%$.
Shinkawa et al. [71]	**Text:** ASR tr., manual correction and annotation (fillers, false starts). **Audio:** microphone synchronisation.	Wrapping (selection): *ROC-AUC*. **Text-based:** *PoS*, lex diversity, semantic (cosine), syntactical (dependency). **Acoustic:** prosodic (temporal). + Gait fts.	Supervised learning. **Classification:** binary (HC-MCI) w/ linear SVM.	B/L: MMSE scores Metrics: *acc, ss, sp, F1*. CV: LOO. Hold-out set: unreported.	B/L:*acc*=76.5% (*F1*=0.667) Speech: *acc*=76.5% (0.733). Gait: *acc*=76.5% (0.667) Multimodal: *acc*=82.4% (0.813).
Tanaka et al. [94]	Avatar system: MMDAgent[78]. **Text:** manual utt tr and annotation, tokenisation. **Audio:** microphone gain set to 70dB. Separate video from audio.	Filtering (selection): *p*-values. **Text-based:** *PoS*, lex diversity (*TTR*), pragmatics (*UoL: hes*). **Acoustic:** prosodic (temporal, F_0, energy), vocal quality, dialogue (*TT*). + Image fts.	Supervised learning. **Classification:** binary (HC-AD) w/ linear SVM and LR.	B/L: unreported. Metrics: *AUC, acc*. CV: LOO. Hold-out set: unreported.	SVM: $AUC = 0.93$ ($acc = 83\%$); LR: $AUC = 0.91$ ($acc = 79\%$).
Thomas et al. [66]	**Text:** manual tr.	Ft set reduction: unreported. **Text-based:** *PoS*, lex diversity (*TTR, BI, HS*), semantic (clause-like unit, *n*-grams).	Supervised learning. **Classification:** binary (HC-severe/mild) and 4-way (HC, mild, moderate, severe) w/ CNG[79] and CWF.	B/L: ZeroR Metrics: *acc*. CV: unreported. Hold-out set: unreported.	HC-severe: B/L=63.6%, CWF=94.5%. HC-mild: B/L=58.8%, CWF=75.34-way: B/L=33.5%, CWF=50%.
Tóth et al. [99]	**Text:** ortographic and phonetic manual tr and annotation.	Filtering (selection): *p*-values. **Acoustic:** prosodic (temporal), ASR (*FP*). Automatic and manual extraction.	Supervised learning. **Classification:** binary (HC-MCI) w/ NB, RF and linear SVM (SMO).	B/L: manual, no ft set reduction Metrics: *acc,ss,sp,F1,AUC* CV: LOO. Hold-out set: unreported.	B/L: $F1 = 0.75$, $accwa = 71.4\%$ w/ SVM. Top-26, automatic: $F1 = 0.788$, $acc = 75\%$ Top-23 fts: $acc = 89\%$.
Tröger et al. [77]	**Audio:** manual segmentation based on signal intensity, 25-28dB; silence length, 0.25-0.5s; minimum sound length, 0.1s.	Filtering (selection): mutual info. **Acoustic:** prosodic (temporal). Silence/sound segments, syllable information.	Supervised learning. **Classification:** binary (HC-AD) w/ SVM (RBF).	B/L: no ft set reduction Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	
Tröger et al. [96]	**Text:** manual and ASR tr. **Audio**: manual segmentation based on signal intensity.	Filtering (selection): clinical relevance. **Text-based:** *BoW, PoS*, lex diversity, semantic (neural word embeddings: distance). **Acoustic:** prosodic (temporal).	Supervised learning. **Classification:** binary (SCI-CI) w/ SVM.	B/L: no ft set reduction Metrics: *AUC, ss, sp, VFER* (ASR). CV: LOO. Hold-out set: unreported.	$VFER = 33.4\%$. Manual tr: $AUC = 0.852$. ASR tr: $AUC = 0.855$.
Weiner et al. [74]	**Text:** manual tr. Speaker segmentation (audio alignment). **Audio:** VAD segmentation (HMM recognizer).	Ft set reduction: unreported. **Acoustic:** prosodic (temporal).	Supervised learning. **Classification:** 3-way (HC-AACD[80]-AD) w/ LDA (SVD, no shrinkage).	B/L: unreported Metrics: *acc,UAR,pc,rc,F1*. CV: stratified 3-fold. Hold-out set: unreported.	$acc = 85.7\%$. $UAR = 0.66$ $F1_{HC}$=0.92, $F1_{AD}$=0.80, $F1_{AACD}$=0.33.
Weiner and Schultz [68]	**Text:** manual tr. Speaker segmentation (audio alignment). **Audio:** VAD segmentation (HMM recognizer).	Ft set reduction: unreported. **Acoustic:** prosodic (temporal).	Supervised learning. **Classification:** binary (no change-change[81]) w/ LDA (SVD, no shrinkage).	B/L: naively estimated *F1* Metrics: *acc, pc, rc, F1*. CV: stratified 6-fold. Hold-out set: unreported.	$Acc = 80.4\%$. No change / Change: $F1_{B/L}$=0.81, LDA=0.87 $F1_{B/L}$=0.48, LDA=0.64.
Yu et al. [97]	**Audio:** discard poor quality audio files, cross-session averaging.	Filtering (selection): Cohen's *d*. **Acoustic:** prosodic (temporal, F_0), spectral (formants)	Supervised learning. **Classification:** binary (HC-CI) w/ SVM and GC.	B/L: SVF score Metrics: *AUC*. CV: LPO. Hold-out set: yes (no %)	B/L: $AUC = 0.54$ GC, $AUC = 0.58$ SVM. GC: $AUC = 0.73$. SVM: $AUC = 0.75$.

[78] http://www.mmdagent.jp/

[79] CNG: Common *N*-grams approach. CWF: Common Word Frequencies.

[80] AACD: Age-associated cognitive decline.

[81] Intra-personal change measured by subtracting early speech vector from the later speech vector, and normalising resulting vector to unit length.

Clinical applicability

This table summarises our assessment of the potential implications and applications of findings of each reviewed paper as regards research and clinical use. The table is structured as follows:

- **Research implications**:
  - Research Novelty: whether at the time of publication the study described a new dataset, proposed a new set of features, implemented a new method or applied an existing one for a different task;
  - Study Replicability: *low*, *partial* or *full*, depending on how well the procedure is described and whether data or data identifiers are available). *Low* refers to cases where both data is unavailable and method description is incomplete or unsatisfactory; *partial* to cases where either is the case, and *full* when both data and methods are available and satisfactorily described.
  - Results generalisability: *low*, *moderate* and *high*, depending on whether the analysis is specific to the task, and/or there have been any extrinsic validation procedures and/or robust evaluation techniques are in place (i.e. train-test, CV, baseline). *Low* refers to cases where the analysis is indeed specific to the task, and therefore difficult to apply to other tasks (e.g. when relying heavily on content features). In *low* generalisability studies there are no extrinsic validation procedures (e.g. pilot in clinical settings) and the evaluation techniques are insufficient (e.g. CV is in place, but no train-test and/or appropriate baseline comparisons). The improvement of one of these features would bring the study up to *moderate*, and further improvements would make its generalisability *high*. Given the state of the field, no study is 100% generalisable, hence why we have used this terminology instead of the same we used for replicability. For generalisability to be *high*, most conditions need to be met except for the extrinsic validation, since it is still very uncommon in the field that studies are carried within a clinical setting.
- **Clinical potential**: external validation is outlined if present. That is, whether the procedure has actually been attempted in real life (*yes*); or is, at least, embedded in a device, or the experimental design envisions realistic clinical testing at some stage (*in-design*). This column also includes potential applications (i.e. early screening for new cases of SCI or similar, monitoring disease progression or supporting diagnosis of MCI and AD), potential outcomes for global health (i.e. language of study) and potential for the methodology to be remotely applicable (no, suggested potential, yes when tried or purposefully designed with that in mind).
- **Risk of bias**: Feature balance (*no/partial/yes*), suitable metrics (*yes/no*, i.e. whether metrics other than overall accuracy are reported when data are class-imbalanced), contextualized results (*yes/no*, i.e. whether an appropriate baseline is provided in order to put results into perspective), overfitting (*yes/no*, i.e. whether cross-validation and/or hold-out set procedures are implemented). With regards to sample size, we specify three ranges that ranges: $ds \leqslant 50$, $ds \leqslant 100$ and $ds > 100$.
- **Strengths/Limitations**: several characteristics are listed with a yes/no answer, "yes" indicating strength and "no" indicating limitation. These characteristics are:
  - spontaneous speech: speech data is naturally generated, generated in response to an open-answer question or a narrative task, or generated in response to a scripted cognitive task (i.e. verbal fluency or counting). Speech is considered spontaneous when it is natural and when its prompted by open-answered or narrative tasks. That is, for example, the Cookie Theft picture description would be spontaneous (although not natural), whereas reading sentences from a screen saying as many animals as possible within 60 seconds is not spontaneous (nor natural).
  - conversational speech: whether the study includes dialogue data or only monologue.
  - automation: the only characteristic that observes a 'middle' stage. Method automation can be labeled as *no*, when the only automated procedure is the ML task; *partial*, when aspects of the procedure other than the ML task, such as feature set reduction, are also automated; or *total*, when everything is automated including preprocessing (e.g. ASR is used for transcription).
  - content-independence: whether the model for feature generation relies heavily on content features of the data (e.g. lexical or high level *n*-gram are often closely related to the way in which spoken language was prompted).

* Transcription-free: text analysis usually requires transcripts. Whether manual or ASR, transcribing procedures entail many restrictions. Manual transcription is time-consuming, whereas ASR transcription have limited performance on impaired speech, and they need to be trained to a specific language, therefore adding an extra step to the method.

Table 8: Clinical applicability

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Beltrami et al. [91]	**Novelty:** preliminary results of new project (OPLON). **Replicability:** partial. **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Italian sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** no. **Overfitting:** hold-out set, no CV. **Sample size:** $ds \leqslant 50$ ($n = 39$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial (manual tr). **Content-independence:** no. **Transcription-free:** no.
Ben Ammar and Ben Ayed [95]	**Novelty:** speech samples only. Compare three ft selection processes. **Replicability:** partial (unreported n). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** no hold-out set, no CV. **Sample size:** $ds > 100$ ($m = 484$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial (manual tr). **Content-independence:** no. **Transcription-free:** no.
Bertola et al. [57]	**Novelty:** graph analysis, MCI subtypes, 3-way classification. **Replicability:** partial (unclear performance metrics). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** disease progression. **Global Health:** Brazilian Portuguese words. **Remote application:** no.	**Feature balance:** yes[82]. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 100$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Chien et al. [82]	**Novelty:** ft selection based on suitability, trainability and generalizability. **Replicability:** partial (*ad hoc* fts & data). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Chinese syllables $\rightarrow$ generalisable to Taiwanese and Hakka. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*AUC*). **Contextualised results:** no. **Overfitting:** hold-out set, no CV. **Sample size:** $ds \leqslant 100$ ($n = 60$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.
Clark et al. [67]	**Novelty:** new fluency scores. Inclusion of MRI data. 4-year follow-up. Ensemble classifier. **Replicability:** full. **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** disease progression. **Global Health:** US English words. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 158$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
D'Arcy et al. [93]	**Novelty:** ASR and prosodic fts (in 2008). **Replicability:** partial (incomplete data information and procedure). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Irish English sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** no CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 87$)	**Spontaneous speech:** some. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Dos Santos et al. [90]	**Novelty:** complex networks enriched w/ word embeddings. Multi-view and ensemble classifiers. **Replicability:** full. **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English and Brazilian Portuguese sentences. **Remote application:** suggested potential.	**Feature balance:** *Pitt* and *Cs* CB. **Suitable metrics:** yes (CB $\rightarrow$ *acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 40 - 86$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Duong et al. [69]	**Novelty:** discourse analysis, cluster analysis. **Replicability:** partial (incomplete procedure). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** French sentences. **Remote application:** no.	**Feature balance:** age only. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** N/A. Reliability test. **Sample size:** $ds \leqslant 100$ ($n = 99$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Egas López et al. [62]	**Novelty:** i-vector approach, spectral fts only. **Replicability:** full **Generalisability:** high (2 ds, task-independent model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Hungarian sentences. **Remote application:** no.	**Feature balance:** yes[83]. **Suitable metrics:** yes (*acc, F1*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 75$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.

[82] aMCI: amnesic single-domain; a+mdMCI: amnesic multiple-domain. Class-balance depends on whether they are considered 1 or 2 groups.
[83] Class-balance depends on whether MCI and AD are considered 1 group (CI, better results) or 2 groups.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Espinoza-Cuadros et al. [83]	**Novelty:** prosodic fts only. Transcribed MEC. **Replicability:** full. **Generalisability:** moderate (task-independent model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Cuban Spanish sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 19$)	**Spontaneous speech:** yes. **Conversational speech:** no[84]. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** no.
Fraser et al. [7]	**Novelty:** multimodal language data and eye-tracking. Cascaded classifiers. **Replicability:** full. **Generalisability:** moderate (different data types)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** G & E only. **Suitable metrics:** yes (*AUC, acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 55$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Fraser et al. [87]	**Novelty:** topic models, multilingual word embeddings (English, Swedish). **Replicability:** full. **Generalisability:** high (different languages).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** multilingual model $\rightarrow$ higher performance. **Remote application:** no.	**Feature balance:** *Pitt* only. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 67 - 116$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Fraser et al. [9]	**Novelty:** comprehensive model (text-based and acoustic fts). **Replicability:** full. **Generalisability:** moderate (task-specific model).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English (*Pitt*). **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 264$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Gonzalez-Moreira et al. [89]	**Novelty:** Mild dementia. Specific tool and software[85]. **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Cuban Spanish sentences. **Remote application:** no.	**Feature balance:** class only. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 20$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Gosztolya et al. [63]	**Novelty:** custom phone-based ASR, phonetic seg. **Replicability:** partial (incomplete procedure). **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Hungarian phonemes. **Remote application:** no.	**Feature balance:** yes[86]. **Suitable metrics:** yes (*acc, UAR*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 75$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** unclear. **Content-independence:** no. **Transcription-free:** no.
Guinn et al. [98]	**Novelty:** dialogue data, pragmatic fts. **Replicability:** partial (no pp IDs). **Generalisability:** moderate (representative data).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English dialogues. **Remote application:** no.	**Feature balance:** yes **Suitable metrics:** yes (*acc, UAR*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 56$)	**Spontaneous speech:** yes. **Conversational speech:** no[87]. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Guo et al. [92]	**Novelty:** comprehensive model, incl perplexity fts from LM. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** no **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 268$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.

[84] Database contains conversational speech but it is not included in the analysis.
[85] DCGrab v3-0. Allows storing clinical and demographic data for each patient, as well as their voice.
[86] Class-balance depends on whether MCI and AD are considered 1 (CI, better performance) or 2 groups.
[87] Database contains conversational speech but specific dialogue features are not included in the analysis.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Haider et al. [11]	**Novelty:** comprehensive standard ft sets, enhanced data. ADR method. **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** yes **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 164$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Kato et al. [64]	**Novelty:** two-phase system w/ prosodic and physiological fts (cerebral blood flow). **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Japanese sentences. **Remote application:** no.	**Feature balance:** no **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 48$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Khodabakhsh and Demiroğlu [84]	**Novelty:** analyse ft pairs. Dialogue data. **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Turkish dialogues. **Remote application:** no.	**Feature balance:** class only **Suitable metrics:** yes (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 54$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Konig et al. [102]	**Novelty:** dynamic time warping for ft extraction. **Replicability:** full. **Generalisability:** high (investigated w/ unseen data).	**External validation:** no. **Potential application:** disease progression. **Global Health:** French sentences. **Remote application:** suggested potential.	**Feature balance:** gender only **Suitable metrics:** yes (*EER*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 64$)	**Spontaneous speech:** no (SVF). **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Lopez-de Ipiña et al. [80]	**Novelty:** preliminary results of new project (AZTIAHO). Emotional response fts. **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Multilingual model. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc, CER*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 10$)	**Spontaneous speech:** yes. **Conversational speech:** no[88]. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Lopez-de Ipiña et al. [79]	**Novelty:** emotional temperature and fractal dimension fts. **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Multilingual model. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc, CER*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 40$)	**Spontaneous speech:** yes. **Conversational speech:** no[89]. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Lundholm Fors et al. [59]	**Novelty:** incl SCI pps, syntactic complexity only. **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Swedish sentences. **Remote application:** no.	**Feature balance:** age only. **Suitable metrics:** yes (*F1*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 90$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Luz [61]	**Novelty:** vocalisation fts only. **Replicability:** low (unreported *n*). **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*acc, AUC, F1*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($m = 398$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Luz et al. [10]	**Novelty:** turn-taking fts. Dialogue data. **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*acc, AUC, F1*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($m = 38$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.

[88] Database contains conversational speech but specific dialogue features are not included in the analysis.
[89] Database contains conversational speech but specific dialogue features are not included in the analysis.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Martinez de Lizarduy et al. [60]	**Novelty:** preliminary results of acoustic decision support system (ALZUMERIC). **Replicability:** partial. **Generalisability:** high.	**External validation:** in-design. **Potential application:** diagnosis support. **Global Health:** Multilingual model. **Remote application:** suggested potential.	**Feature balance:** not all three ds. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 40 - 100$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Meilan et al. [100]	**Novelty:** acoustic fts only. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Spanish sentences. **Remote application:** no.	**Feature balance:** age and educ only. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 66$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Mirheidari et al. [86]	**Novelty:** doctor-patient consultation. Conversational fts. **Replicability:** full. **Generalisability:** moderate.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK English conversations. **Remote application:** suggested potential.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 30$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Mirheidari et al. [85]	**Novelty:** doctor-patient consultation, human-robot interaction. Word-vector repr, conversational fts. Several ds. **Replicability:** low. **Generalisability:** moderate.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK/US English conversations. **Remote application:** suggested potential.	**Feature balance:** unreported. **Suitable metrics:** unclear (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** varied ($n = 40 - 255$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Mirheidari et al. [8]	**Novelty:** compare doctor-patient consultation w/ human-robot interaction. Conversational analysis fts. **Replicability:** full. **Generalisability:** moderate.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK English conversations. **Remote application:** suggested potential.	**Feature balance:** class only. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 12 - 30$)	**Spontaneous speech:** no. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Mirheidari et al. [12]	**Novelty:** human-robot interaction for cognitive ast. 4-way classification. **Replicability:** full. **Generalisability:** low.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK English conversations. **Remote application:** suggested potential.	**Feature balance:** class only. **Suitable metrics:** yes (*acc, AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 12 - 30$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Mirzaei et al. [49]	**Novelty:** two-stage ft selection. Acoustic fts only. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** French sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 48$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.
Nasrolahzadeh et al. [65]	**Novelty:** HOS analysis of speech data. Best 4-way classifier (AD stages). **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Persian sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 60$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Orimaye et al. [75]	**Novelty:** comprehensive linguistic fts, incl n-grams approach. **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** suggested potential.	**Feature balance:** class only. **Suitable metrics:** no (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, unclear hold-out set. **Sample size:** $ds > 100$ ($n = 198$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Prud'Hommeaux and Roark [73]	**Novelty:** automatic word alignment for scoring recall task. **Replicability:** partial. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 124$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Prud'hommeaux and Roark [70]	**Novelty:** automatic graph-based word alignment for scoring recall task. **Replicability:** partial. **Generalisability:** high (translate to *Pitt*).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 235$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Rentoumi et al. [88]	**Novelty:** written data. **Replicability:** partial. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Greek sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 60$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Roark et al. [78]	**Novelty:** combine speech fts and recall cognitive scores. Late onset MCI. **Replicability:** partial. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 74$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Rochford et al. [76]	**Novelty:** dynamic minimum pause threshold estimation (pause distribution). **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Irish English sentences. **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*acc, AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 187$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.
Sadeghian et al. [43]	**Novelty:** compare combinations of manual, custom ASR and MMSE fts. **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** educ only. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 72$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Satt et al. [61]	**Novelty:** compare combinations of manual, custom ASR and MMSE fts. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Greek sentences and syllables. **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*EER*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 89$).	**Spontaneous speech:** yes (some). **Conversational speech:** no. **Automation:** no. **Content-independence:** yes. **Transcription-free:** yes.
Shinkawa et al. [71]	**Novelty:** multimodal data (gait and speech). **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Japanese sentences. **Remote application:** suggested potential.	**Feature balance:** age only. **Suitable metrics:** yes (*acc, F1*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 34$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Tanaka et al. [94]	**Novelty:** human-robot interaction. Dialogue and image data (multimodal approach). **Replicability:** full. **Generalisability:** low.	**External validation:** in-design. **Potential application:** diagnosis support. **Global Health:** Japanese conversations. **Remote application:** yes.	**Feature balance:** yes. **Suitable metrics:** yes (*acc, AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 29$).	**Spontaneous speech:** no. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Thomas et al. [66]	**Novelty:** custom common *n*-grams algorithm. 4-way classification. **Replicability:** low. **Generalisability:** low.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Canadian English conversations. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** no CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 95$).	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** unclear. **Content-independence:** no. **Transcription-free:** no.
Tóth et al. [99]	**Novelty:** custom phone-based ASR, phonetic seg. Compare automatic and manual approach. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Hungarian phonemes. **Remote application:** no.	**Feature balance:** gender & educ. **Suitable metrics:** yes (*acc, F1, AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 84$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** no.
Tröger et al. [77]	**Novelty:** infrastructure-free system, potentially remote and longitudinal within-subjects. Acoustic fts only. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** in-design. **Potential application:** diagnosis support. **Global Health:** French words/sentences. **Remote application:** yes (simulated).	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 115$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** no.
Tröger et al. [96]	**Novelty:** simulated telephone-based screening (SVF). **Replicability:** full. **Generalisability:** low.	**External validation:** in-design. **Potential application:** disease progression. **Global Health:** French words. **Remote application:** yes (simulated).	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 166$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Weiner et al. [74]	**Novelty:** custom VAD algorithm. Longitudinal dialogue data[90], 3-way classification. **Replicability:** low. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** German conversations. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*acc, UAR*). **Contextualised results:** no. **Overfitting:** stratified CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 74$).	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** no.
Weiner and Schultz [68]	**Novelty:** prediction of within-subjects cognitive change. Custom VAD algorithm. Longitudinal dialogue data. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** German conversations. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** stratified CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 51$).	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** no.
Yu et al. [97]	**Novelty:** telephone-based cognitive ast. 4-year longitudinal collection[91]. Compare speech and cognitive scores. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** US English sentences. **Remote application:** yes.	**Feature balance:** demogr, no class. **Suitable metrics:** yes (*UAC*). **Contextualised results:** yes. **Overfitting:** CV & hold-out set. **Sample size:** $ds > 100$ ($n = 165$).	**Spontaneous speech:** some. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.

[90] Database contains longitudinal samples of conversational speech. However dialogue features are not included in the analysis, and samples by one pp are treated as different pps $\rightarrow$ subject dependence)

[91] Cross-observation averaging: discards longitudinal information, althought does not introduce subject dependence.