

Original Paper

Early Detection of Alzheimer's Disease and Related Dementias From Spontaneous Speech Using Foundation Speech and Language Models: Comparative Evaluation

Jingyu Li^{1*}, MS; Lingchao Mao^{1*}, PhD; Xi Mao^{2*}, PhD; Hairong Wang³, PhD; Zhendong Wang⁴, PhD; Xuelei Sherry Ni⁵, PhD

¹H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States

²Department of Economics, The University of Texas Rio Grande Valley, Edinburg, TX, United States

³Operations Research and Industrial Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, United States

⁴Shandong Provincial Key Medical and Health Laboratory of Digital Psychiatry, Shandong Mental Health Center, Jinan, China

⁵School of Data Science and Analytics, Kennesaw State University, Marietta, GA, United States

*these authors contributed equally

Corresponding Author:

Xuelei Sherry Ni, PhD

School of Data Science and Analytics

Kennesaw State University

680 Arntson Drive

Room 3411C

Marietta, GA, 30060

United States

Phone: 1 4705782551

Email: sni@kennesaw.edu

Abstract

Background: Alzheimer's disease and related dementias (ADRD) are progressive neurodegenerative conditions where early detection is critical for timely intervention and care planning. However, current diagnostic methods are often inaccessible, costly, and delayed, especially for underserved populations. There is a growing need for scalable, noninvasive tools that can support timely diagnosis. Spontaneous speech contains rich acoustic and linguistic markers that can serve as noninvasive behavioral markers for cognitive decline. Foundation models, pretrained on large-scale audio or text data, generate high-dimensional embeddings that encode rich contextual and acoustic information.

Objective: This study benchmarks open-source foundation language and speech models to evaluate their effectiveness in detecting ADRD from spontaneous speech as a potential solution for early, noninvasive, and scalable ADRD detection.

Methods: In this study, we used the Pioneering Research for Early Prediction of Alzheimer's and Related Dementias EUREKA (PREPARE) Challenge dataset, which consists of audio recordings from over 1600 participants with 3 distinct categories of cognitive decline: healthy control (HC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). We further excluded samples that are non-English, nonspontaneous speech, or of poor quality. Our final samples included 703 (59.13%) HC, 81 (6.81%) MCI, and 405 (34.06%) AD cases. We systematically benchmarked 18 open-source foundation speech and language models to classify cognitive status into 3 categories (HC, MCI, or AD). Post hoc interpretability analysis was performed for the best-performing model using Shapley additive explanations linking high-dimensional embeddings with explainable acoustic and linguistic markers.

Results: Whisper-medium model achieved the highest performance among speech models at 0.731 accuracy and 0.802 area under the curve, while Bidirectional Encoder Representations from Transformers with pause annotation achieved the top accuracy of 0.662 and 0.744 area under the curve among language models. Overall, ADRD detection based on state-of-the-art automatic speech recognition model-generated audio-embeddings outperformed other models, and the inclusion of nonsemantic information, such as pause patterns, consistently improved the classification performance of text-embedding-based models.

Conclusions: Our work presents a comprehensive comparative evaluation of state-of-the-art speech and language models for AD and MCI detection on a large, clinically relevant dataset. Embeddings derived from acoustic models, which capture both

semantic and acoustic information, show promising performance and highlight the potential for developing a more scalable, noninvasive, and cost-effective early detection tool for ADRD.

(*JMIR Form Res* 2026;10:e79411) doi: [10.2196/79411](https://doi.org/10.2196/79411)

KEYWORDS

Alzheimer's disease; clinical screening; comparative evaluation; dementia detection; foundation speech models; language models; speech biomarkers; spontaneous speech analysis

Introduction

Alzheimer Disease and Early Detection

Alzheimer disease and related dementias (ADRD) are progressive brain disorders that gradually impair memory, thinking, and daily functioning. ADRD typically progresses through 3 stages: healthy control (HC) or preclinical, mild cognitive impairment (MCI), and formal Alzheimer disease (AD) or dementia [1]. MCI, while not always a precursor to AD, is an important early indicator and target for intervention [2].

Because ADRD progresses slowly and irreversibly, and new US Food and Drug Administration–approved drugs like lecanemab and donanemab are effective only in early stages, timely detection is essential to delay symptoms and expand treatment options [3,4]. Although advances in brain imaging, blood biomarkers, and artificial intelligence have improved early detection tools, global health care systems remain underequipped. For instance, limited access to positron emission tomography/magnetic resonance imaging and dementia specialists leads to projected diagnostic delays of up to 10 years in the United Kingdom [5]. A 6-country study found that less than half of individuals who reported memory problems received a formal diagnosis, and few underwent advanced testing—even when seeing a specialist [6]. In the United States, early detection for ADRD is also inadequate and often delayed. Only 8% of Medicare enrollees receive a diagnosis of MCI, and more than 99% of primary care providers and clinics are identifying fewer cases than expected. Even specialists such as geriatricians missed over 80% of early cases [7]. Estimates suggest that 14% of MCI cases progress to dementia annually [8], and over 7 million Medicare beneficiaries may miss treatment opportunities [9]. Missed and late diagnoses often disproportionately affect underrepresented populations, such as racial and ethnic minorities, rural communities, and those with lower socioeconomic status. The worsening health disparities in Alzheimer care and outcomes highlight the critical need for accessible and scalable diagnostic solutions to support early detection of ADRD [10,11].

Language Analysis for ADRD Detection

To close the gap between need and access, health care systems are exploring tools that support earlier, faster, and more affordable ADRD detection. One promising direction involves analyzing paralinguistic and linguistic patterns of speech, since language is among the earliest and most visible domains affected by ADRD [12,13].

Common language symptoms in patients with AD include naming difficulties, repetition, vague word use, inappropriate

pronouns, etc [14,15]. These often lead to fluent yet uninformative speech, with reduced coherence and incomplete sentences [16]. Temporal speech changes such as increased hesitations and slower tempo can also occur, reflecting cognitive challenges in planning and organizing language [17,18].

Accordingly, linguistic and paralinguistic features have been widely studied as noninvasive behavioral markers of ADRD [19-21]. Markers such as semantic fluency, speech recognition, and acoustic features have proven effective in distinguishing normal aging from early MCI [21-23]. Integrating these tools into telehealth workflows, for example, through virtual memory testing and teleneuropsychology, offers a promising approach for remote screening and patient triaging. Such services have been well-received by patients and have helped reduce diagnostic delays [24,25].

Modern Foundation Speech and Language Models for ADRD Detection

Traditional feature engineering approaches for ADRD detection have relied on manually engineered features (eg, lexical diversity, syntactic complexity, and coherence from transcripts) and acoustic features (eg, pause duration, reaction time, loudness, and speech rate) [19-21,23,26,27]. These features are then fed into machine learning classifiers. However, these pipelines require manual feature engineering, and their performance can depend on controlled environments where speech samples are collected. Robin et al [22] emphasized the importance of rigorously evaluating recording quality and data processing accuracy. Recent advances in deep learning, especially in natural language processing, have led to highly capable foundation speech and language models. These models are pretrained on large-scale datasets. These models learn to capture rich contextual information and produce high-dimensional representations of raw speech or texts, known as “embeddings,” thus simplifying the labor-intensive and human-dependent preprocessing step. These embeddings capture semantic, syntactic, and prosodic signals and have shown promise in ADRD detection [28-30]. Embeddings generated by large language and speech models also demonstrate greater flexibility, robustness to varied inputs, and even multilingual capabilities. This shift supports more generalizable and scalable detection in telehealth or low-resource settings. Shakeri and Farmanbar [13] provide a detailed review of the natural language processing in AD research in their 2025 paper.

Two common approaches exist for extracting these embeddings for ADRD detection from deep learning models. One approach is to transcribe speech to text and then use a language model to extract text embeddings [28,29]. However, the transcript-based approach captures only semantic information, while neglecting

important nonsemantic information such as paralinguistic features and pause patterns [29,30]. To address these gaps, some researchers have manually annotated speech pauses in transcripts to preserve nonsemantic information contained in patients' speech data [15,29,30]. This step has been shown to improve prediction accuracy, but it introduces additional complexity, lacks flexibility, and captures limited features of speech. Therefore, current research has been limited to binary ADRD detection using small or curated datasets.

An increasingly promising alternative is to extract embeddings directly from raw audio using pretrained automatic speech recognition (ASR) models trained for both recognition and representation tasks. These models preserve both semantic and paralinguistic cues without requiring transcription. Among them, OpenAI's Whisper has demonstrated state-of-the-art performance across varied languages and acoustic conditions [16]. However, despite their potential, these models have not yet been rigorously benchmarked for ADRD detection in large and real-world speech datasets.

Objective

In this study, we conduct a comprehensive comparative evaluation of modern foundation speech and language models for ADRD detection. Rather than developing advanced models to achieve higher prediction accuracy, the goal of this study is to establish a comprehensive benchmark by systematically evaluating a range of open-source foundation models and comparing their baseline performance in ADRD detection on a large clinical dataset. These results offer an initial reference point for applying these toolkits to ADRD detection.

Specifically, we compare ADRD detection based on text and audio embeddings generated by 18 state-of-the-art open-source language and speech models. We evaluate whether the high-dimensional embeddings generated by these models can serve as effective behavioral markers for ADRD detection. To the best of our knowledge, neither the Bidirectional Encoder Representations from Transformers (BERT) family models nor the speech models evaluated in this study have been systematically and comprehensively assessed for ADRD detection on a large-scale dataset. Our experiments use the Pioneering Research for Early Prediction of Alzheimer's and Related Dementias EUREKA (PREPARE) Challenge dataset—a large, diverse, and publicly available speech corpus curated from DementiaBank containing over 1600 samples to ensure relevance to a diverse population. Notably, our findings identify the medium-sized Whisper model as the top performer in ADRD detection from speech and demonstrate its advanced language understanding to capture linguistic and acoustic markers associated with ADRD. We further discuss an acoustic-based ADRD detection pipeline with strong potential to advance early detection and intervention compared to the traditional diagnosis timeline (Figure S1.1 in [Multimedia Appendix 1](#) [31-34]). This study highlights the promise of foundation speech models for early ADRD detection and offers practical guidance for the development of scalable and noninvasive ADRD screening tools.

Methods

The PREPARE Challenge Audio Data

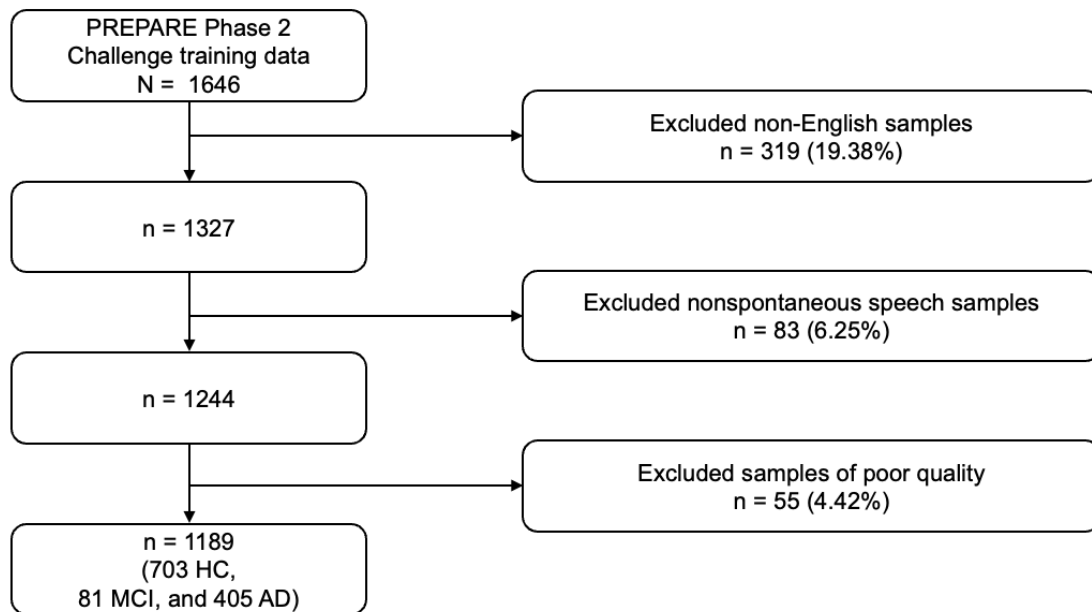
The audio data used in this study are available as part of the PREPARE Phase 2 Challenge-Acoustic Track [35]. This dataset was curated from the DementiaBank [36], a well-established clinical research repository containing multilingual audio recordings of participants performing a wide range of spontaneous speech tasks, including picture description, verbal fluency, and sentence construction. Participants contributing to DementiaBank are recruited through affiliated memory clinics and research sites and, in some cases, community outreach/promotional materials. All contributed datasets involve institutional review board's (IRB) approval and informed consent, as specified in the TalkBank/DementiaBank website [37]. Diagnostic labels (HC, MCI, or AD) are assigned using established clinical criteria. The PREPARE consortium subsequently curated these data, ensuring deidentification and creating consistent metadata before public release.

The original curated dataset consists of audio recordings from 2086 participants, divided into 1646 training samples and 412 testing samples during the PREPARE Phase 2 Challenge. In this study, we used only the provided training dataset, as the cognitive status labels for the testing data were not released.

The dataset comprised speech samples from multiple languages and various language tasks, including sentence reading and spontaneous speech tasks. In the spontaneous speech task, participants were asked to describe the Cookies Theft picture for English speakers, which is a popular language impairment assessment tool originally used in the Boston Diagnostic Aphasia Examination protocol [17]. Since speech and language models primarily analyze speech content, we focused our analysis exclusively on spontaneous speech samples. To identify spontaneous speech samples, we applied density-based spatial clustering of applications with noise on BERT embeddings of the speech transcriptions, followed by human evaluation to classify clusters to language tasks [18,38]. See more details in Figure S2.1 in [Multimedia Appendix 1](#). The transcription and language of recordings were detected using the Whisper-small model [16].

Given that the majority of the recordings were in English, we excluded 319 non-English samples. Additionally, we removed 55 recordings of poor quality, defined as those with participant speech lasting less than 3 seconds or excessive background noise, based on human review. Demographic information of each participant, including age and gender, is also available. A detailed, step-by-step description of the data inclusion process is shown in [Figure 1](#). The final dataset comprised recordings from 1189 participants (≤ 30 seconds each), including 703 classified as HC, 81 as MCI, and 405 as AD. The final sample ($n=1189$) included 680 female and 509 male individuals, with a mean age of 75.178 (SD 8.430) years.

Figure 1. Data inclusion criteria. Speech samples were excluded if they were non-English, nonspontaneous, or of poor audio quality. The final dataset included 1189 samples: 703 from healthy controls (HC), 405 from individuals with Alzheimer disease (AD), and 81 from individuals with mild cognitive impairment (MCI).



Modeling Pipelines

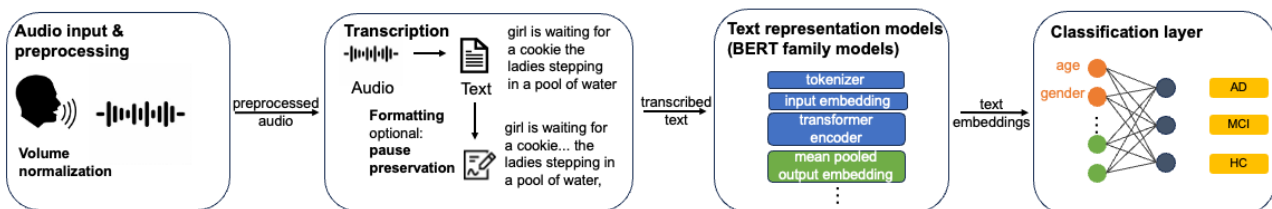
Overview

We explore two deep learning-based pipelines to extract embeddings from foundation models and classify cognitive status of the participants from audio recordings: (1) a *text-based pipeline*, where a language model analyzes the transcribed content of the speech to identify linguistic patterns associated with cognitive decline and (2) an *audio-based pipeline*, where a speech model transforms audio signals into embeddings that capture both the linguistic and acoustic patterns characteristic of ADRD. Figure 2 shows the detailed workflow for both pipelines.

To ensure the comprehensiveness of our benchmarking study, we compared the performance of the 2 deep learning-based pipelines with that of the traditional feature engineering pipeline, where classification models were trained on widely used preextracted paralinguistic and linguistic features reported in the literature including Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [31], Computational Paralinguistics Challenge [32], Mel-Frequency Cepstral Coefficients [39], and the Linguistic Inquiry and Word Count [33]. Detailed description of these 4 groups of features can be found in Table 1 and Multimedia Appendix 2. More details about this pipeline are provided in section S3 of the Multimedia Appendix 1.

Figure 2. Overview of the modeling pipeline for Alzheimer disease and related dementias detection from spontaneous speech samples. (A) Text-based Modeling Pipeline: the system takes and preprocesses audio input, transcribes audio into text (optional: pause is annotated), extracts sentence-level embeddings from transcribed text using a text representation model, and feeds embeddings into a neural network to determine patient’s cognitive status into 3 categories: healthy control (HC), mild cognitive impairment (MCI), or Alzheimer disease (AD). (B) Audio-based Modeling Pipeline: the system takes and preprocesses audio input, extracts sentence-level embeddings from audio data directly using an audio representation model, and feeds embeddings into a neural network to determine the patient’s cognitive status. The extracted embeddings are concatenated with demographic information before feeding them into the classification layer. BERT: Bidirectional Encoder Representations from Transformers.

(A) Text-based Modeling Pipeline



(B) Audio-based Modeling Pipeline

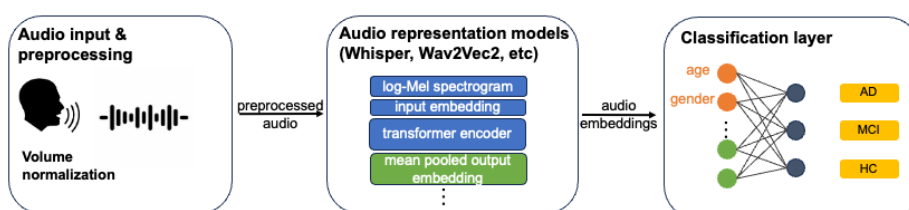


Table 1. Traditional acoustic and linguistic features.

Feature set	Features, n	Acoustic and linguistic variables
eGeMAPS ^a [31]	89	eGeMAPS features provide a compact set of acoustically and physiologically motivated speech descriptors, including pitch, loudness/intensity, spectral shape, formant frequencies, voice quality measures (jitter, shimmer, and HNR ^b), harmonic amplitude ratios, formants, and temporal dynamics.
ComParE ^c [32]	66	ComParE features provide a comprehensive representation of speech by capturing pitch and phonation, voice perturbations (jitter and shimmer), periodicity and noisiness (harmonics-to-noise ratio and harmonicity), loudness and amplitude, temporal dynamics (zero-crossing rate), and spectral energy distribution.
LIWC ^d [33]	108	LIWC features capture a wide range of linguistic and psychological dimensions from text. It includes scores quantifying analytic thinking, social status or confidence, authenticity, emotional tone, as well as usage of terms related to psychological processes, emotion, cultural and personal dimensions, health, and personal needs or drives.
MFCC ^e [34]	20	MFCCs capture the perceptual spectral envelope of speech by summarizing the distribution of energy across frequency on the Mel scale. Lower-order MFCCs reflect coarse spectral structure (eg, spectral tilt and formant distribution), while higher-order coefficients capture fine spectral details and rapid variations.

^aeGeMAPS: Geneva Minimalistic Acoustic Parameter Set.

^bHNR: harmonics-to-noise ratio.

^cComParE: Computational Paralinguistics Challenge.

^dLIWC: Linguistic Inquiry and Word Count.

^eMFCC: Mel-Frequency Cepstral Coefficient.

Our objective is to systematically evaluate a broad set of emerging open-source speech and language models and establish baseline performance benchmarks of these models in ADRD detection. Our central research question is whether deep learning-based pipelines, particularly the audio-based pipeline with state-of-the-art models, outperform the traditional feature engineering pipeline that has been commonly used in previous literature. In this context, we define a “strong” result as one in which a model demonstrates superior performance relative to other benchmarked models and achieves comparable or higher levels of discrimination that are considered clinically meaningful in the literature. Our benchmarks can serve as an initial reference point for future methodological improvements and a foundation for developing practical tools to be used in downstream clinical applications [13,28,29,40-42].

Text-Based Pipeline

Overview

For the text-based pipeline, we first transcribed the preprocessed audio recordings into text using an ASR system, followed by the extraction of embeddings using pretrained language models for downstream classification. This approach emphasizes the linguistic content of speech for dementia diagnosis and can benefit from the abundance of textual data available on the internet, public datasets, and open-domain corpora. To incorporate acoustic characteristics into text embeddings, we further encoded pauses manually within the transcribed text to generate pause-aware embeddings.

Audio Preprocessing

Each raw audio clip (48 kHz) was resampled to 16 kHz, as higher frequencies in the audio recordings typically do not contain relevant speech. Audio amplitudes were normalized, and the beginning and end silence were trimmed. Because recordings were collected in various acoustic environments, we used DeepFilterNet to suppress background noise and enhance audio clarity [43].

Audio Transcription

We used OpenAI’s Whisper model to transcribe audio recordings into text [16]. Preprocessed audio files were loaded and transformed into log-Mel spectrograms, which were fed into the Whisper model for transcription. Encoding pauses in the transcripts has shown superior predictive power in identifying AD status from audio data [29]. To capture speech disfluencies relevant to cognitive decline, we included an optional step of annotating word pause and sentence pause in the transcripts. Specifically, word pauses were detected by measuring the time gap between consecutive words within a segment, and a pause marker was inserted if the duration exceeded a predefined threshold of 0.05 seconds. Sentence pauses were identified by calculating the silence duration between adjacent transcription segments, with pauses above the threshold labeled accordingly. After transcription, the text underwent further preprocessing, including formatting and tokenization, to ensure compatibility with text representation models. Specifically, we removed punctuation, adjusted text casing based on model requirements, and recoded word and sentence pauses for transcripts with the optional pause annotation. Similar to the approach introduced in Yuan et al [29], we categorized pauses into 3 bins: short (under 0.5 seconds); medium (0.5-2 seconds); and long (over 2 seconds), which were represented using punctuation marks—commas (,), periods (.), and ellipses (...), respectively to represent different levels of disfluencies in speech. The processed text underwent proper tokenization before being served as input to text representation models.

Text Embedding Models

BERT [38] and BERT-like models have been highly effective for a wide range of natural language processing tasks, including AD classification. BERT is a transformer-based language model designed to learn contextualized word representations by leveraging bidirectional self-attention to capture associations among words. Each attention head processes elements within

a sequence and generates a new sequence by computing a weighted sum of the transformed input representations. The base model consists of 12 layers, 12 attention heads, and 110 M parameters, while the larger version has 24 layers, 16 attention heads, and 340 M parameters. We also included 2 BERT models adapted for the medical domain in our benchmarking study. BioBERT [44] is a domain-specific adaptation of BERT, pretrained on large-scale biomedical corpora, including PubMed abstracts and full-text articles. We also evaluated BioClinicalBERT, a model further pretrained on clinical notes [45]. While BioBERT and BioClinicalBERT retain the transformer architecture and bidirectional pretraining approach of the base version of BERT, their exposure to domain-specific corpora allows the model to generate more relevant embeddings for medical and clinical applications. In our implementation, mean-pooled sentence-level embeddings were used for downstream classification.

Audio-Based Pipeline

Overview

For the audio-based pipeline, we directly extracted speech embeddings from the recordings using foundation speech models designed from speech processing. This approach benefits from capturing phonetic, prosodic, and acoustic features relevant to speech impairments associated with dementia diagnosis.

Audio Preprocessing

Each raw audio clip (48 kHz) was resampled to 16 kHz, matching the training input of most audio models, and audio amplitudes were normalized. Clips were then padded or truncated to an equal length of 30-second segments. The preprocessed audio inputs were passed into model encoders to obtain mean-pooled embeddings for each clip. For Whisper models, we converted each resampled clip into 80-channel (Whisper-medium or smaller) or 128-channel (Whisper-large or larger) log-Mel spectrogram representation on 25-millisecond windows with a stride of 10 milliseconds, following the original preprocessing of OpenAI [16].

Audio Embedding Models

We used a range of open-source ASR models and self-supervised speech models to learn meaningful speech representations from audio data. Whisper [16] is a state-of-the-art open-source transformer-based ASR model, trained on a large-scale multilingual dataset. It consists of an encoder-decoder architecture where the encoder processes raw audio into latent representations of audio features, and the decoder generates transcriptions. The family of Whisper models varies in size, with the larger versions offering improved transcription accuracy at the cost of higher computational demand. In our study, we demonstrated cases of using both encoder and decoder outputs, where the encoder output served as embedded features, and the decoder output served as audio transcripts for further processing.

We also compared a few prior open-source deep learning models for speech representation. Wav2Vec2 [46] learns speech representations through a 2-stage training process: a multilayer convolutional neural network feature extractor encodes raw speech into latent representations, followed by a transformer

encoder that refines these representations via masked speech prediction. HuBERT [47] uses a masked prediction approach where it learns to predict cluster-based speech units from unlabeled audio. This self-supervised learning method allows the model to capture speech structures without requiring transcriptions. Unispeech [48] improves upon prior models such as Wav2Vec2 and HuBERT by integrating both unlabeled speech pretraining and supervised fine-tuning within a single model. WavLM [49] extends the Wav2Vec2 framework by incorporating denoising and masked speech modeling objectives in pretraining. Data2Vec further generalizes self-supervised learning by predicting latent representations instead of specific speech tokens, thus creating a more unified architecture across multiple modalities, including speech [50].

Classifier and Covariates

Both pipelines were evaluated and compared on their ability to accurately classify AD. We mean-pooled sentence-level embeddings to form clip-level embeddings. We incorporated the patient's demographic information, including age and sex, by concatenating these 2 scalars to the final encoder embeddings of all the models. These modified embeddings then served as input features to a single-layer feedforward neural network, which was trained to predict patients' cognitive status: HC, MCI, or AD.

Experiment Setup

Model Architecture and Hyperparameters

For the text-based pipeline, we used Whisper-small and Whisper-medium for transcription. We did not consider larger Whisper models, as transcripts generated by Whisper-medium did not yield performance improvements over those by Whisper-small. We used pretrained BERT (uncased), BioBERT, and BioClinicalBERT to extract embeddings from transcribed text. Each transcript was tokenized using BERT's default tokenizer. The text was padded to a max length of 512 tokens or truncated if too long. The final embedded features extracted from these 3 BERT family of models are of shape (768).

For the audio-based pipeline, we evaluated 15 deep learning models designed for speech processing: Whisper (tiny, base, small, medium, and large) [16], WavLM (base, base-plus, and large) [49], Wav2Vec2 (base and large) [46], HuBERT (base and large) [47], Unispeech (large) [48], and Data2VecAudio (base and large) [50]. The extracted embeddings varied in dimensionality, with Whisper-tiny embeddings of shape (512); Whisper-base, small, WavLM, Wav2Vec2, and HuBERT embeddings of shape (768); Whisper-medium, Unispeech, and Data2VecAudio embeddings of shape (1024); and Whisper-large embeddings of shape (1280). The models were pretrained on different data corpora, from the 960-hour LibriSpeech dataset [51] used by (Wav2Vec2, HuBERT, and Data2Vec), to more diverse multisource pretraining corpora used by WavLM, and larger datasets such as the 680,000-hour combination of public and proprietary speech data used by Whisper.

For both pipelines, we used a classification layer consisting of a single-layer feedforward neural network of hidden layer size

80 for tiny to medium speech models and of hidden layer size 128 otherwise.

Cross-entropy loss was used for the multiclass classification. During training, the embedding extractors were kept frozen, and only the classification layer was trained. Models were trained for up to 100 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 0.0005 and ReduceLROnPlateau scheduler. To prevent overfitting, we terminated the training if the validation loss failed to improve for 5 consecutive epochs. All pretrained models were sourced from the HuggingFace Transformers library and trained using PyTorch on an NVIDIA V100 GPU. The data preprocessing steps were performed using the pandas, numpy, audiofile, and librosa packages in Python. Evaluation metrics were computed using the sklearn and torchmetrics packages in Python.

Evaluation

The dataset was randomly divided into 80% training and 20% testing based on stratified sampling on the patients' labels of cognitive status to ensure balanced class proportions. Due to class imbalance, we also oversampled minority classes (AD and MCI in our case) in the training set to ensure balanced class distribution before training. A total of 20% (2/10) of the training set was used for validation. The train-test split was repeated 5 times with different random seeds. We selected metrics commonly used in classification tasks. Accuracy provides an intuitive measure of overall classification performance and is well-suited in multiclass settings. Area under the curve (AUC), unlike other metrics, offers a threshold-independent evaluation of the model's performance, which is important for imbalanced clinical datasets. In this 3-class classification setting, AUC was computed using a one-versus-rest macroaverage approach for each class. For each metric, we reported the mean and SD on the test set across 5 repetitions.

Interpretability Analysis

As deep learning models often function as black boxes, we conducted a post hoc analysis to enhance the interpretability of our results. We first applied Shapley additive explanations (SHAP) [52] to identify embedding dimensions that were most contributing to predicting each of HC, MCI, and AD classes. To interpret the high-dimensional embeddings, we curated a complementary feature list comprising explainable acoustic and linguistic features (eGeMAPS and Linguistic Inquiry and Word Count), widely adopted in prior studies (Table 1). We then

examined the extent to which each top contributing embedding dimension can be explained by these features by mapping these interpretable acoustic features to each embedding. Specifically, we fitted multivariate linear regression models with embeddings as the dependent variable and the curated feature set as independent variables. For each class, we identified features with a statistically significant relationship with each top contributing embedding dimension to interpret its acoustic and linguistic properties. The full list of acoustic and linguistic features and their definition can be found in [Multimedia Appendix 2](#).

Ethical Considerations

This study used open-access, deidentified participant recordings curated from the DementiaBank repository. Original data collection was conducted independently by the contributing institutions under institutional ethical approval and informed consent [37]. Because this research involved only secondary analysis of publicly available, deidentified data, it was determined not to involve human subjects research under the IRB policies of collaborating institutions and in accordance with US federal regulations (45 CFR 46.102). Accordingly, IRB review and approval were not required. All analyses were performed using deidentified data, and no attempt was made to reidentify participants.

Results

Text-Based Pipeline

Table 2 shows the results of BERT, BioBERT, and BioClinicalBERT using transcriptions generated by Whisper-small. We reported average prediction accuracy and AUC across 5 repetitions. Among all configurations, BERT-base with pause annotations on Whisper-small transcriptions achieved the highest accuracy of 0.662 (SD 0.013), while the same model using Whisper-medium transcription achieved the highest AUC of 0.744 (SD 0.014). This suggests that for the same type of model and configuration, incorporating pause annotations consistently led to improved performance in both accuracy and AUC compared to their counterparts without pause annotations. Results of the same BERT family models using transcriptions generated by Whisper-medium are reported in [Multimedia Appendix 1](#) (S.6.1). We did not observe consistent performance gains from using a larger transcription model (Whisper-medium vs Whisper-small).

Table 2. Test results for text- and audio-based models across 5 replications.

Models	Data size ^a	Model size	Accuracy, mean (SD)	AUC ^b , mean (SD)
Text-based pipeline^c				
BERT ^d without pause	3.3 B	110 M	0.657 (0.043)	0.719 (0.026)
BERT with pause	3.3 B	110 M	0.662 (0.013)	0.726 (0.016)
BioBERT without pause	18 B	110 M	0.612 (0.033)	0.720 (0.013)
BioBERT with pause	18 B	110 M	0.648 (0.007)	0.729 (0.011)
BioClinicalBERT without pause	0.5 B	110 M	0.626 (0.039)	0.713 (0.023)
BioClinicalBERT with pause	0.5 B	110 M	0.650 (0.021)	0.712 (0.022)
Audio-based pipeline				
Whisper-tiny.en	680,000 h	39 M	0.673 (0.039)	0.759 (0.010)
Whisper-base.en	680,000 h	74 M	0.707 (0.026)	0.801 (0.020)
Whisper-small.en	680,000 h	244 M	0.701 (0.033)	0.791 (0.018)
Whisper-medium.en	680,000 h	769 M	0.731 (0.020)	0.802 (0.014)
Whisper-large	680,000 h	1.55 B	0.702 (0.028)	0.779 (0.014)
Unispeech-large-1500h-cv	1500 h	94 M	0.659 (0.023)	0.737 (0.014)
WavLM-base	960 h	94 M	0.653 (0.013)	0.755 (0.032)
WavLM-base-plus	94,000 h	94 M	0.663 (0.038)	0.748 (0.018)
WavLM-large	94,000 h	316 M	0.694 (0.016)	0.776 (0.023)
Wav2Vec2-base	960 h	95 M	0.622 (0.028)	0.720 (0.018)
Wav2Vec2-large	960 h	137 M	0.642 (0.021)	0.694 (0.030)
HuBERT-base	960 h	95 M	0.683 (0.027)	0.752 (0.031)
HuBERT-large	960 h	317 M	0.645 (0.025)	0.725 (0.011)
Data2Vec-audio-base	960 h	94 M	0.559 (0.062)	0.532 (0.025)
Data2Vec-audio-large	960 h	317 M	0.629 (0.032)	0.667 (0.030)

^aTraining data size in additional words trained or in hours of audio. BioBERT was trained on additional words of 18 B compared to BERT, and BioClinicalBERT was trained on additional words of 0.5 B compared to BioBERT.

^bAUC: area under the curve.

^cAudio recordings were transcribed by Whisper-small. Refer to the Appendix for results from transcription generated by different sizes of the Whisper family models.

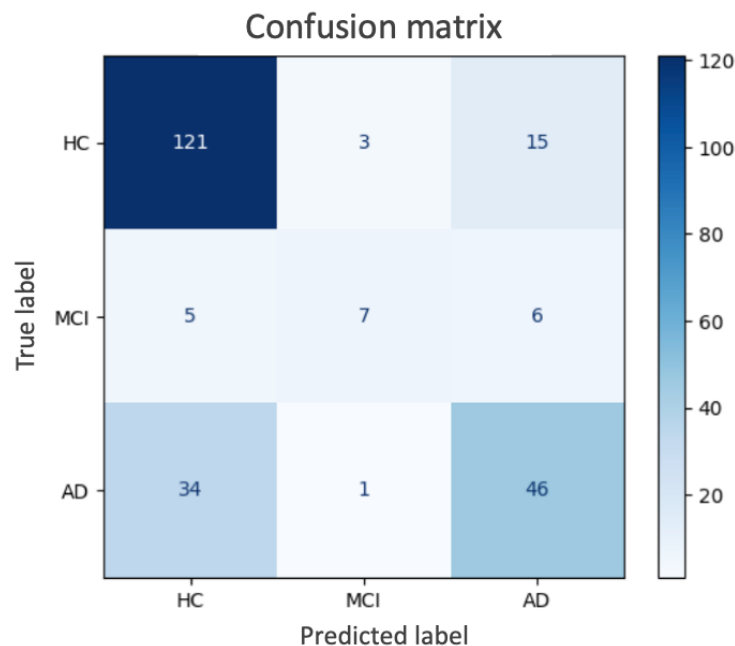
^dBERT: Bidirectional Encoder Representations from Transformers.

Audio-Based Pipeline

Overall, the family of Whisper-based models generally outperformed other audio representation models, likely due to their training on a substantially larger and more diverse multitask dataset (Table 2). Specifically, the medium-sized variant achieved the best performance with an accuracy of 0.731 (SD 0.020) and an AUC of 0.802 (SD 0.014). The 3-class confusion matrix of the best-performing model is presented in Figure 3. Models pretrained on a larger data corpus, such as Whisper, WavLM-base-plus, and WavLM-large, generally outperformed those pretrained on smaller datasets, such as Data2Vec, Wav2Vec, and WavLM-base. Fine-tuning the last layers of Whisper models did not improve performance (Table S4.1 in Multimedia Appendix 1). While performance improved as we

scaled the Whisper model from tiny to medium, Whisper-large (1.55 billion parameters, trained on 680,000 hours) performed slightly worse than Whisper-medium (769 million parameters, trained on the same 680,000 hours), suggesting that scaling beyond a certain point does not necessarily lead to better classification performance. One possibility is that the larger model overfits or overspecializes during pretraining and/or fine-tuning, amplifying domain mismatch from pretraining and producing embeddings that are less aligned with dementia-related acoustic or linguistic patterns. Additionally, the higher-dimensional representations generated by Whisper-large may introduce redundancy or noise that dilutes the subtle discriminative cues needed for this clinical task, making them harder for the classification layer to learn the cues from a small, labeled dataset.

Figure 3. Confusion matrix of the best performing model's (Whisper-medium) predictions on one replication of the test dataset. AD: Alzheimer disease; HC: healthy controls; MCI: mild cognitive impairment.



Comparing results of both text and audio representation models, text-based models achieved relatively similar performance (0.71-0.73 AUC), likely because they share similar BERT-based architectures, whereas audio-based models exhibited a wider performance range (0.53-0.80 AUC) due to greater diversity in model size and design. Overall, audio-based models outperformed text-based models, and text-based models with pause annotation achieved comparable performance with audio representation models, suggesting that acoustic features provide stronger indicators of AD-related changes than semantic features alone. The strongest-performing audio-based Whisper models likely benefit from their ability to capture both acoustic and semantic information, since their decoder predicts text sequences and therefore their encoder must learn representations from raw audio that implicitly preserve acoustic and semantic context. Whisper's advantage may also stem from its substantially larger pretraining corpus and greater model capacity, as Whisper-medium is nearly 7 times larger than the text-based models. While the performance of text-based models may also be affected by ASR errors, recent studies have shown that certain types of transcription errors can, in fact, amplify linguistic irregularities associated with cognitive decline, thereby providing additional predictive signal [41,53].

Results from the traditional feature engineering pipeline are presented in [Multimedia Appendix 1](#) (S3). The modern speech and language models significantly outperformed the traditional approach of using hand-engineered acoustic and linguistic features, demonstrating the strength of the recent deep learning models.

Interpretability Analysis

A visualization of the top contributing embedding dimensions by SHAP is provided in [Multimedia Appendix 1](#) (S7.1). [Figure 4](#) summarizes the acoustic and linguistic markers that have statistically significant relationships with the top 5 embeddings for each class, and the definition of each variable can be found in [Multimedia Appendix 2](#). For HC, the top embeddings are found to have a positive correlation with harmonics-to-noise ratio, which is expected, as a higher harmonics-to-noise ratio reflects clearer, more periodic, and stable vocal production, consistent with preserved laryngeal control and phonatory stability in healthy speakers. Negative correlations with SD of unvoiced segment lengths, zero-crossing rate, L1 norm of auditory spectrum, and SD of shimmer suggest that HC individuals produce more consistent voicing, smoother amplitude variations, and less spectral noise, which fits the literature reporting increased jitter and shimmer in pathological or cognitively impaired speech. For MCI, top contributing embeddings have positive correlations with cycle-to-cycle F0 variation, amplitude variation across cycles, and L1 norm of auditory spectrum, indicating the onset of subtle voice perturbations and increased spectral variability. In AD, embeddings show strong positive correlations with age, jitter, shimmer, and L1 norm of auditory spectrum, alongside negative correlation with zero-crossing rate, confirming the presence of pronounced voice instability, increased aperiodicity, and slower or less articulated speech. Overall, the direction of the correlations aligns with expected physiological and acoustic changes across cognitive decline [21,54]. These results suggest that the high-dimensional embeddings capture rich acoustic and linguistic information that is partially interpretable, extending beyond traditional explainable markers.

Figure 4. Top contributing embedding dimensions that positively and negatively contribute to health control (HC), mild cognitive impairment (MCI), and Alzheimer disease (AD) classification and their statistically related explainable acoustic/linguistic markers. “+” represents contribution toward the positive direction with respect to the class (increased feature value increases the predicted probability of that class). “-“ represents contribution toward the negative direction with respect to the class (decreased feature value increases the predicted probability of that class). Magnitude equals to the absolute value of the coefficient. *P* value is corrected using false discovery rate correction. Age is directly interpretable from the concatenated feature set.

	HC [magnitude, corrected <i>P</i> -value]	MCI [magnitude, corrected <i>P</i> -value]	AD [magnitude, corrected <i>P</i> -value]
+	Embedding dimensions: 305, 20 Related features: • logHNR_sma_de [36.5, <0.001]	Embedding dimensions: 305, 771, 470 Related features: • jitterLocal_sma_de [77151.2, 0.004] • shimmerLocal_sma_de [13605.8, <0.001] • audspect_lengthL1norm_sma_de [2377.3, <0.001]	Embedding dimensions: 1024, 359, 719 Related features: • age • jitterLocal_sma_de [39092.8, 0.008] • shimmerLocal_sma_de [6392.9, <0.001] • audspect_lengthL1norm_sma_de [860.3, 0.01]
-	Embedding dimensions: 470, 359, 931 Related features: • StddevUnvoicedSegmentLength [9.8, <0.001] • pcm_zcr_sma_de [3019.8, 0.01] • audspect_lengthL1norm_sma_de [860.3, 0.01] • shimmerLocaldB_sma3nz_stddevNorm [1.9, 0.01]	Embedding dimensions: 1024, 387 Related features: • age • pcm_zcr_sma_de [3019.8, 0.01]	Embedding dimensions: 305, 771 Related features: • pcm_zcr_sma_de [1855.6, 0.01]

Discussion

Principal Findings

Our results show that acoustic models generally outperform text-only models in dementia detection, highlighting that nonsemantic speech features are useful indicators for dementia detection, echoed by findings in previous literature [29,30,40,55-57]. Notably, when pauses are explicitly annotated in transcripts, text-based language models achieve performance on par with acoustic models. This finding suggests that among nonsemantic features, pauses are particularly informative, capturing subtle disruptions in speech planning and lexical retrieval that correlate with cognitive impairment.

Pauses have long been recognized in clinical research as a sensitive marker of cognitive decline. Silent-pause frequency and duration differ significantly in individuals with AD compared to HC, suggesting planning and retrieval difficulties [58]. These pausing behaviors correlate with neuropathological biomarkers: for example, longer and more frequent pauses have been linked to tau burden in medial temporal and early neocortical regions [59]. Theoretically, pauses are thought to reflect impairments in cognitive functions that decline early in dementia, such as lexical retrieval, working memory, and speech planning. Reduced cognitive resources or resource depletion (eg, working memory and executive function) may force speakers to insert compensatory pauses during more demanding speech tasks [60]. Pause-related patterns are not only theoretically meaningful but also practically useful: they can be passively captured from spontaneous speech collected in mobile or remote applications, enabling low-cost cognitive screening.

We noticed that using traditional acoustic features alone can achieve prediction accuracy comparable to several foundation models [29,55-57]. However, traditional pipelines rely on manual preselection of acoustic and linguistic features, which may not fully capture the multiple speech characteristics, such as timing, intonation, spectral properties, and pause patterns, that jointly reflect cognitive decline. Some of these discriminative acoustic features manifest as subtle but meaningful deviations from typical speech, which can be difficult to identify or quantify manually, particularly in

early-stage disease. In contrast, embedding-based foundation models, particularly those derived from large-scale pretrained speech models, offer greater flexibility and data-driven representation learning. The Whisper-medium model significantly outperformed both the traditional feature-based models and other embedding models, highlighting its strong capacity to capture relevant speech characteristics for AD detection without the need for handcrafted features.

Overall, the proposed audio-based AD diagnosis pipeline achieves promising accuracy across the 3 categories. However, the MCI group is significantly underrepresented in our dataset, and models are less accurate and reliable for MCI detection. In reality, the proportion of AD and MCI in the population stands for a different proportion than modeling population in this study. In particular, in the population aged >65 years, we have 29.7% (354/1189) AD and 5.7% (68/1189) MCI in the modeling population. According to the 2025 AD facts and figures, an estimated 7.2 million Americans aged 65 years and older currently live with AD, with approximately 11% aged 65 years and older having AD [61]. There is no accurate MCI estimated proportion in the population, but it is more than doubled than AD across all populations, according to the 2014, 2016, 2018, and 2020 Health and Retirement Study, that all-cause MCI was consistently 23% in each survey [62].

MCI is hard to detect early because the signs are subtle, and there is no simple test for it. Unlike diseases that show clear signals before symptoms, diagnosing MCI and AD relies on recognizing memory and thinking problems, usually first noticed by the person or their family, then confirmed by a doctor using interviews, tests, and sometimes brain scans. Early diagnosis requires noticing small changes in how long it takes people to do everyday tasks or more obvious problems with daily living. Diagnosis is further complicated because so far, there is no single test that can confirm AD or MCI, which makes the process slow and complex. Primary care doctors can do basic memory tests but may feel unprepared to interpret or explain the results, often resulting in referrals and delayed diagnoses when specialists and equipment aren't locally available [63].

A machine learning-based tool for automatic AD screening via speech holds strong promise for providing an accessible and scalable workflow for early detection. The ability to distinguish

HC from individuals with MCI supports early detection, which is critical for timely intervention and monitoring of cognitive decline. Similarly, accurate identification of HC vs AD and MCI vs AD cases enables timely diagnosis and treatment, as the efficacy of current and emerging therapies to slow disease progression is contingent upon early administration (Figure S1.1 in [Multimedia Appendix 1](#)) [64]. Participants would be guided through simple spoken tasks, such as describing a picture or narrating a recent experience, with spontaneous speech captured directly through their device's microphone. The app would process the audio locally or via a secure cloud service using a pretrained foundation speech model (eg, Whisper) to assess cognitive risk. For healthy individuals, periodic screenings, such as once every 3 months, may suffice for proactive monitoring. If a user's speech is flagged as high risk for MCI, the app would recommend follow-up through an in-clinic cognitive assessment to confirm diagnosis. Early identification of MCI allows for timely intervention, personalized treatment planning, and patient education. Once diagnosed, individuals could continue using the app at a higher frequency (eg, weekly) to monitor their risk score progression and treatment response. This approach supports continuous, personalized cognitive health management in a noninvasive, scalable, and cost-effective manner. By supporting early, low-cost, and remote screening, foundation speech models offer a path toward more accessible Alzheimer caregiving and improved clinical outcomes across diverse real-world environments.

Comparison With Prior Work

While our best model achieved an accuracy of 0.73, the classification accuracy in AD detection using audio data is slightly higher in several prior studies [28,29,41,42]. Previous study reported an accuracy of 0.68 for eGeMAPS, 0.72 using Wav2Vec2, and 0.73 using data2vec [65], with 1 paper reaching 0.89 [29]. We believe that the difference is due to the diversity of our dataset and the greater complexity of our classification task. Our dataset includes nearly 1200 samples with class imbalance, whereas many prior works on machine learning for AD detection used the Alzheimer's Dementia Recognition Through Spontaneous Speech 2020 challenge dataset, which is a curated dataset with around 200 samples balanced by age, gender, and diagnosis category [40]. While added variability increases the difficulty of achieving high accuracy, the dataset used in this study reflects a larger and more diverse population, and the type of audio data is also more scalable and easier to obtain in real-world settings. Moreover, most prior studies that achieved higher accuracy focused on a simpler binary classification problem (HC vs AD), while we have a more demanding 3-class classification problem (HC, MCI, and AD). MCI is particularly harder to classify, because it lies on a clinical and biological continuum between normal cognition and AD. The lower accuracy for MCI is also consistent with prior literature [66].

Limitations

There are several limitations in this study. First, our dataset may not fully represent the broader population in terms of language background, education level, or dialect variation, which could affect acoustic characteristics and thus limit the generalizability of our findings. Second, the MCI class is significantly underrepresented in our dataset. As a result, the model performs more accurately and reliably when distinguishing HC from AD than when detecting MCI. Although we applied a commonly used machine learning strategy, oversampling minority classes in the training data, we observed no meaningful performance improvement. Future studies should aim to collect larger and more balanced samples of MCI participants to enhance both generalizability and clinical use. Additionally, while embeddings extracted from foundation models function as black boxes with limited interpretability. While we conducted post hoc interpretability analysis, more advanced explainability techniques can be explored in the future.

Future research could also explore multilingual models to include non-English samples from the dataset, such as Spanish, Gallego, and Chinese samples. Additionally, incorporating multimodal feature sets, such as participants' socioeconomic, medical history, and lifestyle information, could provide a more interpretable and comprehensive analysis of the individual's cognitive status and enhance model performance. As ADRD is a progressive disorder, it would be interesting to collect longitudinal speech data in the future and develop models that are effective at detecting subtle within-subject changes and tracking disease progression over time. Lastly, the study can be extended to examine the stigma associated with ADRD diagnosis, as stigma can significantly deter people from accepting screening or disclosing their status, and the best strategies to enhance the acceptance of audio-based screening tools once they are available.

Conclusions

Our study presents a comprehensive benchmark for ADRD early detection across several foundation speech and language models and traditional machine learning models on a large, diverse, and clinically relevant dataset. Compared to pretrained language models and more traditional machine learning models, pretrained speech models such as Whisper capture both semantic and nonsemantic acoustic cues that are indicative of cognitive decline and have shown strong predictive performance even without manual transcription. Among the benchmarked speech and language models, Whisper-medium achieved the best performance on this large dataset. Our findings suggest that both semantic and nonsemantic information are crucial for ADRD detection. Future acoustic-based ADRD detection tools should incorporate both semantic and nonsemantic features. The use of acoustic-based automated ADRD detection tools offers a scalable, noninvasive, and cost-effective approach for population-level surveillance that supports earlier intervention and treatment. Technologies such as telehealth and mobile health further enhance their potential for real-world clinical deployment.

Acknowledgments

We want to thank DrivenData Pioneering Research for Early Prediction of Alzheimer's and Related Dementias EUREKA (PREPARE) Challenge organizing team for providing the data and making this study possible.

Funding

This work was supported in part by the University of Texas Rio Grande Valley Vackar College of Business & Entrepreneurship Faculty Development Fund. No other external funding was received for this study.

Data Availability

The data used in this study are available through the DementiaBank DrivenData Challenge website [67]. Users may register on TalkBank [37] and DrivenData [68] to request access. The code used for this study is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JL, LM, XM, HW, XSN

Data Curation: JL, LM

Formal Analysis: JL, LM

Investigation: JL, LM

Methodology: JL, LM

Project Administration: XM

Supervision: XSN

Validation: ZW

Visualization: JL, LM, HW, XM

Writing – Original Draft: JL, LM, HW, ZW, XM, XSN

Writing – Review & Editing: JL, LM, ZW, XM, XSN

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary material: visual overview; clustering of BERT embeddings; additional details for the traditional feature engineering pipeline; additional details for the Whisper family of models; additional details for the interpretability analysis.

[\[DOCX File , 1038 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Variable definition of 4 groups of acoustic and linguistic features.

[\[XLSX File \(Microsoft Excel File\), 25 KB-Multimedia Appendix 2\]](#)

References

1. Jack Jr CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14(4):535-562. [\[FREE Full text\]](#) [doi: [10.1016/j.jalz.2018.02.018](https://doi.org/10.1016/j.jalz.2018.02.018)] [Medline: [29653606](https://pubmed.ncbi.nlm.nih.gov/29653606/)]
2. Petersen RC, Roberts RO, Knopman DS, Boeve BF, Geda YE, Ivnik RJ, et al. Mild cognitive impairment: ten years later. *Arch Neurol*. 2009;66(12):1447-1455. [\[FREE Full text\]](#) [doi: [10.1001/archneurol.2009.266](https://doi.org/10.1001/archneurol.2009.266)] [Medline: [20008648](https://pubmed.ncbi.nlm.nih.gov/20008648/)]
3. Van Dyck CH, Swanson CJ, Aisen P, Bateman RJ. Lecanemab in early Alzheimer's disease. *New Engl J Med*. 2023;388(1):9-12. [\[FREE Full text\]](#)
4. Sims JR, Zimmer JA, Evans CD, Lu M. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. *JAMA*. 2023;330(6):512-527. [\[FREE Full text\]](#) [doi: [10.1001/jama.2023.21109](https://doi.org/10.1001/jama.2023.21109)] [Medline: [38112817](https://pubmed.ncbi.nlm.nih.gov/38112817/)]
5. Mattke S, Shi Z, Hanson M, Mitchell S, Lynch C, MacLean Kalonji K, et al. Estimated investment need to increase England's capacity to diagnose eligibility for an Alzheimer's treatment to G7 average capacity levels. *J Prev Alzheimers Dis*. 2024;11(4):1022-1029. [\[FREE Full text\]](#) [doi: [10.14283/jpad.2024.24](https://doi.org/10.14283/jpad.2024.24)] [Medline: [39044513](https://pubmed.ncbi.nlm.nih.gov/39044513/)]
6. Roth S, Burnie N, Suridjan I, Yan JT, Carboni M. Current diagnostic pathways for Alzheimer's disease: a cross-sectional real-world study across six countries. *J Alzheimers Dis Rep*. 2023;7(1):659-674. [\[FREE Full text\]](#) [doi: [10.3233/ADR230007](https://doi.org/10.3233/ADR230007)] [Medline: [37483324](https://pubmed.ncbi.nlm.nih.gov/37483324/)]

7. Liu Y, Jun H, Becker A, Wallick C, Mattke S. Detection rates of mild cognitive impairment in primary care for the United States Medicare population. *J Prev Alzheimers Dis.* 2024;11(1):7-12. [FREE Full text] [doi: [10.14283/jpad.2023.131](https://doi.org/10.14283/jpad.2023.131)] [Medline: [38230712](https://pubmed.ncbi.nlm.nih.gov/38230712/)]
8. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia-meta-analysis of 41 robust inception cohort studies. *Acta Psychiatr Scand.* 2009;119(4):252-265. [FREE Full text] [doi: [10.1111/j.1600-0447.2008.01326.x](https://doi.org/10.1111/j.1600-0447.2008.01326.x)] [Medline: [19236314](https://pubmed.ncbi.nlm.nih.gov/19236314/)]
9. Mattke S, Jun H, Chen E, Liu Y, Becker A, Wallick C. Expected and diagnosed rates of mild cognitive impairment and dementia in the U.S. Medicare population: observational analysis. *Alzheimers Res Ther.* 2023;15(1):128. [FREE Full text] [doi: [10.1186/s13195-023-01272-z](https://doi.org/10.1186/s13195-023-01272-z)] [Medline: [37481563](https://pubmed.ncbi.nlm.nih.gov/37481563/)]
10. Gilmore-Bykovskiy AL, Jin Y, Gleason C, Flowers-Benton S, Block LM, Dilworth-Anderson P, et al. Recruitment and retention of underrepresented populations in Alzheimer's disease research: a systematic review. *Alzheimers Dement (N Y).* 2019;5(1):751-770. [FREE Full text] [doi: [10.1016/j.trci.2019.09.018](https://doi.org/10.1016/j.trci.2019.09.018)] [Medline: [31921966](https://pubmed.ncbi.nlm.nih.gov/31921966/)]
11. Glymour MM, Manly JJ. Lifecourse social conditions and racial and ethnic patterns of cognitive aging. *Neuropsychol Rev.* 2008;18(3):223-254. [FREE Full text] [doi: [10.1007/s11065-008-9064-z](https://doi.org/10.1007/s11065-008-9064-z)] [Medline: [18815889](https://pubmed.ncbi.nlm.nih.gov/18815889/)]
12. Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front Aging Neurosci.* 2018;10:369. [FREE Full text] [doi: [10.3389/fnagi.2018.00369](https://doi.org/10.3389/fnagi.2018.00369)] [Medline: [30483116](https://pubmed.ncbi.nlm.nih.gov/30483116/)]
13. Shakeri A, Farmanbar M. Natural language processing in Alzheimer's disease research: systematic review of methods, data, and efficacy. *Alzheimers Dement (Amst).* 2025;17(1):e70082. [FREE Full text] [doi: [10.1002/dad2.70082](https://doi.org/10.1002/dad2.70082)] [Medline: [39935888](https://pubmed.ncbi.nlm.nih.gov/39935888/)]
14. Zhu Y, Tran B, Liang X, Batsis JA, Roth RM. Towards interpretability of speech pause in dementia detection using adversarial learning. 2022. Presented at: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore. URL: <https://europepmc.org/abstract/MED/37064829> [doi: [10.1109/icassp43922.2022.9747006](https://doi.org/10.1109/icassp43922.2022.9747006)]
15. Gómez-Zaragozá L, Wills S, Tejedor-García C, Marín-Morales J. Alzheimer disease classification through ASR-based transcriptions: exploring the impact of punctuation and pauses. *ArXiv.* Preprint posted online on June 6, 2023. [doi: [10.21437/interspeech.2023-1734](https://doi.org/10.21437/interspeech.2023-1734)]
16. Radford A, Kim JW, Xu T, Brockman G. Robust speech recognition via large-scale weak supervision. 2023. Presented at: International Conference on Machine Learning; Dec 6, 2022; Seoul, South Korea. URL: <https://proceedings.mlr.press/v202/radford23a.html>
17. Goodglass H, Kaplan E. Boston Diagnostic Aphasia Examination Booklet. Philadelphia. Lea & Febiger; 1983.
18. Ester M, Kriegel HP, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise. Association for the Advancement of Artificial Intelligence. URL: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf?source=post_page [accessed 2026-04-09]
19. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis.* 2015;49(2):407-422. [FREE Full text] [doi: [10.3233/JAD-150520](https://doi.org/10.3233/JAD-150520)] [Medline: [26484921](https://pubmed.ncbi.nlm.nih.gov/26484921/)]
20. Meilán JGG, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord.* 2014;37(5-6):327-334. [FREE Full text] [doi: [10.1159/000356726](https://doi.org/10.1159/000356726)] [Medline: [24481220](https://pubmed.ncbi.nlm.nih.gov/24481220/)]
21. König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst).* 2015;1(1):112-124. [FREE Full text] [doi: [10.1016/j.dadm.2014.11.012](https://doi.org/10.1016/j.dadm.2014.11.012)] [Medline: [27239498](https://pubmed.ncbi.nlm.nih.gov/27239498/)]
22. Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of speech-based digital biomarkers: review and recommendations. *Digit Biomark.* 2020;4(3):99-108. [FREE Full text] [doi: [10.1159/000510820](https://doi.org/10.1159/000510820)] [Medline: [33251474](https://pubmed.ncbi.nlm.nih.gov/33251474/)]
23. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine.* Nov 2020;28:100583. [FREE Full text] [doi: [10.1016/j.eclinm.2020.100583](https://doi.org/10.1016/j.eclinm.2020.100583)] [Medline: [33294808](https://pubmed.ncbi.nlm.nih.gov/33294808/)]
24. Harrell KM, Wilkins SS, Connor MK, Chodosh J. Telemedicine and the evaluation of cognitive impairment: the additive value of neuropsychological assessment. *J Am Med Dir Assoc.* 2014;15(8):600-606. [doi: [10.1016/j.jamda.2014.04.015](https://doi.org/10.1016/j.jamda.2014.04.015)] [Medline: [24913209](https://pubmed.ncbi.nlm.nih.gov/24913209/)]
25. Cullum CM, Hynan L, Grosch M, Parikh M, Weiner M. Teleneuropsychology: evidence for video teleconference-based neuropsychological assessment. *J Int Neuropsychol Soc.* 2014;20(10):1028-1033. [FREE Full text] [doi: [10.1017/s1355617714000873](https://doi.org/10.1017/s1355617714000873)]
26. Meilán JJ, Martínez-Sánchez F, Carro J, Sánchez JA. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *Span J Psychol.* 2012;15(2):487-494. [doi: [10.5209/rev_sjop.2012.v15.n2.38859](https://doi.org/10.5209/rev_sjop.2012.v15.n2.38859)] [Medline: [22774422](https://pubmed.ncbi.nlm.nih.gov/22774422/)]
27. Viliam R, Shona D, Nils P, Ian R, Richard R. The use of telephone speech recordings for assessment and monitoring of cognitive function in elderly people. *ISCA.* 2009:1. [FREE Full text] [doi: [10.21437/interspeech.2009-282](https://doi.org/10.21437/interspeech.2009-282)]
28. Haulcy RM, Glass J. Classifying Alzheimer's disease using audio and text-based representations of speech. *Front Psychol.* 2021;11:624137. [FREE Full text] [doi: [10.3389/fpsyg.2020.624137](https://doi.org/10.3389/fpsyg.2020.624137)] [Medline: [33519651](https://pubmed.ncbi.nlm.nih.gov/33519651/)]

29. Yuan J, Cai X, Bian Y, Ye Z. Pauses for detection of Alzheimer's disease. *Front Comput Sci*. 2021;2:624488. [[FREE Full text](#)]
30. Zhu Y, Obyat A, Liang X, Batsis JA, Roth RM. WavBERT: Exploiting Semantic and Non-semantic Speech using Wav2vec and BERT for Dementia Detection. *Interspeech*. 2021;2021:3790-3794. [[FREE Full text](#)] [doi: [10.21437/interspeech.2021-332](https://doi.org/10.21437/interspeech.2021-332)] [Medline: [37063977](https://pubmed.ncbi.nlm.nih.gov/37063977/)]
31. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Trans Affective Comput*. 2016;7(2):190-202. [[FREE Full text](#)] [doi: [10.1109/taffc.2015.2457417](https://doi.org/10.1109/taffc.2015.2457417)]
32. Schuller B, Steidl S, Batliner A. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. *Interspeech*. 2013. [[FREE Full text](#)] [doi: [10.21437/interspeech.2013-56](https://doi.org/10.21437/interspeech.2013-56)]
33. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC. Mahway: Lawrence Erlbaum Associates. 2001. URL: http://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf [accessed 2026-04-21]
34. McFee B, Raffel C, Liang D, Ellis DP. librosa: audio and music signal analysis in Python. 2015. Presented at: 14th Python in Science Conference (SciPy 2015); July 6-12, 2015:18-24; Austin, Texas. URL: <https://www.academia.edu/download/40296500/librosa.pdf> [doi: [10.25080/majora-7b98e3ed-003](https://doi.org/10.25080/majora-7b98e3ed-003)]
35. PREPARE challenge - phase 2: model arena (acoustic track). DrivenData. URL: <https://www.drivendata.org/competitions/299/competition-nih-alzheimers-acoustic-2/> [accessed 2024-04-15]
36. Lanzi AM, Saylor AK, Fromm D, Liu H, MacWhinney B, Cohen ML. DementiaBank: theoretical rationale, protocol, and illustrative analyses. *Am J Speech Lang Pathol*. 2023;32(2):426-438. [[FREE Full text](#)] [doi: [10.1044/2022_ajslp-22-00281](https://doi.org/10.1044/2022_ajslp-22-00281)]
37. DementiaBank. 2025. URL: <https://talkbank.org/dementia/> [accessed 2026-04-09]
38. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. Presented at: 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); June 1, 2019; Minneapolis, Minnesota. URL: <https://aclanthology.org/N19-1423.pdf>
39. Abdul ZK, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. *IEEE Access*. 2022;10:122136-122158. [[FREE Full text](#)] [doi: [10.1109/access.2022.3223444](https://doi.org/10.1109/access.2022.3223444)]
40. Luz S, Haider F, Fuente Garcia SDL, Fromm D. Alzheimer's dementia recognition through spontaneous speech. *Frontiers Media SA*. 2021. URL: <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2021.780169/pdf> [accessed 2026-04-09]
41. Li C, Xu W, Cohen T, Pakhomov S. Useful blunders: can automated speech recognition errors improve downstream dementia classification? *J Biomed Inform*. 2024;150:104598. [[FREE Full text](#)] [doi: [10.1016/j.jbi.2024.104598](https://doi.org/10.1016/j.jbi.2024.104598)] [Medline: [38253228](https://pubmed.ncbi.nlm.nih.gov/38253228/)]
42. Balagopalan A, Eyre B, Rudzicz F, Novikova J. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *ArXiv*. Preprint posted online on July 26, 2020. [[FREE Full text](#)] [doi: [10.21437/interspeech.2020-2557](https://doi.org/10.21437/interspeech.2020-2557)]
43. Schroter H, Escalante-B AN, Rosenkranz T, Maier A. DeepFilterNet: a low complexity speech enhancement framework for full-band audio based on deep filtering. 2022. Presented at: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 22, 2022; Shenzhen, China. URL: <https://arxiv.org/pdf/2110.05588> [doi: [10.1109/icassp43922.2022.9747055](https://doi.org/10.1109/icassp43922.2022.9747055)]
44. Lee, J, Yoon W, Kim S, Kim SD. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. [[FREE Full text](#)] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)]
45. Alsentzer E, Murphy JR, Boag W, Weng WH. Publicly available clinical BERT embeddings. *ArXiv*. Preprint posted online on June 20, 2019. [[FREE Full text](#)] [doi: [10.18653/v1/w19-1909](https://doi.org/10.18653/v1/w19-1909)]
46. Baevski A, Zhou Y, Mohamed A, Auli M. Wav2Vec 2.0: a framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems*. 2020. Presented at: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; December 6-12, 2020:12449-12460; Virtual. URL: <https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf>
47. Hsu W, Bolte B, Tsai YH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans Audio Speech Lang Process*. 2021;29:3451-3460. [[FREE Full text](#)] [doi: [10.1109/taslp.2021.3122291](https://doi.org/10.1109/taslp.2021.3122291)]
48. Wang C, Wu Y, Qian Y, Kumatani K. Unispeech: unified speech representation learning with labeled and unlabeled data. In: *International Conference on Machine Learning*. 2021. Presented at: Proceedings of the 38th International Conference on Machine Learning, PMLR; July 18-24, 2021; Virtual event. URL: <http://proceedings.mlr.press/v139/wang21y/wang21y.pdf>
49. Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing. *IEEE J Sel Top Signal Process*. 2022;16(6):1505-1518. [[FREE Full text](#)] [doi: [10.1109/jstsp.2022.3188113](https://doi.org/10.1109/jstsp.2022.3188113)]
50. Baevski A, Hsu WN, Xu Q, Babu A. Data2vec: A general framework for self-supervised learning in speech, vision and language. 2022. Presented at: Proceedings of the 39th International Conference on Machine Learning; July 17-23, 2022; Baltimore, Maryland, USA. URL: <https://proceedings.mlr.press/v162/baevski22a/baevski22a.pdf>

51. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. 2015. Presented at: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); April 19-24, 2015; Brisbane. URL: http://www.danielpovey.com/files/2015_icassp_librispeech.pdf [doi: [10.1109/icassp.2015.7178964](https://doi.org/10.1109/icassp.2015.7178964)]
52. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017. Presented at: Advances in Neural Information Processing Systems; December 4-9, 2017; California, USA. URL: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
53. Kang J, Li J, Li J, Wu X. Not all errors are equal: investigation of speech recognition errors in Alzheimer's disease detection. 2024. Presented at: IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP); November 7-10, 2024; Beijing, China. URL: <https://ieeexplore.ieee.org/abstract/document/10800483> [doi: [10.1109/iscslp63861.2024.10800483](https://doi.org/10.1109/iscslp63861.2024.10800483)]
54. Little M, McSharry P, Hunter E, Spielman J. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Nature Precedings*. 2008. URL: <https://www.nature.com/articles/npre.2008.2298.1.pdf> [accessed 2026-04-09]
55. Haider F, de la Fuente S, Luz S. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE J Sel Top Signal Process*. 2020;14(2):272-281. [FREE Full text] [doi: [10.1109/jstsp.2019.2955022](https://doi.org/10.1109/jstsp.2019.2955022)]
56. Tóth L, Gosztolya G, Vincze V, Hoffmann I. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. 2015. Presented at: INTERSPEECH; September 6-10, 2015; Dresden, Germany. URL: https://www.isca-archive.org/interspeech_2015/toth15_interspeech.pdf [doi: [10.21437/interspeech.2015-568](https://doi.org/10.21437/interspeech.2015-568)]
57. Fraser KC, Rudzicz F, Graham N, Rochon E. Automatic speech recognition in the diagnosis of primary progressive aphasia. Association for Computational Linguistics; 2013. Presented at: Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies; April 8, 2026:47-54; Grenoble, France. URL: <https://aclanthology.org/W13-3909/>
58. Gayraud F, Lee H, Barkat-Defradas M. Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. *Clin Linguist Phon*. 2011;25(3):198-209. [doi: [10.3109/02699206.2010.521612](https://doi.org/10.3109/02699206.2010.521612)] [Medline: [21080826](https://pubmed.ncbi.nlm.nih.gov/21080826/)]
59. Young CB, Smith V, Karjadi C, Grogan S, Ang TFA, Insel PS, et al. Speech patterns during memory recall relates to early tau burden across adulthood. *Alzheimers Dement*. 2024;20(4):2552-2563. [FREE Full text] [doi: [10.1002/alz.13731](https://doi.org/10.1002/alz.13731)] [Medline: [38348772](https://pubmed.ncbi.nlm.nih.gov/38348772/)]
60. Wei HT, Kulzhabayeva D, Erceg L, Robin J, Hu YZ, Chignell M, et al. Cognitive components of aging-related increase in word-finding difficulty. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*. 2024;31(6):987-1019. [FREE Full text] [doi: [10.1080/13825585.2024.2315774](https://doi.org/10.1080/13825585.2024.2315774)] [Medline: [38353604](https://pubmed.ncbi.nlm.nih.gov/38353604/)]
61. Alzheimer's Association. 2015 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2015;11(3):332-384. [doi: [10.1016/j.jalz.2015.02.003](https://doi.org/10.1016/j.jalz.2015.02.003)] [Medline: [25984581](https://pubmed.ncbi.nlm.nih.gov/25984581/)]
62. Tahami Monfared AA, Hummel N, Chandak A, Khachatrian A, Zhang R, Zhang Q. Prevalence estimation of dementia/Alzheimer's disease using health and retirement study database in the United States. *J Prev Alzheimers Dis*. 2024;11(5):1183-1188. [doi: [10.14283/jpad.2024.114](https://doi.org/10.14283/jpad.2024.114)] [Medline: [39350362](https://pubmed.ncbi.nlm.nih.gov/39350362/)]
63. American perspectives on early detection of Alzheimer's disease in the era of treatment. Alzheimer's Association. 2025. URL: <https://www.alz.org/getmedia/3d226bf2-0690-48d0-98ac-d790384f4ec2/alzheimers-facts-and-figures-special-report.pdf> [accessed 2026-04-09]
64. Nakashima S, Sato K, Niimi Y, Ihara R, Suzuki K, Iwata A, et al. Alzheimer's Disease Neuroimaging Initiative. Therapeutic time window of disease-modifying therapy for early Alzheimer's disease. *Alzheimers Dement (N Y)*. 2025;11(2):e70102. [doi: [10.1002/trc2.70102](https://doi.org/10.1002/trc2.70102)] [Medline: [40453977](https://pubmed.ncbi.nlm.nih.gov/40453977/)]
65. Agbavor F, Liang H. Artificial intelligence-enabled end-to-end detection and assessment of Alzheimer's disease using voice. *Brain Sci*. 2022;13(1):28. [FREE Full text] [doi: [10.3390/brainsci13010028](https://doi.org/10.3390/brainsci13010028)] [Medline: [36672010](https://pubmed.ncbi.nlm.nih.gov/36672010/)]
66. Richter-Laskowska M, Sobotnicka E, Bednorz A. Cognitive performance classification of older patients using machine learning and electronic medical records. *Sci Rep*. 2025;15(1):6564. [FREE Full text] [doi: [10.1038/s41598-025-90460-y](https://doi.org/10.1038/s41598-025-90460-y)] [Medline: [39994339](https://pubmed.ncbi.nlm.nih.gov/39994339/)]
67. DementiaBank. URL: <https://talkbank.org/dementia/DrivenData/> [accessed 2026-04-10]
68. DrivenData. URL: <https://www.drivendata.org/> [accessed 2026-04-10]

Abbreviations

- AD:** Alzheimer disease
- ADRD:** Alzheimer disease and related dementias
- ASR:** automatic speech recognition
- AUC:** area under the curve
- BERT:** Bidirectional Encoder Representations from Transformers
- eGeMAPS:** Geneva Minimalistic Acoustic Parameter Set
- HC:** healthy control
- IRB:** institutional review board
- MCI:** mild cognitive impairment

PREPARE: Pioneering Research for Early Prediction of Alzheimer's and Related Dementias EUREKA

SHAP: Shapley additive explanations

Edited by A Mavragani; submitted 20.Jun.2025; peer-reviewed by M Split, J Li, C-Y Chen; comments to author 24.Sep.2025; revised version received 18.Dec.2025; accepted 22.Dec.2025; published 13.May.2026

Please cite as:

Li J, Mao L, Mao X, Wang H, Wang Z, Ni XS

Early Detection of Alzheimer's Disease and Related Dementias From Spontaneous Speech Using Foundation Speech and Language Models: Comparative Evaluation

JMIR Form Res 2026;10:e79411

URL: <https://formative.jmir.org/2026/1/e79411>

doi: [10.2196/79411](https://doi.org/10.2196/79411)

PMID:

©Jingyu Li, Lingchao Mao, Xi Mao, Hairong Wang, Zhendong Wang, Xuelei Sherry Ni. Originally published in JMIR Formative Research (<https://formative.jmir.org>), 13.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://formative.jmir.org>, as well as this copyright and license information must be included.