Multilingual Natural Language Processing Approaches for Alzheimer's Detection: A Scoping Review

Anonymous ACL submission

Abstract

Spoken language analysis is a compelling tool for detecting signs of Alzheimer's disease (AD). However, most language-based AD detection resources are only available in English, leaving multilingual or crosslingual AD detection understudied. We review the current state of the field with respect to this topic, compiling recent approaches, constraints, and potential solutions from original work published across engineering, natural language processing, and medical databases from 2004 to the present. From the 776 search results, we identified 42 articles meeting predefined eligibility criteria and summarized their findings. Promising results are reported in almost all studies, but few are integrated into clinical practice. The main limitations of the field are poor standardization, the lack of benchmark data repositories (which in turn hinders direct results comparison), and some disconnect between the study goals and the clinical applications. Active efforts from the research community and industry to close these gaps would support robust integration of research across languages into clinical practice.

1 Introduction

Recently, the natural language processing community has collectively devoted increased attention to the detection of Alzheimer's disease (AD) and related cognitive impairment (Ding et al., 2024). This is a welcome development, creating opportunity for transformative support technologies in innovative language use contexts. Popular challenges have been organized to stimulate research on automated dementia detection from text (Luz et al., 2020b) and acoustic (Luz et al., 2021a) features, accelerating progress. Smaller bodies of work have investigated the relationship between linguistic features and biomarkers of AD (Hajjar et al., 2023; Farzana et al., 2024) and tracking AD progression over time (Gkoumas et al., 2023).

Unfortunately, this flurry of progress has been constrained almost entirely to the English language (Luz et al., 2023a). This follows trends across the NLP community; although precise estimates vary, it is widely accepted that the vast majority of NLP research is conducted using English text (Søgaard, 2022). However, it is also problematic: people experience AD across the globe, and less than 20% of the world population speaks English (Bahji et al., 2023). This means that most of the world is missing out on the benefits of convenient, cost-effective, and customized technologies resulting from contemporary spoken-language AD detection research.

Some research has begun to examine non-English or multilingual approaches for these settings (Luz et al., 2023a). Unique challenges arise when developing systems for languages less resourced (Hedderich et al., 2021), and these can be compounded by applying them to a task already under-resourced (Farzana and Parde, 2023). In this review, we synthesize existing research in NLP and adjacent research communities on the automated detection of Alzheimer's disease in crosslingual and multilingual settings to answer four questions:

- What are the characteristics of currently available multilingual AD detection datasets?
- How is multilingual AD detection currently modeled?
- How well do multilingual AD detection approaches currently perform?
- What are the challenges and limitations of these studies to date?

Through our review, framed using PRISMA scoping review guidelines (Tricco et al., 2018), we hope to provide a starting point for other researchers interested in pursuing multilingual AD detection. We make all resources resulting from

our review publicly available to further promote literacy in this area and facilitate follow-up by others.

2 Materials and Methods

2.1 Initial Search

We conducted our search using the following electronic databases: *ISCA-SPEECH*,[1] *ACL Anthology*,[2] *ACM*,[3] *IEEE*,[4] *PubMed*, *Computer Speech & Language*,[5], *Springer*,[6], *Frontiers*,[7] and *ResearchGate*.[8] We selected these databases to ensure comprehensive coverage of our target material. Some of the sources (ISCA-SPEECH, ACL Anthology, and Computer Speech & Language) are prominent databases for publications in the NLP and broader speech and language processing community. Others (ACM and IEEE) are popular databases for computer science research in general. PubMed is a premier database for healthcare publications, and ResearchGate is a social network platform for researchers to share their work.

We were specifically interested in reviewing papers that focus on analyzing spoken language data (e.g., speech recordings and transcripts) from elderly people with AD, mild cognitive impairment (MCI),[9] and age-matched healthy controls (HC). Unlike most previous surveys on the detection of Alzheimer's disease (Voleti et al., 2019; de la Fuente Garcia et al., 2020; Yang et al., 2022), we focused solely on articles that dealt with crosslingual and multilingual approaches to AD detection. We define *crosslingual* approaches as those that learn from data in one language to make predictions in another, and *multilingual* approaches as those that learn from multiple languages to make predictions in all of those languages.

For our initial search, we used the following keywords: `((dementia OR Alzheimer* OR cognit* OR "decline" OR impair*) AND (screen* OR diagnos* OR monitor* OR speech OR audio OR voice OR "multilingual" OR "crosslingual" OR "cross-lingual" OR language*))`. We divided the databases among co-authors, assigning each co-author to apply the predetermined search keywords to retrieve papers from their assigned database(s). This resulted in the retrieval of 776 papers, of which 107 were duplicates. Figure 1 summarizes our search process. We conducted our final search for this review on December 16, 2025.

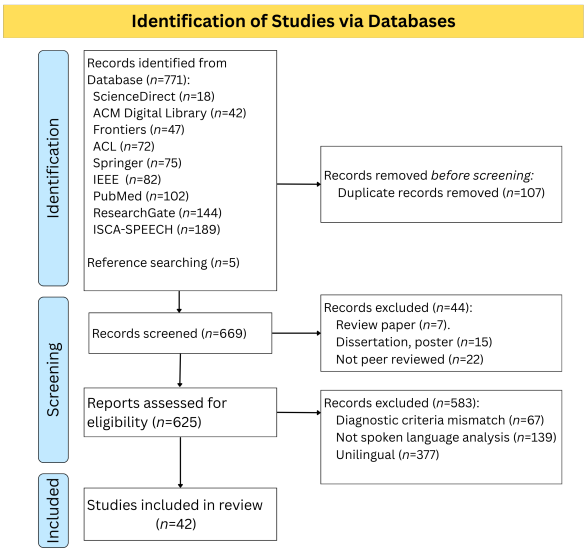

Figure 1: Study selection process, from initial search retrieval ($n = 776$) to final included studies ($n = 42$).

2.2 Step 2: Selection Process

Our exclusion criteria were: (1) studies that did not use corpora from at least two languages, considering different dialects of a language (i.e., American, Canadian, and British English (Gao et al., 2025)) as the same language; (2) studies that did not focus on speech or text data; (3) studies without AD or MCI samples (i.e., studies focusing on other neurocognitive disorders such as aphasia or Parkinson's disease); (4) papers that were not written in English; and (5) studies that were not full peer-reviewed academic papers (e.g., dissertations or posters). While it is possible that exclusion criterion (4) overlooked some relevant papers in non-English venues, setting this criterion was necessary to ensure that we were able to manually review all included studies in a fair and equitable manner. All authors were fluent English speakers, but did not have overlapping proficiency in languages other than English.

The screening process (see Figure 1) was independently undertaken by two co-authors to determine which of the remaining 669 papers to retain according to the predefined inclusion and exclusion criteria. In the first phase of screening, the

[1] https://www.isca-speech.org/archive/
[2] https://aclanthology.org/
[3] https://dl.acm.org/
[4] https://ieeexplore.ieee.org/Xplore/home.jsp
[5] https://www.sciencedirect.com/journal/computer-speech-and-language
[6] https://link.springer.com/
[7] https://www.frontiersin.org/journals/aging-neuroscience
[8] https://www.researchgate.net/
[9] MCI is the intermediate stage between expected age-related cognitive decline and Alzheimer's diagnosis. MCI may or may not ever convert to AD.

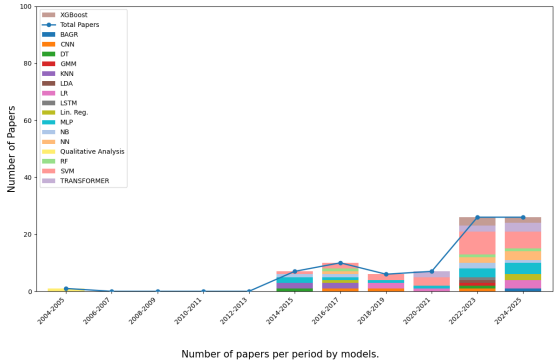

Figure 2: Frequency distribution of multilingual Alzheimer's detection studies per period by machine learning models. *SVM* are most frequently used for AD detection.

two co-authors screened the titles and abstracts excluding 44 documents, resulting in 625 remaining papers. As an outcome of full-text screening, 584 documents were excluded due to meeting predefined exclusion criteria. This resulted in a final set of 42 studies included in the review. Disagreements among two reviewers at any of the stages were discussed, and when necessary, a third author was consulted for resolution. We summarize the included papers by year of publication in Figure 2.

2.3 Collating, Summarizing, and Reporting Results

For all included papers, data was extracted by the lead author in close consultation with other authors and stored in a spreadsheet. We extracted data for each paper based on the following characteristics:

- **Dataset Characteristics:** We extracted information on the datasets used in each study, including their names, sizes, supported task types, languages, availability and access links (if provided), diagnostic label distribution, the inclusion of audio and/or transcripts, and the presence of cognitive scores (e.g., MMSE, MoCA). Results are reported in §3.1.

- **Feature Characteristics:** We collected details on the speech processing pipeline, including automatic speech recognition (ASR) for transcription, language identification, and the extraction of linguistic, acoustic, or phonetic features from both transcripts and audio. We also noted the tools and models used throughout the pipeline. More details are in §3.2.

- **Modeling Techniques:** We gathered data on the machine learning approaches used for AD prediction, including the training strategy (uni-, multi-, or crosslingual), evaluation methods, and outcomes. More details are in §3.3.

As an outcome of this process, we summarized key trends and insights from the papers across languages and levels of granularity. We also discuss system challenges along with actionable recommendations for future research in §4.

2.4 Standardization

The steps outlined above seek to enforce standardization across search and study selection, data extraction, and data synthesis and comparison. All studies were screened using the identical eligibility criteria defined a priori. The same variables were collected (when available) from each study. To measure performance, accuracy (for classification) and root mean squared error (for regression) were extracted preferentially, although other performance metrics were accepted when preferred metrics were unavailable. This approach ensured consistent evaluation of the included studies, enabling robust synthesis of evidence while accounting for language-specific challenges or differences.

3 Results

We organize our findings into subsections focusing on *Dataset Characteristics*, *Feature Characteristics*, *Modeling Techniques*, and *Challenges* for multilingual AD detection. This structure was chosen to synthesize key themes and trends emerging from the literature in each area. Our review resulted in a roadmap and broad taxonomy of methods and modeling decisions for language-agnostic and language-specific AD detection, shown in Figure 4.

3.1 Dataset Characteristics

Across the reviewed studies, several interesting patterns emerge regarding their dataset characteristics. Details regarding the datasets used in reviewed papers are in Table 1. The most prominent observation is the heavy reliance on a small set of benchmark corpora, particularly the Pitt Corpus from DementiaBank and its derivatives like ADReSS and ADReSS-M (Fraser et al., 2019a; Luz et al., 2023b, 2024b). The derivatives are balanced in terms of participant age, gender, and education, making them popular choices for AD detection.

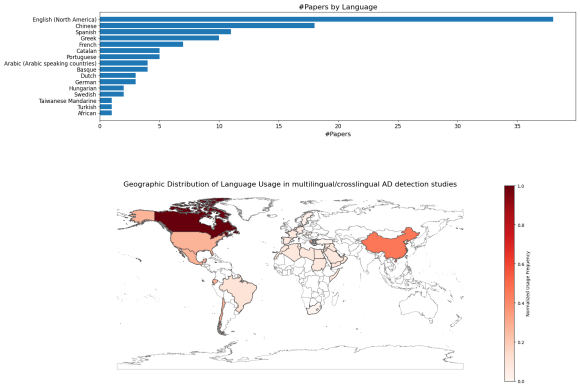

Figure 3: Language distribution in multilingual AD studies. Frequency = studies. North America = English; Arabic regions shown due to unspecified origins.

Although the Pitt Corpus and ADReSS are English-only, the ADReSS-M corpus includes both English and Greek recordings (Jin et al., 2023; Tamm et al., 2023). We show the distribution of datasets used in the papers across languages in Figure 3. We observed that most non-English datasets are small, consisting of fewer than 50 participants. Aside from the shared task datasets (e.g. ADReSS-M or TAUKADIAL), most are unbalanced in demographics (Gosztolya et al., 2021; Santos et al., 2017; Pérez-Toro et al., 2022). This imbalance limits generalizability and raises concerns that differences in speech might actually reflect variations in age or education level rather than in disease status itself.

A second consistent pattern involves the types of tasks used to elicit speech. Picture description tasks, particularly using the Cookie Theft Picture, are the most common, appearing in the majority of datasets (Fraser et al., 2019b; Santos et al., 2017; Pérez-Toro et al., 2022). Other frequent tasks include semantic verbal fluency (e.g., naming animals within a time limit), story narration (like the Cinderella story), and narrative recall (immediate or delayed retelling of short stories) (Santos et al., 2017). Some corpora also incorporate reading-aloud tasks, though these are less common. Picture description and verbal fluency likely dominate due to their diagnostic sensitivity and ease of administration. Many datasets focus on one or two tasks, which facilitates consistency in feature extraction (Martinez et al., 2017; Fraser et al., 2019b).

The majority of datasets include audio recordings, and some also provide transcripts, either manually created or derived using ASR. Some datasets offer additional multimodal content, including video or longitudinal follow-ups (Ortiz et al., 2024; Ablimit et al., 2022; Barrera et al., 2024). Most speech data was gathered through in-person assessments, often in clinical or research lab settings. While some well-known corpora are available through repositories like DementiaBank, many datasets are only accessible by request, and several do not report availability at all (Pérez-Toro et al., 2023, 2022; Martinez et al., 2017). This inconsistency in sharing practices makes it difficult to compare models and slows progress toward more reproducible, generalizable AD detection systems. There is a clear need for more accessible, diverse, and demographically balanced datasets to support robust multilingual research in this space.

3.2 Feature Characteristics

Many multilingual AD detection pipelines develop rich feature extractors. In general, these feature extractors are either (1) language-agnostic or (2) language-specific. Language-agnostic pipelines are designed to be independent of specific languages and capture general patterns present in the speech data. Language-specific pipelines first identify the language of the data, and then extract language-dependent information (e.g., linguistic features). As shown in the roadmap followed by multilingual and crosslingual AD detection papers (Figure 4), the process begins with raw speech input that is preprocessed and converted into acoustic signals. Language-agnostic and language-specific features are extracted from the preprocessed data. These features are then used to train multilingual, crosslingual, or monolingual models; feature selection and optional model ensembling are applied prior to language-agnostic or language-specific inference.

3.2.1 Language-Agnostic Features

Language-agnostic features are often acoustic, and can be extracted using engineered or supervised/self-supervised processes. Audio preprocessing involves sampling the speech signal at a fixed rate, normalizing volume, LLM-based anomaly detection, and chunking utterances in overlapping frames. Studies leveraging engineered acoustic features focused on spectrogram-based MFCC features (Favaro et al., 2024), disfluency/silence features using voice activity detection (VAD) (Mei et al., 2023; Chen et al., 2023b), speech timing features (Pérez et al., 2024), speech tempo (Luz et al., 2024b; Mei et al., 2023; Lopez et al., 2013; Gosztolya et al., 2021; Kálmán et al., 2022; Luz et al., 2024b), emotional re-

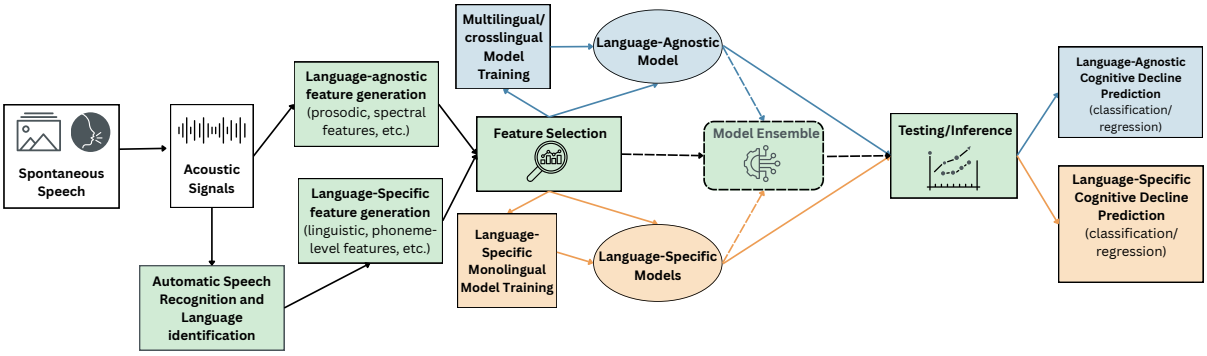

Figure 4: Diagram of the modeling framework highlighting language-agnostic and language-specific pathways for cognitive decline prediction. White boxes indicate input, green boxes show shared processing steps, blue and orange boxes represent language-agnostic and language-specific paths, respectively, and dotted outlines are optional.

sponse (Pérez-Toro et al., 2023), temporal features, (Lopez et al., 2014; Martinez et al., 2017; Kálmán et al., 2022), acoustic embeddings (Jin et al., 2023), and generated acoustic and paralinguistic features (Schäfer et al., 2022; Lindsay et al., 2021a; Mei et al., 2023; Lindsay et al., 2021b). Non-acoustic engineered features focused on patient-level demographic characteristics, such as age, gender, and education level (Mei et al., 2023; Jia et al., 2025a; Azadmaleki et al., 2025).

Supervised and self-supervised features are typically generated using pre-trained language models, for example by generating high-dimensional acoustic embeddings from preprocessed audio waveforms. Popular tools for generating these types of feature have included Whisper (Radford et al., 2023), Wav2vec 2.0 (Baevski et al., 2020), multilingual pre-trained XLSR-53 (Conneau et al., 2021) HuBERT (Hsu et al., 2021), and Prosody UnitY2 (Seamless et al., 2023). Multilingual ASR (e.g., Whisper) enabled automatic transcription that, when paired with prompting, supports a scalable and language-agnostic AD detection pipeline (Jia et al., 2025a). These tools were used across many studies (Luz et al., 2024a; Gosztolya and Tóth, 2024; Pérez et al., 2024; Pérez-Toro et al., 2022; Jin et al., 2023; Mei et al., 2023; Pérez-Toro et al., 2023; Duan et al., 2024; Jia et al., 2025a).

3.2.2 Language-Specific Features

To first identify language, language-specific feature extractors often leverage deep learning-based ASR systems (Agbavor and Lia, 2024; Luz et al., 2024a; Duan et al., 2024; Gosztolya et al., 2021; Pérez et al., 2024). 15 studies used ASR to generate either phone-level transcripts with nonverbal labels (i.e., filled pauses or coughs) (Gosztolya et al., 2021) or token-level transcripts from which text features were extracted (Agbavor and Lia, 2024; Pérez et al., 2024). Whisper ASR was used to capture speech patterns using fine-grained transcriptions with pauses and filler words (Duan et al., 2024).

Language-specific textual feature extraction involved both supervised/self-supervised and engineered features. Popular multilingual pre-trained language models for this purpose included XLM-RoBERTa-base (Conneau et al., 2020), Distilbert-base-multilingual-cased (Sanh et al., 2019), Text2vec-base-multilingual (Xu, 2023), XLM-Roberta-Large-Vit-L-14 (Radford et al., 2021), LaBSE (Feng et al., 2022), Lealla (Mao and Nakagawa, 2023), and Multilingual-e5-large (Wang et al., 2024). Language-specific versions of BERT (Devlin et al., 2019), RoBERTA (Zhuang et al., 2021), Voyage-large-v2[10], and FLag[11] were also used to generate fixed-length embeddings from ASR-generated token-level transcripts (Agbavor and Lia, 2024; Duan et al., 2024).

Engineered features typically captured lexicosyntactic and lexicosemantic information (Favaro et al., 2024; Fraser et al., 2019a). For instance, Li et al. (2019) transformed Mandarin Chinese feature vectors to English feature vectors using a parallel out-of-domain corpus to learn mappings between the two languages. To address crosslingual distributional shift in feature spaces, Hoang et al. (2024) applied crosslingual alignment by extracting embeddings from backtranslated text while maintaining the semantic information

[10] https://huggingface.co/voyageai/voyage-large-2-instruct/tree/57bd79078480bd6a36669407bfbc59bb4ac1ddcf

[11] https://huggingface.co/BAAI/bge-large-zh-v1.5

between two languages. Domain adaptation was also applied in crosslingual settings (Fraser et al., 2019a). To select the best features across languages, Li et al. (2019) applied joint feature selection, and Martinez et al. (2017); Kálmán et al. (2022); Lindsay et al. (2021b,b); **?** used the nonparametric Mann-Whitney U test or Kruskal–Wallis non-corrected significance test.

3.3 Modeling Techniques

The papers formulated their modeling tasks as either: (1) classification (16 studies categorized AD versus healthy control subjects, 17 studies categorized MCI versus healthy subjects, and 2 studies performed multiclass classification across different pre-AD and AD stages); or (2) regression (14 studies performed MMSE score prediction). All papers used supervised learning for classification and regression. Figure 2 summarizes the machine learning models applied across papers. At a broad level, *language-agnostic modeling* involved training a single model on the feature set irrespective of the language of the data, whereas *language-specific modeling* tailored the modeling process to specific languages after language identification and feature extraction. During inference, language-specific pipelines were also activated for the final prediction dependent on the identified language.

Overall, we observed three training settings: (1) *multilingual* model training and testing, in which two or more languages were incorporated in the same pool of training samples; (2) *crosslingual* training, in which the model was trained in one language and tested in another; and (3) *monolingual* training, in which different models were trained for different languages using different training sets, and then were applied to test instances dependent on the sample's identified language. Language-specific modeling uses monolingual training, whereas language-agnostic modeling may use multilingual or crosslingual training.

3.3.1 Language-Agnostic Modeling

Preliminary studies show promise in using crosslingual data for AD detection (Fraser et al., 2019a; Guo et al., 2020; Duan et al., 2024; Li et al., 2019; Pérez-Toro et al., 2023); when transforming data to a shared domain with concept-based language model features, AUCs of 89% (French) and 64% (English) were reported (Fraser et al., 2019a). Li et al. (2019) used a large parallel corpus to align lexicosyntactic features in English and Mandarin, attaining Spearman $\rho = 0.549$ and outperforming monolingual models. BERT-based monolingual representations with a contrastive autoencoder reached 81.6% accuracy in Mandarin AD detection (Guo et al., 2020).

Multilingual training with multimodal features has recently outperformed crosslingual models (UAR up to 70% for English–Spanish) (Pérez-Toro et al., 2023). Multilingual training with backtranslated embeddings achieved 45.1% balanced accuracy for MCI and RMSE=2.578 for MMSE prediction in TAUKADIAL (Hoang et al., 2024). Overall, English-to-other-language crosslingual performance (e.g., Hungarian, French, Spanish) is better, likely due to larger English datasets.

Since data availability tends to be especially low for non-English AD detection (see Table 1), many reviewed papers explored the use of external data to boost performance (Fraser et al., 2019a; Duan et al., 2024; Lindsay et al., 2021a; Guo et al., 2020). Increasing the amount of source data when the target domain data was small was found to improve performance in multilingual training with domain adaptation (Fraser et al., 2019a; Duan et al., 2024). Lindsay et al. (2021a) found that crosslingual increase in clinical data (combining English and Swedish MCI samples) was more effective than simply increasing language-specific healthy control data. A similar finding was that the addition of bilingual representation for data learning in the cross language improved performance over other methods when using a large general domain English corpus (Guo et al., 2020).

3.3.2 Language-Specific Modeling

Ten papers focused on language-specific modeling, either aggregating performance across languages or reporting language-specific performance separately (Agbavor and Lia, 2024; Favaro et al., 2024; Tsai et al., 2021; Santos et al., 2017; Schäfer et al., 2022; Gosztolya et al., 2021; Tsai et al., 2021; Kálmán et al., 2022; Pérez-Toro et al., 2022; Lindsay et al., 2021a). Language-specific ensembles of models ranked first and second in the TAUKADIAL challenge for MMSE prediction and MCI classification respectively (Agbavor and Lia, 2024). Monolingual training of models (NN, SVR, XGB Regressor (XGBR), and Bagging Regressor (BAGR)) followed by ensembling also achieved their best accuracy (75.0%; RMSE=2.44) over multilingual training (Favaro et al., 2024).

When comparing performance across languages,

several studies reported better performance on a non-English dataset (i.e., Chinese, Portuguese, and Hungarian) than in English (Tsai et al., 2021; Santos et al., 2017; Gosztolya et al., 2021). A paper performing AD detection in three languages achieved the best AUC in German (compared to French and Dutch) when training an SVM with different linguistic embeddings (Lindsay et al., 2021a). Another study reported that monolingual performance in both English and Spanish ($F_1 = 81\%$ both) was better than multilingual training ($F_1 = 78\%$) for AD classification (Pérez-Toro et al., 2022).

With monolingual training, strong language dependencies in English were observed for linguistic embeddings, whereas a lower language dependency was observed for acoustic embeddings (facilitating stronger performance in Spanish) while detecting AD across English and Spanish samples. Spanish's simpler phonetics may explain this, but language-specific grammar and semantics warrant further study(Pérez-Toro et al., 2022). In addition to feature dependencies and multilingual model training, narrative length plays a role: SVM was found to perform the highest for English and Portuguese datasets with medium-length narratives, but did not outperform much simpler models for very short transcripts (Santos et al., 2017).

3.3.3 Ensemble Models

Finally, ensemble models generally led to improved multilingual AD classification outcomes. Ensemble models combine the predictions of multiple algorithms, with the idea that multiple weak learners merge to form a stronger prediction. Seven of the reviewed papers used ensemble approaches. In the TAUKADIAL (Favaro et al., 2024) and ADReSSm (Luz et al., 2023b) challenges, the best performing models applied ensembles of language-specific models (Agbavor and Lia, 2024; Jin et al., 2023). Most ensemble approaches used majority voting for classification and averaging for regression (Favaro et al., 2024; Chen et al., 2023b; Mei et al., 2023; Lindsay et al., 2024; **?**). CONSEN (Jin et al., 2023) updated AD and MMSE scores complementarily and simultaneously until a threshold was reached, which outperformed more standard ensembling in crosslingual settings but not in multilingual settings (Jin et al., 2023; Favaro et al., 2024).

4 Discussion

Our review revealed common limitations across multilingual AD detection settings. We summarize those here, followed by recommendations for addressing them in future work.

Limited Generalization across Languages. The papers reported strong performance from language-specific monolingual models combined through ensemble approaches; however, accuracy and MMSE prediction performance both declined sharply when the models were applied to unseen languages. This highlights the challenge of achieving generalization across languages without adaptation or alignment. Due to limited non-English corpora, Fraser et al. (2019a) showed that domain adaptation enables data-efficient learning, allowing models trained on high-resource languages like English to transfer effectively to low-resource ones. Multilingual training often gains more from adding data in new languages than from additional same-language data. While fine-tuning multilingual pretrained models improves performance on represented—especially low-resource—languages, frozen pretrained models typically exhibit stronger generalization to under-represented or unseen languages, revealing a trade-off between in-language gains and cross-lingual robustness (Jia et al., 2025a). Five papers found that monolingual and multilingual models outperformed crosslingual alternatives, particularly when the language tasks differed across languages. Prosodic features were most relevant for AD/MCI detection in mono- and multilingual setups, but less transferable in crosslingual settings. However, silence or pause-based acoustic features remain relatively robust across languages, as reported by 13 papers. Embedding-level multimodal fusion appears most promising for multilingual and crosslingual training. Eight studies show that fusing acoustic and linguistic features (via concatenation or self-attention) enhances multilingual MMSE prediction (RMSE 1.87 in TAUKADIAL) and MCI/AD classification (UAR 83%), though some report better MCI performance from acoustic-only models.

Risk of Bias. Achieving balance across class, age, gender, and education in multilingual datasets is crucial for unbiased detection of AD, yet only 8 corpora in the reviewed papers meet this criterion for all or some factors. To handle class imbalance, studies commonly employ subsampling, stratified cross-validation, and evaluation metrics such as UAR or AUROC. Two studies with imbalanced data report accuracy only. Thirty-six studies use cross-validation and/or held-out sets to prevent overfitting, but nine multilingual or crosslingual

studies rely on small (<50) non-English datasets, making tuning and evaluation difficult. Except for TAUKADIAL (Luz et al., 2024a), ADReSS-M (Luz et al., 2024b), PREPARE (Azadmaleki et al., 2025), ADReSS (Luz et al., 2020a), ADReSSo (Luz et al., 2021b), NCMMSC (Chen et al., 2023a), no reviewed corpus includes a held-out test set, indicating a substantial risk of model overfitting.

Actionable Recommendations. Based on our findings in the review, we recommend that several topics take priority in near-future multilingual AD detection research. Addressing these points will foster easier, more accessible translation of existing research progress to global clinical practice.

First, we recommend immediate focus on creating a diverse, **multilingual, multimodal benchmark dataset** (extending efforts such as those behind TAUKADIAL) that is heterogeneous in terms of language resources, diagnoses (e.g., different stages of AD and/or MCI), and cognitive test scores. The benchmark dataset should be balanced in terms of age, gender, and education. A larger, heterogeneous benchmark dataset would support more rigorous evaluation practices and enable the assessment of model performance across linguistically diverse populations, contributing to the development of more equitable and generalizable methods. Moreover, rich multimodal datasets spanning speech, video, clinical records, imaging, social determinants of health, and digital text can enable LLM-based pipelines, and detection models to support scalable Alzheimer's disease screening and longitudinal monitoring in clinical settings.

Second, the majority of the corpora ($n = 14$) in reviewed papers used semi-spontaneous speech tasks (e.g., picture description) to elicit language data. Although such tasks are effective as a stimulus compared to verbal fluency tasks, they often need to be adapted according to the country and culture of the participants, which limits broader generalization. We recommend that future multilingual AD research should focus on the **spontaneous speech** present in natural conversations. This data would be more straightforward to collect passively, continuously, and longitudinally without need for adaptation to new language collection settings.

Third, focusing on cohorts already diagnosed with AD is less useful from the perspective of eventual practical deployment, and the reviewed studies generally reported strong performance on AD detection already. Future work should prioritize **early onset detection**, refocusing on identifying preclinical stages of AD. Multilingual corpora should integrate early biomarkers such as cerebrospinal fluid and blood-based plasma markers, which are valuable indicators of early AD. Multilingual speech features linked to biomarkers (SLaCAD (Farzana et al., 2024)) may support less invasive, more globally accessible early detection.

Finally, although most reviewed papers claim some level of automation, only 18 employ a fully automatic pipeline spanning LLM-enabled audio processing, ASR-based transcription, and downstream classification or regression. The rest use partial automation, typically for feature extraction or classification, while relying on manual transcription, audio processing, or result evaluation. Across studies, common acoustic features (ComParE, eGeMAPS, emobase, and acoustic embeddings) and text-based features (word embeddings and frequency-based metrics such as type–token ratio) are observed, yet overall feature sets remain highly heterogeneous. Automation and generalization are further hindered by the need for language identification and speech-to-text processing, especially in low-resource languages. Although eight studies demonstrate the effectiveness of AI-based remote cognitive health monitoring, designing such platforms requires standardized, user-friendly, language-agnostic, and transcription-free end-to-end systems. We recommend an explicit focus on **standardization and automation** in near-term multilingual AD detection research.

5 Conclusion

In this work, we conducted the first scoping review on automated multilingual and crosslingual AD detection. Given the prevalence of studies on general AD detection, surprisingly limited work has been done on multilingual and crosslingual approaches, likely because most resources for training AD detection models are in English (de la Fuente Garcia et al., 2020). After careful review of 42 papers meeting inclusion criteria from among 776 retrieved, we find promise of a field that could beneficially transform clinical practice. Almost all studies report relatively high performance, despite the difficulties inherent to this under-resourced space. This leaves us optimistic for eventual real-world global integration of automated AD detection systems, especially with targeted focus toward the identified actionable recommendations.

Limitations

Our work is limited by several factors. First, we conduct the survey for a specific time period (2004 to 2025), limited to databases relevant to the speech processing and medical domains. Despite our attempts at thorough coverage of these databases, we may have missed some relevant sources, limiting the breadth of our survey.

Second, when studying the reproducibility of the reviewed work, we were unable to report the availability of all data sources and code repositories since some were not publicly available. We tried to contact the respective authors of such works, which in some cases yielded helpful new information, but was not always successful.

Third, we did not perform an independent risk assessment for the included papers. A compelling direction for future work, although beyond the scope of this scoping review, may be to operationalize a systematic review tool such as QUADAS-2 (Whiting et al., 2011) for the AI context to assess items such as whether different varieties of included languages are adequately represented, whether AD diagnoses are based on clinical assessments and (if so) whether the diagnostic criteria are consistent across languages, whether sociolinguistic confounds are acknowledged, and many other factors. Finally, it is unclear to what extent the findings in this survey can extend to languages for which no corpora were identified in reviewed papers. Pending the pursuit of this research topic by an increasingly global community, an interesting avenue for follow-up work would be to validate these findings at a future time horizon, when a greater variety of languages are hopefully represented.

Ethical Considerations

Automated models for AD detection from spoken language present potential benefits in real-world scenarios: they offer opportunity to expand healthcare access, minimize cost of care, and reduce caregiver burden. However, they may also pose risks if used in unintended ways. We consider intended use of our study findings to extend to the following:

- People may use these findings to study language differences between individuals with and without AD, as a way of building further understanding of the condition.
- People may use the findings developed in this work to further their own research into low-resource NLP tasks, including those associated with this and other healthcare problems.
- People may use the findings developed in this work to build early warning systems to flag individuals about potential AD symptoms, provided that the technology is not misconstrued as an alternative to human care.

We reiterate our caution against building systems that function as intended or perceived replacements for human medical care. Future research pursuing these goals should be done in careful coordination with clinical professionals and other stakeholders.

References

Ayimnisagul Ablimit, Catarina Botelho, Alberto Abad, Tanja Schultz, and Isabel Trancoso. 2022. Exploring dementia detection from speech: Cross corpus analysis. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Felix Agbavor and Hualou Lia. 2024. Multilingual prediction of cognitive impairment with large language models and speech analysis. *Brain Sciences*, 14:1292.

Hossein Azadmaleki, Yasaman Haghbin, Sina Rashidi, Mohammad Nezhad, Ali Zolnour, and Maryam Zolnoori. 2025. Speechcare: dynamic multimodal modeling for cognitive screening in diverse linguistic and speech task contexts. *npj Digital Medicine*, 8.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint*.

Anees Bahji, Laura Acion, Anne-Marie Laslett, and Bryon Adinoff. 2023. Exclusion of the non-english-speaking world from the scientific literature: Recommendations for change for addiction journals and publishers. *Nordic Studies on Alcohol and Drugs*, 40(1):6–13.

Benjamin Barrera, Daeun Lee, Zaima Zarnaz, Jinyoung Han, and Seungbae Kim. 2024. The interspeech 2024 taukadial challenge: Multilingual mild cognitive impairment detection with multimodal approach. In *Interspeech 2024*, pages 967–971.

Tracy G. Beckett. 2004. Language and dementia in bilingual settings: Evidence from two case studies. Master's thesis, University of Cape Town. Master's thesis.

Edward Campbell, RaÃºl YÃ¡Ã±ez MesÃa, Laura Docio-Fernandez, and Carmen GarcÃa-Mateo. 2021. Paralinguistic and linguistic fluency features for alzheimer disease detection. *Computer Speech & Language*, 68:101198.

B. Ceyhan, S. Bek, and T. Önal Süzek. 2024. Machine learning-based prediction models for cognitive decline progression: A comparative study in multilingual settings using speech analysis. *JAR Life*, 13:43–50.

Xu-Chu Chen, Wei-Qiang Zhang, and Yong Ma. 2023a. Raw waveform-based end-to-end Alzheimer's disease detection method. *Acta Electronica Sinica*, 51(12):3582–3590.

Xuchu Chen, Yu Pu, Jinpeng Li, and Wei-Qiang Zhang. 2023b. Cross-lingual alzheimer's disease detection based on paralinguistic and pre-trained features. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sofia de la Fuente Garcia, Craig Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 78:1–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. 2024. Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):325.

Junwen Duan, Fangyuan Wei, Hong-Dong Li, and Jin Liu. 2024. Pre-trained feature fusion and matching for mild cognitive impairment detection. In *Proc. Interspeech 2024*, pages 962–966.

Shahla Farzana and Natalie Parde. 2023. Towards domain-agnostic and domain-adaptive dementia detection from spoken language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11965–11978, Toronto, Canada. Association for Computational Linguistics.

Shahla Farzana, Edoardo Stoppa, Alex Leow, Tamar Gollan, Raeanne Moore, David Salmon, Douglas Galasko, Erin Sundermann, and Natalie Parde. 2024. SLaCAD: A spoken language corpus for early Alzheimer's disease detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14877–14897, Torino, Italia. ELRA and ICCL.

Anna Favaro, Tianyu Cao, Najim Dehak, and Laureano Moro-Velazquez. 2024. Leveraging universal speech representations for detecting and assessing the severity of mild cognitive impairment across languages. In *Proc. Interspeech 2024*, pages 972–976.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kathleen C. Fraser, Nicklas Linz, Bai Li, Kristina Lundholm Fors, Frank Rudzicz, Alexandra König, Jan Alexandersson, Philippe Robert, and Dimitrios Kokkinakis. 2019a. Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3659–3670, Minneapolis, Minnesota. Association for Computational Linguistics.

Kathleen C. Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2019b. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139.

Kunxiao Gao, Anna Favaro, Najim Dehak, and Laureano Moro-Velázquez. 2025. Adceleb: A longitudinal speech dataset from public figures for early detection of alzheimer's disease. *Interspeech 2025*.

Dimitris Gkoumas, Matthew Purver, and Maria Liakata. 2023. Reformulating NLP tasks to capture longitudinal manifestation of language disorders in people with dementia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15904–15917, Singapore. Association for Computational Linguistics.

Gábor Gosztolya, Réka Balogh, Nóra Imre, José Vicente Egas-López, Ildikó Hoffmann, Veronika Vincze, László Tóth, Davangere P. Devanand, Magdolna Pákáski, and János Kálmán. 2021. Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech. *Computer Speech Language*, 69:101215.

Gábor Gosztolya and László Tóth. 2024. Combining acoustic feature sets for detecting mild cognitive impairment in the interspeech'24 taukadial challenge. pages 957–961.

Zhiqiang Guo, Zhaoci Liu, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. 2020. Text classification by contrastive learning and cross-lingual data augmentation for Alzheimer's disease detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6161–6171, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ihab Hajjar, Maureen Okafor, Jinho D Choi, Elliot Moore, Anees Abrol, Vince D Calhoun, and Felicia C Goldstein. 2023. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 15(1):e12393.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Laura Hernández-Domínguez, Sylvie Ratté, Boyd Davis, and Charlene Pope. 2016. Conversing with the elderly in Latin America: a new cohort for multimodal, multilingual longitudinal studies on aging. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 16–21, Berlin. Association for Computational Linguistics.

Bao Hoang, Yijiang Pang, Hiroko Dodge, and Jiayu Zhou. 2024. Translingual language markers for cognitive assessment from spontaneous speech. In *Proc. Interspeech 2024*, pages 977–981.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Kaichen Jia, Jinpeng Li, Ke Li, and Wei-Qiang Zhang. 2025a. Whisper-based multilingual alzheimer's disease detection and improvements for low-resource language. pages 549–553.

Kaichen Jia, Jinpeng Li, Ke Li, and Wei-Qiang Zhang. 2025b. Whisper-based multilingual alzheimer's disease detection and improvements for low-resource language. pages 549–553.

Longbin Jin, Yealim Oh, Hyunseo Kim, Hyuntaek Jung, Hyo Jin Jon, Jung Eun Shin, and Eun Yi Kim. 2023. Consen: Complementary and simultaneous ensemble for alzheimer's disease detection and mmse score prediction. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2.

János Kálmán, Davangere P Devanand, Gábor Gosztolya, Réka Balogh, Nóra Imre, László Tóth, Ildikó Hoffmann, Ildikó Kovács, Veronika Vincze, and Magdolna Pákáski. 2022. Temporal speech parameters detect mild cognitive impairment in different languages: Validation and comparison of the speech-gap test® in english and hungarian. *Current Alzheimer Research*, 19:373–386.

Bai Li, Yi-Te Hsu, and Frank Rudzicz. 2019. Detecting dementia in Mandarin Chinese using transfer learning from a parallel corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1991–1997, Minneapolis, Minnesota. Association for Computational Linguistics.

Hali Lindsay, Giorgia Albertin, Louisa Schwed, Nicklas Linz, and Johannes Tröger. 2024. Cross-lingual examination of language features and cognitive scores from free speech. In *Proceedings of the Fifth Workshop on Resources and Processing of linguistic, paralinguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 16–25, Torino, Italia. ELRA and ICCL.

Hali Lindsay, Philipp Müller, Insa Kröger, Johannes Tröger, Nicklas Linz, Alexandra Konig, Radia Zeghari, Frans RJ Verhey, and Inez HGB Ramakers. 2021a. Multilingual learning for mild cognitive impairment screening from a clinical speech task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 830–838, Held Online. INCOMA Ltd.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021b. Language impairment in alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in Aging Neuroscience*, Volume 13 - 2021.

Karmele Lopez, Jes Alonso, Jordi SolCasals, Nora Barroso, Patricia HenrÃquez RodrÃguez, Marcos Faundez-Zanuy, Carlos Travieso, M. Ecay, Pablo Martinez-Lage, Unai Martinez-de Lizarduy, Harkaitz Eguiraun Martinez, and A. Ezeiza. 2013. On automatic diagnosis of alzheimerâ€™s disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7.

Karmele Lopez, JesÃºs Alonso, Carlos Travieso, Jordi Sol-Casals, Aitzol Ezeiza, Marcos Faundez-Zanuy, Blanca Beitia, and Pilar Calvo. 2014. Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of alzheimer's disease. *Neurocomputing*, 150.

Karmele Lopez, Unai Martinez-de Lizarduy, Pilar M Calvo, Jiri Mekyska, Blanca Beitia, Nora Barroso, Ainara Estanga, Milkel Tainta, and Mirian Ecay-Torres. 2018. Advances on automatic speech analysis for early detection of alzheimer disease: A non-linear multi-task approach. *Curr Alzheimer Res*, 15(2):139–148.

Saturnino Luz, Sofia de la Fuente Garcia, Fasih Haider, Davida Fromm, Brian Macwhinney, Alyssa Lanzi, Ya-Ning Chang, Una Chou, and Yi-Chien Liu. 2024a. Connected speech-based cognitive assessment in chinese and english.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020a. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of INTERSPEECH 2020*, Shanghai, China.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020b. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021a. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Proc. Interspeech 2021*, pages 3780–3784.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021b. Detecting cognitive decline using speech only: The adresso challenge. *medRxiv*.

Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. 2023a. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. *arXiv preprint*.

Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian Macwhinney. 2023b. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge.

Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. 2024b. An overview of the adress-m signal processing grand challenge on multilingual alzheimer's dementia recognition through spontaneous speech. *IEEE open journal of signal processing*, 5:738 – 749.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Unai Martinez, Pilar SalomÃ³n, Pedro Gomez, Mirian Torres, and Karmele Lopez. 2017. Alzumeric: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens*, 4:037.

Kangdi Mei, Xinyun Ding, Yinlong Liu, Zhiqiang Guo, Feiyang Xu, Xin Li, Tuya Naren, Jiahong Yuan, and Zhenhua Ling. 2023. The ustc system for adress-m challenge. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2.

Thomas Melistas, Lefteris Kapelonis, Nikos Antoniou, Petros Mitseas, Dimitris Sgouropoulos, Theodoros Giannakopoulos, Athanasios Katsamanis, and Shrikanth Narayanan. 2023. Cross-Lingual Features for Alzheimer's Dementia Detection from Speech. In *Proc. INTERSPEECH 2023*, pages 3008–3012.

David Ortiz, José Rodríguez, and David Tomás. 2024. Cognitive insights across languages: Enhancing multimodal interview analysis. pages 952–956.

Paula Andrea Pérez, Tomás Arias-Vergara, Philipp Klumpp, Tobias Weise, Maria Schuster, Elmar Noeth, Juan Rafael Orozco-Arroyave, and Andreas Maier. 2024. Multilingual speech and language analysis for the assessment of mild cognitive impairment: Outcomes from the taukadial challenge. *Proc. Interspeech 2024*, pages 982–986.

Paula A. Pérez-Toro, Tomás Arias-Vergara, Franziska Braun, Florian Hönig, Carlos A. Tobón-Quintero, David Aguillón, Francisco Lopera, Liliana Hincapié-Henao, Maria Schuster, Korbinian Riedhammer, Andreas Maier, Elmar Nöth, and Juan Rafael Orozco-Arroyave. 2023. Automatic Assessment of Alzheimer's across Three Languages Using Speech and Language Features. In *Proc. INTERSPEECH 2023*, pages 1748–1752.

Paula Andrea Pérez-Toro, Philipp Klumpp, Abner Hernandez, Tomas Arias, Patricia Lillo, Andrea Slachevsky, Adolfo Martín García, Maria Schuster, Andreas K. Maier, Elmar Noeth, and Juan Rafael Orozco-Arroyave. 2022. Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings. In *Proc. Interspeech 2022*, pages 2483–2487.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Leandro Santos, Edilson Anselmo Corrêa Júnior, Osvaldo Oliveira Jr, Diego Amancio, Letícia Mansur, and Sandra Aluísio. 2017. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1284–1296, Vancouver, Canada. Association for Computational Linguistics.

Simona Schäfer, Elisa Mallick, Louisa Schwed, Alexandra König, Jian Zhao, Nicklas Linz, Timothy Hadarsson Bodin, Johan Skoog, Nina Possemis, Daphne ter Huurne, Anna Zettergren, Silke Kern, Simona Sacuiu, Inez Ramakers, Ingmar Skoog, and Johannes Tröger. 2022. Screening for mild cognitive impairment using a machine learning classifier and the remote speech biomarker for cognition: Evidence from two clinically relevant cohorts. *Journal of Alzheimer s Disease*, 91:1165–1171.

Seamless, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bastiaan Tamm, Rik Vandenberghe, and Hugo Van Hamme. 2023. Cross-lingual transfer learning for alzheimer's detection from spontaneous speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2.

Andrea C Tricco, Erin Lillie, Wasifa Zarin, Kelly K O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah DJ Peters, Tanya Horsley, Laura Weeks, and 1 others. 2018. Prisma extension for scoping reviews (prisma-scr): checklist and explanation. *Annals of internal medicine*, 169(7):467–473.

Austin Tsai, Sheng-Yi Hong, Lihung Yao, Wei-Der Chang, Li-Chen Fu, and Yu-Ling Chang. 2021. An efficient context-aware screening system for alzheimer's disease based on neuropsychology test. *Scientific Reports*, 11:18570.

Rohit Voleti, Julie Liss, and Visar Berisha. 2019. A review of automated speech and language features for assessment of cognition and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *CoRR*, abs/2401.00368.

Penny F Whiting, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, Patrick MM Bossuyt, and QUADAS-2 Group*. 2011. Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, 155(8):529–536.

Ming Xu. 2023. text2vec: A tool for text to vector.

Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhen-Hua Ling. 2022. Deep learning-based speech analysis for alzheimer's disease detection: a literature review. *Alzheimer's Research & Therapy*, 14.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

Table 1: Detailed Data information. Abbreviations: *domain*:dom., *Gothenburg*:Got., *Karolinska*:kar, *OpenSubtitles*:OS., *available*:avail., *not available*:NA, *Speech Biomarker for Cognition*: SB-C, *Mini Mental State Examination*: MMSE, *Classification*: cls., *Regression*:reg., age associated cognitive decline:AACD, *Narrative recall*:NR, *Biographical interview*:BI, *Reading task*: RT, *Logical memory test*: LM, *Cookie theft picture description*:CTP, *Cat rescue picture description*:CR, *Coming & going picture description*:C&G, *Lion lying with a cub in the desert while eating.*:LWC, *Semantic verbal fluency*:SVF, *Voice Assistant Interactions*:VAI, Language code: *English*:EN, *Spanish*:ES, *Swedish*:SWE, *Chilean Spanish*:ES-CL, *Mandarin Chinese*:Mnd. CHI, *Taiwanese Mandarin Chinese*: Tai. Mnd., *Hungarian*:HU, *Greek*:GRK, *German*:GER, *Dutch*:DE, *Catalan*:CAT, *Basque*:BSQ, *Arabic*:ARB, *French*:FR, *Portuguese*:PRT, *Turkish*:TR*African*:AFR

Study	Database, task & balance	Data annotation	Task type (Cls./Reg.)	Data avail.	Language
(Pérez et al., 2024; Agbavor and Lia, 2024; Luz et al., 2024a; Ortiz et al., 2024; Gosztolya and Tóth, 2024; Duan et al., 2024; Favaro et al., 2024; Hoang et al., 2024; Barrera et al., 2024)	**TAUKADIAL** (*n*=507(train/ test:/387 (129 subjects)/120 (40 subjects), EN:PD (CTP, CR, C&G); CHI: PD (3 pictures on Taiwanese culture) (rec.)); age & gender balanced	MCI/HC: train:(222/165), test:(63/57);MMSE: avail.	MCI & MMSE reg.	avail.[12]	train & test: EN (246), CHI (261).
(Luz et al., 2023b; Jin et al., 2023; Tamm et al., 2023; Mei et al., 2023; Melistas et al., 2023; Chen et al., 2023b; Luz et al., 2024b; **?**)	**ADReSS-M** (*n*=279(train/dev/ test:225/8/46, PD (CTP, LWC) (rec.)); age, gender balanced and adjusted for education	AD/HC: train:(110/115), dev:(4/4), test:(22/24);MMSE: avail.	AD vs. HC & MMSE reg.	avail.[22]	train: EN, dev:GRK, test:GRK.
(Santos et al., 2017)	**Pitt**(*n*=86, PD (CTP)), **CN** (*n*=40, SN), **ABCD** (*n=43*, NR), rec. & tr. for all; balanced: no	HC/MCI: Pitt: (43/43), CN: (20/20), ABCD: (20/23); Pitt:MMSE, CN:NA, ABCD:NA;	MCI vs. HC	CN, ABCD: avail. as upon request; Pitt: avail.[22]	Pitt: EN, CN: PT-BR, ABCD:PT-BR
(Fraser et al., 2019b)	**Got.**:*n*=67, **Kar.**(*n*=96, SS (Written)) **Pitt**(*n*=116, PD(CTP) (rec. & tr.)); age and gender balanced (Pitt), education adjusted (Pitt, Got., Kar.).	Got / Kar / Pitt: HC: 6/96/97; MCI: 31/NA/19; Pitt & Got.:MMSE	MCI vs. HC	Got & Kar:unreported, Pitt:avail.[22]	Got & Kar:SWE., Pitt: EN
(Lopez et al., 2014)	**AZTIAHORE**, *n*=40 (subset of AZTIAHO[13], SS video); AD class gender balanced, AD and HC not age balanced	HC/AD_{ES}/AD_{IS}/ HC/AD_{AS}:20/4/10/6; MMSE:NA	AD vs. HC (multiclass)	unreported	EN, FR, ES, CAT, BSQ, CHI, ARB, PRT

Continued on next page

[12] https://talkbank.org/dementia/

[13] 50 HC and 20 AD subjects. AD has 3 severity stages: ES (early), IS (intermediate) and AS (advanced);

Table 1 – *Continued from previous page*

Study	Database, task & balance	Data annotation	Task type (cls./Reg.)	Data availability	Language
(Lopez et al., 2013)	**AZTITXIKI:** *n*=10 (subset of AZTIAHORE, SS video); not balanced	$HC/AD_{ES}/AD_{IS}/AD_{AS}$: 5/1/2/2; MMSE:NA	HC vs. AD (ES, IS, AS) (multiclass)	unreported	EN, FR, ES, CAT, BSQ, CHI, ARB, PRT
(Martinez et al., 2017)	**ALZUMERIC** CVF (AN)[14]:*n*=225 PD:*n*=18 (rec.), SS (AZTIAHORE):*n*=40 ,video all; balanced: CVF and PD not gender balanced; SS gender balanced across classes; age balance not reported	CVF :HC/MCI:187/38, PD:HC/AD:12/6, SS:HC/AD: 20/20; MMSE:NA	PD and SS : AD vs HC AN : MCI vs HC	unreported	EN, FR, ES, CAT, BSQ, CHI, ARB, PRT
(Li et al., 2019)	**Pitt**:*n*=551 PD(CTP), **Lu**:*n*=49 , PD (CTP), SVF, PN ; tr., **OS** *n*=50k narrations (parallel corpus); age, gender and education not balanced (Pitt, Lu)	Pitt:HC/AD:241/310 tr. & rec. Lu:HC/AD(score exhibit various degrees of dementia):NA/49 , OS:NA; Pitt:MMSE, Lu:NA	AD vs. HC	Pitt: avail.[22], Lu:avail., OS:avail.	Pitt EN, OS: CHI/EN, Lu: Tai. Mand.
(Guo et al., 2020)	**Pitt**:*n*=498, PD(CTP) tr. & rec., **other**:*n*=208 PD (CTP), **OS** *n*=9.9m lines dialogs (parallel corpus), tr. ; age and gender not balanced (Pitt, other)	Pitt:HC/AD:242/256, other:HC/AD:104/104, OS:HC/AD:NA/NA; Pitt, other:MMSE, OS:NA	AD vs. HC	OS:avail., other:unreported	Pitt EN, OS: CHI/EN, other: Mand. CHI
(Gosztolya et al., 2021)	**corpus-1**:(*n*=33) **corpus-2**: (*n*=33), (NR) both corpus; not balanced for age, gender, education	corpus-1:HC/MCI:19/14, corpus-2:HC/MCI:20/13; corpus-1&2: MMSE, CDT, GDS score	MCI vs. HC	corpus-1&2:unreported	corpus-1:EN, corpus-2:HU
(Campbell et al., 2021)	**AcceXible**: *n*=154 ,CVF (AN, LF), **ADReSS**: *n*=156 (PD (CTP)) rec. & tr.; AcceXible: not balanced for age, ADReSS: age, gender balanced	AcceXible:HC/AD:81/73, ADReSS (train):HC/AD:54/54, ADReSS (test):HC/AD:24/24; AcceXible: MMSE, CDT, GDS score, ADReSS: MMSE	AD vs. HC	AcceXible:unreported, ADReSS:avail.[22]	AcceXible:ES, ADReSS:EN
(Pérez-Toro et al., 2023)	**Pitt**: *n*=186, PD (CTP), **corpus-3**:*n*=56 (PD (CTP)) rec. & tr., **corpus-4**: *n*=205 (PD), rec. & tr. ; corpus-3&4: not balanced for age and gender, Pitt: age, gender balanced	Pitt:HC/AD:93/93, corpus-3:HC/AD:27/29, corpus-4:HC/AD:58/147; Pitt: MMSE, corpus-1&2:unreported	AD vs. HC	corpus-1&2:unreported	Pitt:EN, corpus-1:ES, corpus-2:DE
(Pérez-Toro et al., 2022)	**corpus-5:**: *n*= 39,**Pitt:** *n*=186 ;PD (CTP), rec. & tr.; Pitt & corpus-5: age, gender, education matched	corpus-5:HC/AD:18/21, Pitt:HC/AD:93/93; corpus-1: MMSE (for AD class), Pitt: MMSE	AD vs. HC	corpus-5: unreported	corpus-5:ES-CL
(Lindsay et al., 2024)	**Private study**, *n*=138 (PD: 1-min free speech, phone), age balance not fully detailed; MMSE mostly >25	No HC/MCI/AD breakdown; MMSE and SB-C collected	Correlation (MMSE, SB-C), not classification	private	ES, CAT, DE, NL

Continued on next page

[14]Gipuzkoa-Alzheimer Project: https://www.cita-alzheimer.org/es/investigacion/proyectos

Table 1 – *Continued from previous page*

Study	Database, task & balance	Data annotation	Task type (cls./Reg.)	Data availability	Language
(Hernández-Domínguez et al., 2016)	**Carolinas Conversations (Latin America cohort)**, *n*=112 (natural dialogue, video), longitudinal; age/gender balance not specified	AD: ≥57, MCI/Other: 25, HC: 30; clinical diagnosis	Data collection only (no modeling)	Not public yet (planned)	EN, ES (USA, MX, EC)
(Schäfer et al., 2022)	**DeepSpA:** *n*=121, **H70:** *n*=404; RAVLT, SVF via mobile app; speech-based score pipeline	DeepSpA: 52 MCI/69 SCI; H70: 48 MCI/356 HC; MMSE/CDR	MCI vs. HC	avail. upon request	NL, SWE
(Lopez et al., 2018)	**AZTIAHO** , *n* = 283, AN, PD, SS ; multiple environments; age/gender balance varies by task	AN: 38 MCI/187 HC; PD: 6 AD/12 HC; SS: 20 AD/20 HC; MMSE and others	MCI or AD vs. HC (task-dependent)	private	ES, CAT, BSQ, EN, FR, ARB, PRT, CHI
(Kálmán et al., 2022)	**Speech-GAP Test®** English/Hungarian monologue task, *n*=66; phone-based, not fully balanced	EN: 19 HC/14 MCI; HU: 20 HC/13 MCI; MMSE-based criteria	MCI vs. HC	private	EN, HU
(Ablimit et al., 2022)	**ILSE :** *n*=108 (91 participants, BI, rec. & tr.), **ADReSS:** *n*= 156; (CTP), ILSE not balanced	ILSE:HC/AD/AACD:108/16/21, ADReSS train:HC/AD:54/54, test:HC/AD:24/24; ILSE: MMSE (main dataset)	AD vs. HC	ILSE: avail. on request	GER, EN
(Lindsay et al., 2021a)	**Corpus-6**: *n*=132; SVF (AN); balanced for age and education in each language	HC/AD: 27/27(FR), 23/23(GER), 16/16 (DE)	MCI vs. HC	private; clinical data from 3 memory clinics	FR, GER, DE (original), EN (translated)
(Fraser et al., 2019a)	**Pitt:** *n*=550, **corpus-7**: *n*=58 , PD (CTP), rec. & tr.	Pitt:HC/AD:241/309, corpus-7:HC/AD:25/33; Pitt & corpus-7: MMSE	AD vs. HC	Pitt: avail.[15]; corpus-7: private	Pitt:EN, corpus-7:FR
(Beckett, 2004)	Case-study of bilingualism in AD detection; spontaneous conversation with a researcher about past life on predefined topics (e.g. Home, School, Names, Housekeeping, Christmas, Gifts, Romance, Children, Work)	2 AD participants	qualitative analysis	transcripts avail. in the paper	EN, AFR
(Lindsay et al., 2021b)	**ADReSS (subset):** *n*=106, age, gender balanced **EIT-Digital**: *n*=47, educ. balanced, not age ; PD (CTP); rec. & tr.	ADReSS'20:HC/AD:52/54, EIT-Digital:HC/AD:25/22; MMSE avail.	AD vs. HC	ADReSS:avail [16], EIT-Digital: avail.[17]	ADReSS:EN, EIT-Digital: FR

Continued on next page

[15] https://talkbank.org/dementia/
[16] https://talkbank.org/dementia/
[17] upon request to: alexandra.konig@inria.fr

Table 1 – *Continued from previous page*

Study	Database, task & balance	Data annotation	Task type (cls./Reg.)	Data availability	Language
(Azadmaleki et al., 2025)	**PREPARE:** (*n*=2058(train/ test:1646/412), BI, NR, PD, SVF, VAI, RT (rec.), age, gender, educ. not balanced);(**Chou corpus:** *n*=87 (train/val/test:51/17/19)); rec.	PREPARE:AD/MCI/HC: 1140/268/650, Chou corpus:MCI/HC:47/40; PREPARE: MMSE (partially avail. in original dataset), Chou corpus: no MMSE	PREPARE: AD vs. MCI vs. HC, Chou corpus: MCI vs. HC	PREPARE & Chou corpus: avail.[18]	PREPARE: EN, Mnd. CHI, ES, Chou corpus: Mnd. CHI
(Melistas et al., 2023)	**ADReSS-M**; **Ivanova** (*n*=271, RT, not balanced); rec.	Ivanova: AD/HC:74/197, MMSE: not avail.	AD vs. HC	both corpus: avail.[19]	ADReSS-M: EN, GRK, Ivanova: ES
(Jia et al., 2025b)	**ADReSS-M**(*n*=8 PD (LWC)); **ADReSSo'21** (*n*=237 (train/test:166/71), PD (CTP), balanced);**NCMMSC** (*n*=124, PD);**Ivanova** (*n*=360, RT); balance not reported); rec.	ADReSSo'21: AD/HC:train(87/79), test(35/36); NCMMSC:AD/MCI/HC: 26/54/44; Ivanova:AD/MCI/HC: 74/91/197; ADReSS-M: AD/HC:4/4; ADReSSo'21: MMSE; NCMMSC: MMSE not avail.	AD vs. HC MCI in Ivanova corpus reclassified as AD)	ADReSSo'21, ADReSS-M, Ivanova: avail[20], NCMMSC: avail.[21]	ADReSS-M: GRK, ADReSSo'21: EN, NCMMSC: CHI, Ivanova:ES
(Tsai et al., 2021)	**Pitt** PD (CTP), not balanced;**NTUHV** (PD, LM) ;rec. and tr., balanced	AD/HC:Pitt(257/242), NTUHV(40/40) MCI/HC: Pitt(43/43), NTUHV(30/30); MMSE avail.	AD vs. HC & MCI vs. HC	NTUHV: private	EN, CHI
(Ceyhan et al., 2024)	**ADReSS:** (n=153, age and gender balanced); **MUDC:** *n*=60; PD (CTP); rec., tr.,	MUDC:AD/HC:37/23; ADReSS: AD/HC: train:53/52 test:24/24; MMSE avail.	AD VS. HC	MUDC: on request, ADReSS:avail.[22]	ADReSS: EN, MUDC: TR

[18] https://talkbank.org/dementia/
[19] https://talkbank.org/dementia/
[20] https://talkbank.org/dementia/
[21] https://web.ee.tsinghua.edu.cn/satlab/en/gxsj/7552/content/1011.htm?utm_source=chatgpt.com
[22] https://talkbank.org/dementia/