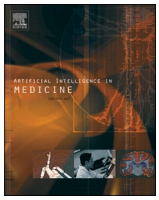
Artificial Intelligence In Medicine 143 (2023) 102624

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Research paper

ADscreen: A speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia

Maryam Zolnoori [a,b,*,1], Ali Zolnour [c,1], Maxim Topaz [a,b]

[a] Columbia University Medical Center, New York, NY, United States of America
[b] School of Nursing, Columbia University, New York, NY, United States of America
[c] School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

ARTICLE INFO

Keywords:
Alzheimer's disease and related dementias
Speech analysis
Natural language processing
Machine learning
Screening algorithm

ABSTRACT

Alzheimer's disease and related dementias (ADRD) present a looming public health crisis, affecting roughly 5 million people and 11 % of older adults in the United States. Despite nationwide efforts for timely diagnosis of patients with ADRD, >50 % of them are not diagnosed and unaware of their disease. To address this challenge, we developed ADscreen, an innovative speech-processing based ADRD screening algorithm for the protective identification of patients with ADRD. ADscreen consists of five major components: (i) noise reduction for reducing background noises from the audio-recorded patient speech, (ii) modeling the patient's ability in phonetic motor planning using acoustic parameters of the patient's voice, (iii) modeling the patient's ability in semantic and syntactic levels of language organization using linguistic parameters of the patient speech, (iv) extracting vocal and semantic psycholinguistic cues from the patient speech, and (v) building and evaluating the screening algorithm. To identify important speech parameters (features) associated with ADRD, we used the Joint Mutual Information Maximization (JMIM), an effective feature selection method for high dimensional, small sample size datasets. Modeling the relationship between speech parameters and the outcome variable (presence/absence of ADRD) was conducted using three different machine learning (ML) architectures with the capability of joining informative acoustic and linguistic with contextual word embedding vectors obtained from the DistilBERT (Bidirectional Encoder Representations from Transformers). We evaluated the performance of the ADscreen on an audio-recorded patients' speech (verbal description) for the Cookie-Theft picture description task, which is publicly available in the dementia databank. The joint fusion of acoustic and linguistic parameters with contextual word embedding vectors of DistilBERT achieved F1-score = 84.64 (standard deviation [std] = ±3.58) and AUC-ROC = 92.53 (std = ±3.34) for training dataset, and F1-score = 89.55 and AUC-ROC = 93.89 for the test dataset. In summary, ADscreen has a strong potential to be integrated with clinical workflow to address the need for an ADRD screening tool so that patients with cognitive impairment can receive appropriate and timely care.

1. Introduction

Alzheimer's disease and related dementias (ADRD) represent a looming public health crisis, affecting roughly 5 million people and 11 % of older adults in the United States [1]. ADRD patients are frequent utilizers of healthcare services in general [2,3] and emergency department services [4,5] in particular, and they incur higher costs of care compared with non-ADRD patients [2,6]. Despite nationwide efforts for timely diagnosis of ADRD, >50 % of these patients remain underdiagnosed and undertreated [7–9]. This is mostly due to patients' inability to recognize early symptoms [10], limited availability of biomarkers (e.g., cerebrospinal fluid, magnetic resonance imaging [11]) [12], and clinicians' insufficient time to assess patients for ADRD [13]. Given the projection of 13.2 million ADRD patients by 2050 [14], development of a robust screening tool for early identification of elderly patients with ADRD has been recognized as a research priority by the National Institute on Aging (NIA) [9,15].

Emerging studies show that patients' spoken language is one of the earliest signs of cognitive impairment, enabling the features of spoken language to act as biomarkers for multiple dimensions of cognitive

* Corresponding author at: Columbia University Medical Center, New York, NY, United States of America.
E-mail addresses: m.zolnoori@gmail.com, mz2825@cumc.columbia.edu (M. Zolnoori).
[1] Authors contributed equally.

https://doi.org/10.1016/j.artmed.2023.102624
Received 9 October 2022; Received in revised form 22 June 2023; Accepted 8 July 2023
Available online 17 July 2023

abilities, including executive functioning, semantic memory, and language [16–18]. Cues of cognitive impairment conveyed in the voice have been empirically documented by measurement of the acoustic waveform (parameters), reflecting the shape of the vocal tract and the patient's abilities to control vocal cord execution during speech. Also, the language of patients with cognitive impairment conveys cues such as low coherence or low information density that mostly occur due to the memory deficit [19]. These language cues can be identified using established methods in the natural language processing domain, such as metrics for measuring information density and contextual word embedding methods for modeling disfluency in patient speech. Additionally, alternations in emotion expression in patients with ADRD can develop in parallel with cognitive deterioration. This alternation can affect the psycholinguistic features of the voice, and in turn the patient ability in communication and expression of their needs to some degree [20,21]. The psycholinguistic cues of speech can be modeled using both nonverbal vocalization (non-word speech) or semantically via language. Changes in nonverbal vocalization can be estimated by different phonatory and articulatory parameters of the acoustic waveform [22]. The identification of semantic psycholinguists cues in language can be achieved by utilizing a set of linguistic features, including lexical-based natural language processing tools specifically designed to analyze the psychological aspects of language [23].

In this research, we developed an innovative screening method called ADscreen for proactive, automated identification of patients at risk for ADRD. This study marks the first time we have developed a unique pipeline to model three primary elements of speech, specifically phonetic motor planning, semantic and syntactic language organization, and psycholinguistic features of patient speech. We utilized this pipeline to process audio recordings of patients' verbal descriptions during the "Cookie-Theft" picture description test (referred to as the Cookie-Theft test) for each element. This audio-recorded dataset is part of the publicly available DementiaBank English Pitt Corpus. Different machine learning (ML) models were trained on the training and evaluated on the test dataset to provide an unbiased evaluation of the performance of the screening algorithm.

2. Related works

To develop a screening algorithm for identifying patients with ADRD, previous studies used different approaches to process the audio-recorded patients' verbal descriptions for the Cookie-Theft test available in the DementiaBank English Pitt Corpus. Some of the studies focused only on the acoustic (voice) or linguistic (transcription of the verbal description) part of speech, while others analyzed both parts and highlighted the importance of each part in detecting patients with ADRD. Additionally, they used different ML architectures and evaluation mechanisms to build the screening algorithm. This section provides a brief review of studies' approaches for generating acoustic and linguistic features and ML architecture used to analyze the features for building an ADRD screening algorithm. We included studies demonstrated promising performance in identifying patients with ADRD.

Balagopalan et al. [24] modeled the acoustic part of speech using acoustic parameters of frequency and spectral domain, and speech fluency. To model the linguistic part, they used lexical and syntactic features (e.g., lexical richness, constituency parsing tree) and proportions of various information content units (as an indicator of memory impairment) used in the patients' verbal descriptions for the Cookie-Theft test. Different ML algorithms were trained on the combination of acoustic and linguistic features. Support Vector Machine (SVM) had the highest performance with an accuracy of 81.5. The authors did not report the importance of each part of the speech in achieving this performance.

Shah et al. [25] investigated the performance of alternate open-access repository of acoustic assessment algorithms, including AVEC-2013 [26], EMO_Large [27], and ComParE-2013 [28] for modeling the acoustic part of speech. For modeling the linguistic part, they used different natural language processing (NLP) techniques such as part of speech (POS) tagging, term frequency-inverse document frequency (TF-IDF) features, and n-grams to quantify syntactic and semantic parameters of the patient's langue. They also computed repetition and filled/non-filled pauses to quantify semantic disfluency. Different machine learning models were trained on acoustic and linguistic feature sets. SVM achieved the highest performance with an accuracy of 65 for acoustic features, an accuracy of 85 for linguistic features, and an accuracy of 83 for the joint combination of acoustic and linguistic features. The authors concluded that the reduction in accuracy might be due to the overfitting of ML models on the feature sets of training data.

Martinc et al. [29] extracted the psycholinguistic cues in the patient's speech using GeMAPS [30] acoustic feature set (see Section 3.5 "Component 4: modeling the patient's psycholinguistic expression" for details of GeMAPS). To model the acoustic part of speech, they used parameters of frequency and spectral domain. To model the linguistic part, they used different NLP techniques, such as TF-IDF, grammatical dependency, universal dependency, and the Doc2Vex text representation model. Authors also used readability features (e.g., Gunning Fog index) to measure complexity in the linguistic part of speech. Different ML algorithms were trained on the linguistic and acoustic feature sets. Logistic Regression had the highest performance, with an accuracy of 57.6 for the acoustic part, an accuracy of 75 for the linguistic component, and an accuracy of 77.08 for the joint combination of acoustic, linguistic, and psycholinguistic feature sets.

Chen et al. [31] extracted the psycholinguistic cues in the patient's speech using Linguistic Inquiry and Word Count [23] (LIWC) and GeMAPS [30] feature sets. Similar to Shah et al. [25], authors used ComParE-2013 [28] to model the acoustic part of speech. To model the linguistic part of speech, the authors used a transformer-based pretrained language model, Bidirectional Encoder Representations from Transformers (BERT). The BERT language model is capable of modeling the conceptual relationship between words and semantic disfluency in the patient's utterances [32]. Several ML algorithms were trained and tested on acoustic and linguistic feature sets (including psycholinguistic cues extracted using GeMAPS and LIWC). Logistic Regression had the highest performance with an accuracy of 71.69 for the acoustic part, an accuracy of 74.65 for the linguistic part (using only BERT), and an accuracy of 81.69 for the joint combination of acoustic, linguistic, and psycholinguistic feature sets.

Rohanian et al. [33] modeled the acoustic part of speech using COVAREP [34], an open-access repository of acoustic assessment algorithms. To model the linguist part of speech, authors used Glove [35], a word embedding technique that generates word embedding vectors from the patients' utterances. To model semantic disfluency, the authors used a deep-learning driven model of incremental detection of disfluency developed by Hough and Schlangen [36]. An ML architecture with Bi-long short-term memory (Bi-LSTM) was trained and evaluated on the linguistic and acoustic feature sets, which achieved an accuracy of 66.6 for the acoustic feature sets, an accuracy of 70.8 for the linguistic feature set, and an accuracy of 79.2 for the joint combination of acoustic and linguistic feature sets.

Pappagari et al. [37] investigated the performance of x-vectors [38], a transformer-based pretrained speech processing model, to process the acoustic part of the speech [38]. The x-vectors is a deep neural network that was originally developed for speaker type identification. It was trained on several datasets [38] for telephone conversation and microphone speech to map variable-length utterances to fixed-dimensional embeddings. For modeling the linguistic part, the authors used the BERT language model, which can model the conceptual relationship between words and semantic disfluency in the patient's utterances [32]. Authors trained and tested gradient boosting machine (GBM) on acoustic and linguistic embedding vectors. GBM achieved an accuracy of 58 for x-vectors, an accuracy of 72.92 for the BERT model, and an accuracy of 75 for the joint combination of x-vectors and the BERT model.

Pompili et al. [39] investigated the performance of x-vectors (see Pappagari et al. [37]) and i-vectors [40], transformer-based pretrained speech processing models, to process the acoustic part of the speech. The i-vectors is a DNN acoustic embedding method trained on VoxCeleb dataset [41], an annotated audio data for speaker identification collected from YouTube. Similar to previous studies [31,37], the authors used BERT contextual embedding to model the linguistics part of speech. Also, POS tagging was used to calculate the distribution of part of speech for each word in the sentences. An ML architecture with a Bi-LSTM network and SVM was trained and tested on acoustic and linguistic feature sets, which achieved an accuracy of 54.17 for x-vectors embeddings, an accuracy of 72.92 for BERT embedding, and an accuracy of 81.25 for the joint combination of acoustic and linguistic embeddings.

Zhu et al. [42] investigated the performance of different transformer-based pretrained speech processing models for generating deep acoustic embeddings from the acoustic part of speech. Specifically, they used MobileNet [43], YAMNet [44], and Speech BERT [45] for this task. MobileNet is a lightweight deep neural network built on depth-wise separable convolutions and trained on the ImageNet dataset [46]. YAMNet has the same architecture as MobileNet but with the difference that it was trained on a human-labeled YouTube audio dataset [47] for audio events. The Speech BERT architecture is similar to the Text BERT architecture (the contextual embedding model), except that the Speech BERT's input is the Mel spectrogram of speech data, and it was trained on the LibriSpeech dataset [48], a large set of audiobooks. For the linguistic part, they used BERT, BERT large, and Longformer. Longformer [49] is an extended version of the BERT language model that scales linearly with the sequence length of the text document to facilitate processing a document with thousands of tokens or longer. For unimodal (acoustic or linguistic part) transfer learning, Speech BERT and Longformer achieved the highest accuracy of 66.67 and 82.08, respectively. For multimodal transfer learning, the joint combination of Speech BERT and Longformer achieved accuracy = 82.9, which was the highest compared to the joint combination of other models.

Koo et al. [50] investigated the acoustic component of speech using VGGish [51], a transformer-based pretrained speech processing model for generating deep acoustic embeddings from the acoustic part of speech. VGGish was trained on a large manually-annotated YouTube videos [47]. The authors also employed the GeMAPS feature set to extract acoustic psycholinguistic cues from the acoustic part of speech. To model the linguistic part of speech, the authors used the XLNet [52] language model, an extended version of the BERT language model. Like BERT, XLNet is able to model the conceptual relationship between words and semantic disfluency in speech utterances. Additionally, the authors quantified semantic impairment in speech utterances using repetitiveness and lexical richness metrics. An ML architecture with CCN and Bi-LSTM networks was trained on acoustic and linguistic features, which achieved an accuracy of 72.92 for VGGish, an accuracy = 81.25 for XLNet, and an accuracy = 81.25 for the joint combination of all acoustic and linguistic features. The Authors concluded that achieving the same performance for only the linguistic feature set (unimodal) and the joint combination of linguistic and acoustic feature sets (multimodal) might be due to the overfitting of the multimodal model on the training dataset.

Syed et al. [53] investigated the performance of alternate open-access repository of acoustic assessment algorithms, including IS10-Paralinguistics feature-set [54] and COVAREP [34] for modeling the acoustic part of speech. The authors also used VGGish [51] for generating acoustic embeddings. For the linguistic part, they used different versions of the BERT language model (e.g., BERT large cased, DistilRoBERTa, and BioMed Roberta) for modeling the contextual relationship between words in the utterances. SVM was trained on the acoustic and linguistic parts of speech, which achieved an accuracy of 64.58 for the acoustic part, an accuracy of 85.42 for the linguistic part, and an accuracy = 79.17 for the joint combination of acoustic and linguistic parts.

Balagopalan et al. [55] utilized lexico-syntactic features to model the linguistic aspects of participant speech. These features were derived from speech-graph, constituency parsing tree, lexical richness, syntactic and semantic features based on picture description content. For the acoustic aspect of speech, the researchers employed Mel-frequency cepstral coefficients (MFCCs), fundamental frequency, and zero-crossing rate-related statistics. They trained various machine learning algorithms using a combination of these acoustic and linguistic features. The SVM model, with the 10 most informative features chosen through the ANOVA method, achieved an accuracy of 81.3 %. Additionally, the authors applied the pre-trained BERT model for the linguistic portion, resulting in a higher accuracy of 83.3 % compared to models trained on manually crafted linguistic and acoustic features.

Weirui Kong et al. [56] modeled the linguistic aspects of speech by employing syntactic and semantic features, as well as psycholinguistic characteristics of participants' language. They projected these features into an embedding space with a specific dimension using an encoder. For the acoustic parameters of speech, the researchers utilized the MFCCs technique. They combined the two modalities using a joint embedding method adapted from Kiros et al. (2014) and built logistic regression classifiers on these feature sets, achieving an accuracy of 70.8 %. In addition to this approach, the study also explored the performance of an end-to-end neural model using hierarchical attention networks (HAN), which allows avoiding any feature engineering. This model achieved an accuracy of 81.5 %. However, when incorporating participant age into the model, the classification performance improved, reaching an accuracy of 86.9 %.

Bertini et al. [57] trained an autoencoder with a multilayer perceptron architecture on the log mel spectrogram of participants' speech audio data. The rationale for employing an autoencoder was to generate a 128-dimensional vector that effectively captures the inherent audio features of Alzheimer's disease patients' vocal production. This code was then utilized to train a multilayer perceptron capable of identifying potential Alzheimer's disease subjects. To enhance the model's performance, the researchers utilized the SpecAugment suite, introduced by Park et al. [58], which transforms log mel spectrograms to increase the input data points. The model achieved an accuracy of 93.3 % (F-1 score = 88.5) on the augmented dataset, demonstrating a 26.4 % improvement compared to its performance on the non-augmented data (accuracy = 73.9, F-1 score = 62.1).

Roshanzamir et al. [59] examined the performance of pre-trained BERT, XLNet, and XLM models for contextual embedding of the linguistic aspect of speech. They found that the BERT large language model, when combined with a logistic regression classifier, achieved the highest accuracy score of 88.08 %. The researchers also investigated the impact of two text augmentation techniques on the performance of contextual word embedding methods: the similar word substitution augmentation method and the sentence removal augmentation method. However, they showed that these methods did not result in any significant overall improvements.

The ADscreen, the screening algorithm we developed, has some key differences from previous studies: (1) ADscreen has a component for modeling the patient's ability in phonetic motor planning, built on informative acoustic parameters associated with ADRD. Compared to previous studies that used transformer-based speech processing models, this component provides an insight into the impairment in acoustic parameters such as alternation in speech fluency. Additionally, we showed that the prediction power of this component (accuracy = 78.87) is higher than the prediction power of transformer-based speech processing models (x-vector, i-vectors, VGGish, MobileNet, YAMNet, Speech BERT) reported in previous studies. (2) To model the linguistic part of speech, we used both the transformer-based pretrained language model (the BERT language model) and domain-related features to quantify the semantic fluency, semantic impairment, and syntactic structure in the speech of the patient with ADRD. This component had a relatively high predictive performance (accuracy = 83.09) compared to

previous studies and can provide an insight into the impairment in the linguistic components of speech. (3) ADscreen has a component for extracting psycholinguistic cues from the patient's speech. This component is particularly important to extract vocal semantic psycholinguistic cues associated with neuropsychiatric symptoms for further evaluation. (4) Finally, ADscreen has an accuracy = 90.14 and F1-score = 89.55 for identifying patients with ADRD measured on the test dataset. This result indicates that ADscreen has a strong potential to be integrated into clinical workflow to raise clinicians' attention to the patient's cognitive status for further evaluation. Details of methodology for the development of ADscreen were provided in the methodology section.

3. Method

ADscreen is built on an analytic pipeline for modeling spontaneous speech by extracting acoustic and linguistic speech parameters. As a part of this pipeline, we used ML algorithms for modeling the relationships among variables for detecting patients with ADRD. Fig. 1 provides a schematic view of the ADscreen analytical components. Sections 3.2–3.5 provide a detailed description of the ADscreen components.

3.1. Data source

We used an audio-recorded speech dataset from the English Pitt DementiaBank, which included spontaneous speech samples from 237 participants during the "Cookie-Theft" picture description test. The "Cookie-Theft" test [60] is a drawing depicting two children stealing cookies behind their mother's back (see Appendix A). This test has been proven effective for assessing cognitive function in several studies [61,62].

Participants were instructed by healthcare providers to describe the drawing or create a story. From the 237 participants, the audio-recorded speech of 166 participants was organized into a training dataset, consisting of 87 ADRD patients (case group) and 79 non-cognitively impaired participants (control group). The remaining 71 participants' audio-recorded speech was organized into a test dataset, with 35 ADRD patients and 36 non-cognitively impaired participants.

The "Cookie-Theft" test study's inclusion criteria mandated participants be at least 44 years old. Table 1 displays demographic information for both development and test datasets, indicating that case group participants were slightly older and more likely to be women than control participants.

Participants also underwent a thorough neuropsychological assessment, including verbal tasks and Mini-Mental State Examination (MMSE). Eligible participants were required to have no history of major nervous system disorders and achieve an initial MMSE score above 10. MMSE scores, ranging from 0 to 30 points, are interpreted as follows: 24–30 (normal cognitive function), 18–23 (mild cognitive impairment), 10–17 (moderate cognitive impairment), and 0–9 (severe cognitive impairment). As per the MMSE scores in Table 1, case groups in both development (MMSE = 17.44 ± 5.33) and test datasets (MMSE = 18.86 ± 5.8) primarily experienced mild to moderate cognitive impairment.

ADRD participants exhibited lower mean word counts compared to the control group in both datasets (see Table 1), suggesting potential difficulties with language and communication. A detailed description of the cohort building process is available in Becker et al. [60], with further information provided in Appendix B of the manuscript.

3.2. Component 1: noise reduction

Individuals' speech that is audio-recorded in laboratory or real-word

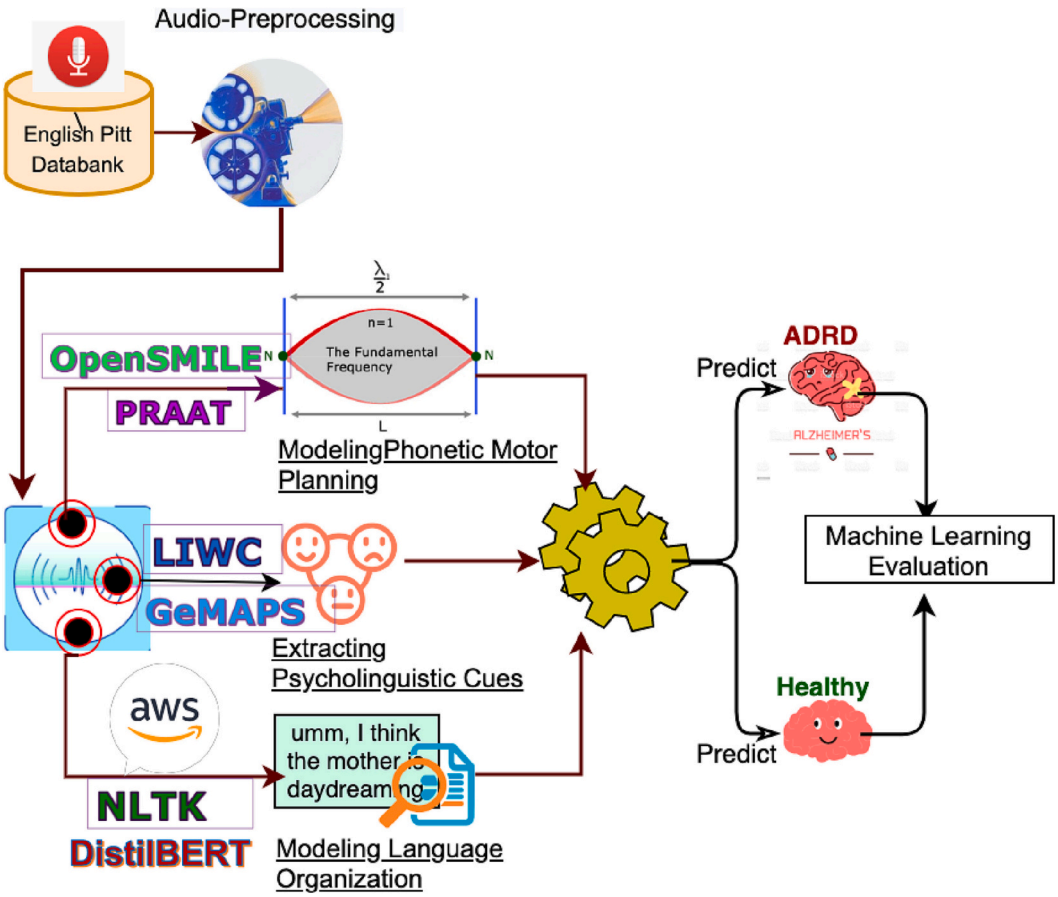


Fig. 1. A schematic view of the ADScreen pipeline.

Table 1
Characteristics of the cohort.

Development dataset		
Attributes	ADRD participant (case) $N = 87$	Non-cognitively impaired participants (control group) $N = 79$
Gender: F/M	58/29	52/27
Age Mean (Std):	69.72 ± 6.8	66.04 ± 6.25
MMSE score Mean (Std):	17.44 ± 5.33	28.99 ± 1.15
Words counts Mean (Std):	88.54 ± 47.92	113.54 ± 69.58

Test dataset		
Attributes	ADRD participant (case) $N = 35$	Non-cognitively impaired participants (control group) $N = 36$
Gender: F/M	21/14	23/13
Age Mean (Std):	68.51 ± 7.12	66.11 ± 6.53
MMSE score Mean (Std):	18.86 ± 5.8	28.91 ± 1.25
Words counts Mean (Std):	92 ± 57	109 ± 56

settings often includes noises that may affect the quality of downstream tasks built on that audio data. The environmental noises can stem from different sources, such as human conversation in the background or a thermal noise from a radio receiver. Overall, noises can be categorized into two main areas: stationary and nonstationary. In stationary noises, statistical parameters of the signal, such as intensity and spectrum shape, remain constant over time, while in nonstationary noises, these parameters change. Including noise in the audio-recorded data affects the accuracy of acoustic assessment algorithms, as well as the accuracy of machine learning classifiers built on computed acoustic feature sets. To eliminate noise in the speech data, we used the iZotope RX8 [63] toolkits. iZotope RX8 is software for noise reduction and noise removal built on deep learning neural network models to eliminate background noises. iZotope RX8 uses deep learning methods to identify and reduce both stationary and nonstationary noises, which showed a good performance in several sound enhancement studies [64,65].

3.3. Component 2: modeling phonetic motor planning (phonetic component)

Impairment in phonetic motor planning in patients with neurodegenerative disorders leads to poor pronunciation, along with alternation in phonological planning and speech rhythm [66–68]. We used acoustic parameters in five domains to model phonetic motor planning.

3.3.1. Alternation in speech fluency

Metrics for evaluation of speech fluency are among the most widely employed measures for assessing cognitive functioning. Different studies show that a subtle impairment in speech fluency indicates changes in the temporal functions of phonation. Individuals with ADRD manifest slower speech and articulation rates, including longer within-word disfluency (occurring due to prolongation of words' sounds), more and longer pauses, as well as inappropriate temporal distribution of pauses [69] in their speech unit. We used the following metrics to model speech fluency: articulation rate (number of phonemes per second without hesitation) [66], speech rate (number of phonemes per second with hesitation) [66], silent pauses (number of speechless intervals at the beginning of and between words) [70], and within-word disfluency (within-word silent pauses and sound prolongations) [71], and voicing probability [72] (indicating a percentage of unvoiced and voiced energy in a speech signal).

3.3.2. Alternation in frequency and spectral parameters of the voice

Fundamental frequency (F0) and resonant frequencies (Formant) [measured in hertz (Hz) or cycles per second (cps)] can provide information about an individual's ability to control vocal fold and tract in speech production [73], and phonological motor planning in turn. F0 is the vibratory rate of the vocal folds, and F0 range is an indicator of the phonation range that an individual can produce. The range is lower in individuals with ADRD compared with non-cognitively impaired individuals. Formant frequencies [74] (F1, F2, F3) are acoustic resonances of the vocal tract that occur due to changes in the position of vocal organs. The perceived quality of vowel pronunciation is the functional relationship among F1, F2, & F3 [75]. If the formants do not change fast enough or are not distinct enough, sounds may become harder for listeners to identify, leading to the perception of mumbling [76]. Patients with ADRD are unable to control high-format frequencies with average tonal oscillations over 500 Hz. This is particularly the case for F3 with tonal oscillations between 1500 Hz and 2500 Hz [77], resulting in the generation of unclear sound in speech.

Spectral parameters of speech are the analysis of discrete frequencies (spectrum frequencies) of the speech signal over desired frequency bins (e.g., 25 ms) [78]. Statistical analysis of the power (energy) of spectrum frequencies over a continuous range of speech can provide important clues about an individual's phonological planning. Energy variation among the frequency spectrum of speech signals can be computed using the Mel Frequency Cepstral Coefficients (MFCCs) metric. MFCC [79] is a widely used approach for the detection of phones (a sound representation of the phoneme) in speech recognition systems. Previous studies show that MFCC has good discrimination power in detecting patients with ADRD [80,81]. Other metrics computed using spectrum frequencies are long-term average spectrum (LTAS) and the spectral center of gravity [82], which captures the spectrum of the glottal source as well as resonant characteristics of the vocal tract [83]. These two metrics were linked to cognitive changes in previous studies [84]. The estimation of formant frequencies and bandwidths are commonly computed using linear predictor analysis via Linear Predictor Coefficients (LPCs). LSP is used to represent LPC due to properties such as smaller sensitivity to quantization noise that make them superior to direct quantization of LPCs.

3.3.3. Alternation in the intensity of the voice

Voice intensity is a function of mass, tension, and biochemical characteristics of the vocal folds, as well as a slight variation in an individual's ability in neural control. Impairment in phonological motor planning is associated with the individual's inability to control the intensity of speech. That inability can negatively affect the articulatory and prosodic aspects of speech, making the sound monotonous, dull, or even meaningless [66]. Mean and variability of intensity in an individual's speech correlates with the perception of vocal loudness and loudness variation. Vocal intensity is measured using metrics of sound pressure level (the Time-average Sound Level Definition), indicating the strength of vocal fold vibration. Variation in loudness can be measured using jitter and shimmer metrics, which are measured by the cycle-to-cycle variations of fundamental frequency and amplitude, respectively. These two metrics are widely used for describing pathological voice quality [85], particularly in patients with cognitive impairment [86]. Additionally, to quantify speech intensity, we computed the Hammarberg Index (the difference between the maximum energy in the

0…2 kHz band and the energy in the 2…5 kHz band) energy concentration [70] (average of spectral frequency content), and the ratio of the energy of the spectral harmonic peak at the second and third formant´s center frequency to the energy of the spectral peak at F0 in voiced regions.

3.3.4. Alternation in the voice quality

The patient's ability in phonetic motor planning affects how the listener perceives the quality of their voice. Despite the increase in voice noise in elderly individuals as a part of normal aging, older individuals with ADRD "lose" part of their vocal noise characteristics and have more harmonic and fluty voices than they did when they were younger [77]. The presence of noise can be measured using the Harmonic to Noise Ratio (HNR) [73]. HNR is the relationship between harmonic sound (periodic component) and noise in the vocal signal (aperiodic component). Voice-breaking [87] is another indicator in individuals' speech of impairment in phonological planning and difficulty in vocal cord execution. This break is a sudden gap in sound that accrues when the thyroarytenoid muscles suddenly decrease their activity, and the cricothyroid muscles begin to function [88]. It is calculated as the frequency of breaks during an utterance (a continuous block of speech without interruption). We also calculated the Voice Quality Index (AVQI) [89], which consists of a weighted combination of time-frequency and frequency-domain metrics that was originally developed to measure the severity of dysphonia.

3.3.5. Alternation in the rhythmic structure of the voice

Studies showed that alternation in rhythmic structure throughout the evolution from healthy aging to ADRD follows a steady pattern parallel to cognitive decline. This disorder can be related to impairment in phonetic-motor planning, leading to poor pronunciation and alteration in syllabic rhythm. Rhythm is defined as an isochronous recurrence of some type of speech unit [90], such as syllabic duration, intensity, and voice breaks. Impairment in phonological planning and progression to ADRD implies conversion to slower speech, less intensive speech, monotone and tremulous voice, and continuous interruptions and breaks. As a result, the speech signal becomes progressively degraded, and the speech itself loses clarity [86,91], creating the impression of choked and hesitant speech sounds. Slowness in speech is measured using metrics of speech fluency, prolonged syllable intervals, and a higher variation in the duration between two successive syllabic intervals (Pairwise Variability Index). Monotonous voice is the result of a reduction in the variation in the breadth of vowel sounds, which can be measured using the Shimmer metric (a measure of the maximum amplitude between two consecutive periods of vibration of the glottis). The greater shimmer in speech production is a characteristic of patients with cognitive impairment [91], which indicates greater instability of amplitude in the sound.

3.4. Component 3: modeling the patient ability in semantic and syntactic levels of language organization (linguistic component)

Language impairment in patients with cognitive impairment is associated with aphasia-like symptoms and memory deficit symptoms, characterized by less dense and inaccurate speech planning [92,93], difficulties in finding words [66,94], simplified syntaxis and semantics [66,95,96], and circumlocution [66,97]. These symptoms can result in communication errors and lower coherence in speech. To model the semantic level of language organization, we used metrics of semantic disfluency and lexical richness. To model the syntactic structure of the language, we used metrics measuring the complexity and components of sentence structure.

3.4.1. Modeling semantic disfluency in speech

Semantic disfluency in speech often characterized by repeated words (repetitiveness) or inappropriate pausing behavior/hesitation during their speech.

Repetitiveness: Previous studies found that patients with ADRD repeat words and phrases more frequently in their verbal responses to the Cookie-Theft test compared with non-cognitively impaired participants [98,99]. To identify repetitiveness in the patients' verbal responses for the Cookie-Theft test, we used two methods: (1) computing the similarity score between clauses in the patient's verbal response: Using bag-of-words, we computed cosine distance between clauses to obtain the similarity score. To improve the accuracy of this computation, we removed stop-words and common occurrences of some words, such as "he" or "is" in clauses, "he is looking at mom," and "he is falling off the stool." A similarity score of "0" indicates that two clauses are identical. The proportion of clauses with a score of "0" and average and standard deviation of computed similarity scores in the patient's verbal response were taken into account. (2) Identifying consecutive duplicate words or phrases in clauses: We used the Regular Expression library of Python programming language to identify duplicate words/terms. The proportion of duplicated words/phrases with reference to the total number of words/phrases in the patient's verbal response was computed.

Pausing behavior: Pauses are the smallest syntactic units that occur at a specific moment in speech production where a form of a content word (e.g., verb/noun) or a lexical concept has to be retrieved from the memory and inserted to complete a clause or an utterance [100,101]. At this time, the lexical and semantic memory needs to be active. By contrast, at the initial boundary of an utterance or clause, a complete semantic and syntactic configuration should be constructed that requires full thought. Such structure does not retrieve from the memory but is creatively produced on occasion. Therefore, identifying disfluencies at a specific syntactic location can show a link between pausing and semantic disfluency/thinking expressed in speech. We modeled the pausing behavior according to four metrics introduced by Lofgren et al. [102]: whether pauses occurred (1) within-clauses, (2) clause-initial, (3) utterance-initial, or (4) whether the pause preceded nouns, verbs, or adjective/adverbs when occurring within-clauses.

To compute these linguistic parameters, we used the Amazon Web Service (AWS) General Transcribe (GT) system to transcribe the audio-recorded speech in the study sample. To calculate pausing metrics, we used Spacy, a module of Python programming language, and the "word timing" information provided by AWS-GT.

3.4.2. Measuring lexical richness (lexical diversity)

Patients with impairment in semantic memory have a less diverse vocabulary and are biased toward and higher frequent words [103,104]. Additionally, they may have "difficulty accessing more diverse nouns and verbs." [105] Previous studies showed that metrics for computing lexical richness are applicable for quantifying impairment in semantic memory in patients with ADRD [92,93]. We computed the lexical richness (content density) score using five following metrics: 1) type-token ratio (TTR), including root type-token ratio (RTTR), corrected type-token ratio (CTTR), moving average type-token ratio (MATTR), [106] index is the total number of unique words divided by the total number of words for each successive window of fixed length; 2) Brunet's Index is the variation in the type of words marked by a part-of-speech tagging tool in a sentence with reference to the total number of words in a sentence [107]; 3) Honore's Index measures the proportion of words used only once with reference to the total number of words [108]; 4) hypergeometric distribution index is a discrete probability distribution that computes the probability of randomly drawing the same word after a number of attempts without replacement [109,110]; and 5) the Measure of Textual Lexical Diversity (MTLD) which reflects the average number of words in a sequence of words for which a certain TTR is maintained [109].

3.4.3. Modeling the syntactic structure

We used part-of-speech tagging (POS) for this task. POS is a method for identifying sentences and types of words (adjectives, adverbs,

articles, nouns, numbers, and verbs) in sentences. We used the outcome of the POS method to compute metrics indicating syntactic complexity, syntactic components, and dependency among syntactic components of sentences using a set of syntactical features. The features include part-of-speech rate, frequency-of-use tagging, action verbs rate, relative pronouns, and negative adverbs rate. Previous studies showed that these metrics are associated with mental disorders and cognitive impairment, and thus they can point toward syntactical indicators of ADRD [76,111].

3.4.4. Modeling disfluency in the patient speech using the BERT language model

Previous studies showed that BERT and its extended versions (e.g., DistilBERT, XLNet) has a knowledge of the structure of disfluency [32,112]. This language model processes disfluency by selectively attending to different parts of the disfluency at different intensities using the key mechanism of attention. This mechanism allows BERT to differentiate between the contextual embeddings of disfluent sentences and their fluent counterparts (see more details in Appendix C).

We modeled the conceptual relationship between words in the patient's utterances (for the Cookie-Theft test) using the BERT language model and its extended versions, DistilBERT, DistilRoBERTa, and XLNet. Next, we evaluated the performance of each language model in detecting patients with ADRD. Appendix D provides more details about this evaluation. DistilBERT had the highest performance with an accuracy = 83.09 and F1-score = 82.35. Therefore, we used this model along with other linguistic domain-related features (explained above) to model the patient's ability in semantic and syntactic levels of language organization.

DistilBERT is a BERT-based small, fast, and light pretrained transformer model. It uses 40 % fewer parameters than the original BERT-base model and runs 60 % faster, and keeps >95 % of BERT's performance, as measured in this study [113]. We only use the cased version of this model with the following hyperparameters: 6-layer, 768-hidden, 12-heads, 65 M parameters [114]. The model implementation is available here [114].

3.5. Component 4: modeling the patient's vocal and semantic psycholinguistic cues

Prior studies have revealed that vocal and semantic psycholinguistic cues present in a patient's language can be indicative of Alzheimer's disease and other related dementias, as these conditions impact a person's cognitive processes [115,116]. By analyzing these cues, healthcare professionals can detect early signs of cognitive decline, track the progress of the disease and treatment. We modeled the vocal psycholinguistic cues using Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [30] and the semantic psycholinguistic cures using Linguistic Inquiry and Word Count (LIWC) 2015 [23].

3.5.1. Extracting vocal psycholinguistic cues

Vocal psycholinguistic cues conveyed in an individual's voice have been empirically documented using a wide range of acoustic parameters that reflect subglottal pressure, vocal tract airflow, and vocal fold vibration [117,118] (e.g., frequency and spectral parameters and parameters measuring intensity and speech fluency). In this study, we used a minimalistic standard parameter set called the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [30] for modeling vocal psycholinguistic cues in participants' speech. Compared with large brute-forced feature sets (e.g., ComParE with 6373 acoustic parameters), GeMAPS has a better generalization capability to an unseen dataset [119]. The GeMAPS parameters are presented in four domains: frequency-related parameters, energy/amplitude-related parameters, spectral (balance/shape/dynamics) parameters, and voice quality parameters. The details of parameters for each domain were explained in the original article [30]. The effectiveness of GeMAPS in modeling the vocal psycholinguistic cues particularly "arousal" and "valence" and its generalization capabilities is clear from previous studies [30,120,121]. Please see Appendix E for more details about the performance of GeMAPS in identifying vocal psycholinguistic cues. We also conducted a statistical association analysis using *t*-test method to demonstrate the relationship between GeMAPS parameters and ADRD within the study's sample population. The findings can be found in Appendix E. According to the findings, out of 88 total acoustic parameters, 57 parameters were significantly associated with ADRD (*P*-value <0.05).

3.5.2. Extracting semantic psycholinguistic cues

In this study, we used Linguistic Inquiry and Word Count (LIWC) 2015 [23] to extract semantic psycholinguistic cues associated with ADRD. LIWC is a manually curated lexical-based natural language processing tool developed by experts in the psychology of language. It contains a large selection of commonly used words and terms organized into 11 top-level categories, including linguistic structure, affective processes, social processes, cognitive processes, perceptual processes, biological processes, drives, relativity, informal language, personal concerns, and time orientation. LIWC has been used in several studies to characterize patients' and clinicians' language [122,123]. For examples words indicating tentativeness, such as "maybe" and "guess/think" belong to the category cognitive process, words indicating conjunctions such as "but" and "whereas" belongs to category of linguistic structure, and words indicating "unease" and "worried" belong to the category of "affective process." In the area of healthcare, the reliability and validity of LIWC in detecting semantic psycholinguistic cues associated with mental and neurological disorders have been verified in several studies [95,124–127]. O'Dea et al. [124] showed that features of the LIWC's linguistic domain, including "tentativeness" and "non-fluencies," were significantly correlated with symptoms of depression and anxiety [124]. Also, Asgari et al. [95] found that linguistic markers from domains of psychological process and linguistic features were associated with the presence of cognitive impairment.

Please see Appendix F for more details about the performance of LIWC in identifying semantic psycholinguistic cues. We also conducted a statistical association analysis using *t*-test method to demonstrate the relationship between LIWC parameters and ADRD within the study's sample population. The findings can be found in Appendix F. Based on these findings, 79 out of the 93 linguistic parameters analyzed exhibited a significant association with ADRD, as evidenced by a *P*-value <0.05.

3.6. Component five: building and evaluating machine learning models

Fig. 2 provides a schematic overview of the phases we used for building and evaluation of machine learning models.

3.6.1. Feature generation phase: processing the acoustic and linguistic parts of the audio-recorded data

To model the patient phonetic motor planning, we used the implementation of acoustic assessment algorithms available in OpenSMILE [128] and PRAAT Vocal Toolkits [129] for five acoustic domains, speech fluency, frequency and spectral parameters, intensity, quality, and rhythmic structure of the voice. Both Toolkits are open-source platforms, including the efficient implementation of the acoustic parameters. We computed the parameters using a frame size of 25 ms. [130] Mean, standard deviation (std), feature quartile, interquartile range, and skewness were computed. We also processed the acoustic part of speech using YAMNet, a transformer-based pretrained speech processing model. However, because of the low performance of this model in detecting patients with ADRD, we did not incorporate the acoustic embedding vectors generated by this model into the feature sets used for building the screening algorithm. See Appendix G for more information about YAMNet and its performance in detecting patients with ADRD.

For the implementation of acoustic parameters of GeMAPS, we also used the acoustic assessment algorithms available in OpenSMILE [128].

For the implementation of the linguistic part of the speech, we first

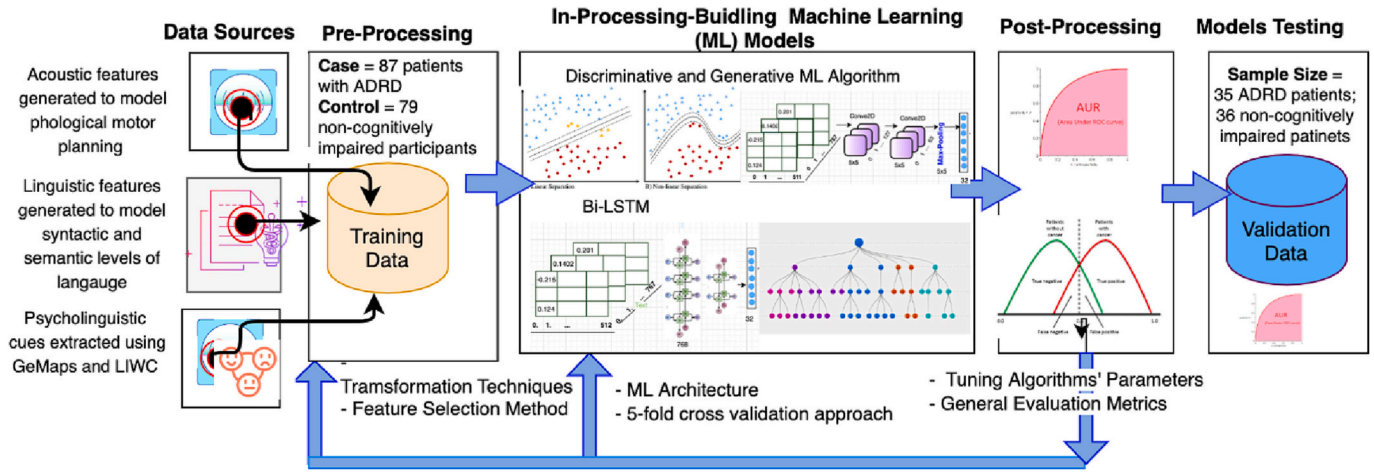


Fig. 2. A schematic view used for building and evaluating machine learning models.

transcribed all the audio-recorded data to text using Amazon Web Service (AWS) General Transcribe (GT). In our previous study [131], we computed the Word Error Rate (WER) for AWS-GT for the transcription of patient-spoken language as 0.26, which was higher than other transcription systems, including AWS-Medical Transcribe (WER = 0.56) and Wave2Vec (WER = 0.98) [132]. Wave2Vec [133] is an open-source automatic transcription system developed by Facebook company. AWS-GT transcription includes the transcription of the spoken word and the timing (start time and end time) associated with it. We applied NLTK and Spacy toolkits of Python programming language to the transcribed data to compute metrics related to repetitiveness, pausing behavior, lexical richness, and syntactic structure of patients' verbal descriptions (see Section 3.4. "Component 3" for details).

We also generated word embedding vectors using the DistilBERT language model with a size of 768 × 512 to model the conceptual relationship between words in the patients' speech. 768 is the size of the hidden layer, and 512 is the max sequence length (see details in the Section 3.4.4. "Modeling disfluency in the patient speech using BERT" and Appendix D). The embedding vectors were incorporated into the acoustic and linguistic feature sets for building the screening algorithm (see Section 3.6.3 "In processing phase" for details).

To extract the semantic psycholinguistic cures, we used Linguistic Inquiry and Word Count (LIWC) version 2015 [23]. See section "Component 4: modeling the patient's vocal and semantic psycholinguistic cues" for details. Table 2 provides the list of generated features for three components (phonetic motor planning, semantic and syntactic levels of language organization, vocal and semantic psycholinguistic cues) in compact form.

3.6.2. *Pre-processing phase*

All variables (acoustic and linguistic features) were centered and scaled using standard scaling, which standardizes the features by removing the mean and by scaling to unit variance. For feature selection, we used Joint Mutual Information Maximization (JMIM) method. JMIM was recently introduced as an effective feature selection method, particularly in high dimensional, small sample size datasets. It is from the family of joint mutual information (JMI)-based methods. In information theory, the mutual information (MI) of two random variables is the amount of information obtained about one random variable (X) by observing the other random variable (Y). This can be quantified as the reduction in entropy of one random variable (Y) given another variable (X) as follows:

$$I(X;Y) = E(Y) - E(Y|X)$$

where $E(Y)$ is entropy of Y. For any discrete variable such Y = (y_1,y_2, …,

Table 2
List of acoustic and linguistic features used for development of models in compact form.

Phonetic more planning	Speech fluency	Articulation rate, speech rate, silent pauses, voicing probability, within-word disfluency
	Frequency and spectral parameters	Fundamental frequency, jitter, pitch, formant frequencies, MFCCs, LPCs, LTAS
	Voice intensity	Voice intensity, loudness, Hammarberg Index, energy concentration, ratio of the energy of the spectral harmonic peak at the second and third formant's center frequency to the energy of the spectral peak at F0 in voiced regions.
	Voice quality	Harmonic to noise ratio, voice-breaking, Voice Quality Index
	Rhythmic structure of the voice	Pairwise Variability Index, prolonged syllable intervals, Shimmer metric
	Acoustic embedding models	YAMNet model
Semantic and syntactic levels of language organization	Repetitiveness	Similarity score between clauses, consecutive duplicate words or phrases in clauses
	Pausing behavior	Pauses occurred (1) within-clauses, (2) clause-initial, (3) utterance-initial, or (4) whether the pause preceded nouns, verbs, or adjective/adverbs when occurring within-clauses.
	Lexical richness	Type-token ratio (TTR) [root type-token ratio (RTTR), corrected type-token ratio (CTTR), moving average type-token ratio (MATTR)], Brunet's Index, Honore's Index, hypergeometric distribution index, textual lexical diversity (MTLD)
	Syntactic structure	Part-of-speech rate, frequency-of-use tagging, action verbs rate, relative pronouns, and negative adverbs rate
	Contextual word embedding features	DistilBERT, BERT base-cased, XLNET, DistilRoBERTa
Vocal and semantic psycholinguistic cues	Parameters defined in eGMAPS and LIWC	

y_N), E(Y) is defined as:

$$E(Y) = -\sum_{i=1}^{N} p(y_i) log(p(y_i))$$

$E(Y|X)$ is the conditional entropy. The conditional entropy is the amount of uncertainty left in variable Y when a variable X is introduced. So, it is less than or equal to the entropy of both variables. Conditional entropy is formulated as follows:

$$E(Y|X) = -\sum_{j=1}^{M} \sum_{i=1}^{N} p(x_i, y_j) log\left(p\left(y_j|x_i\right)\right)$$

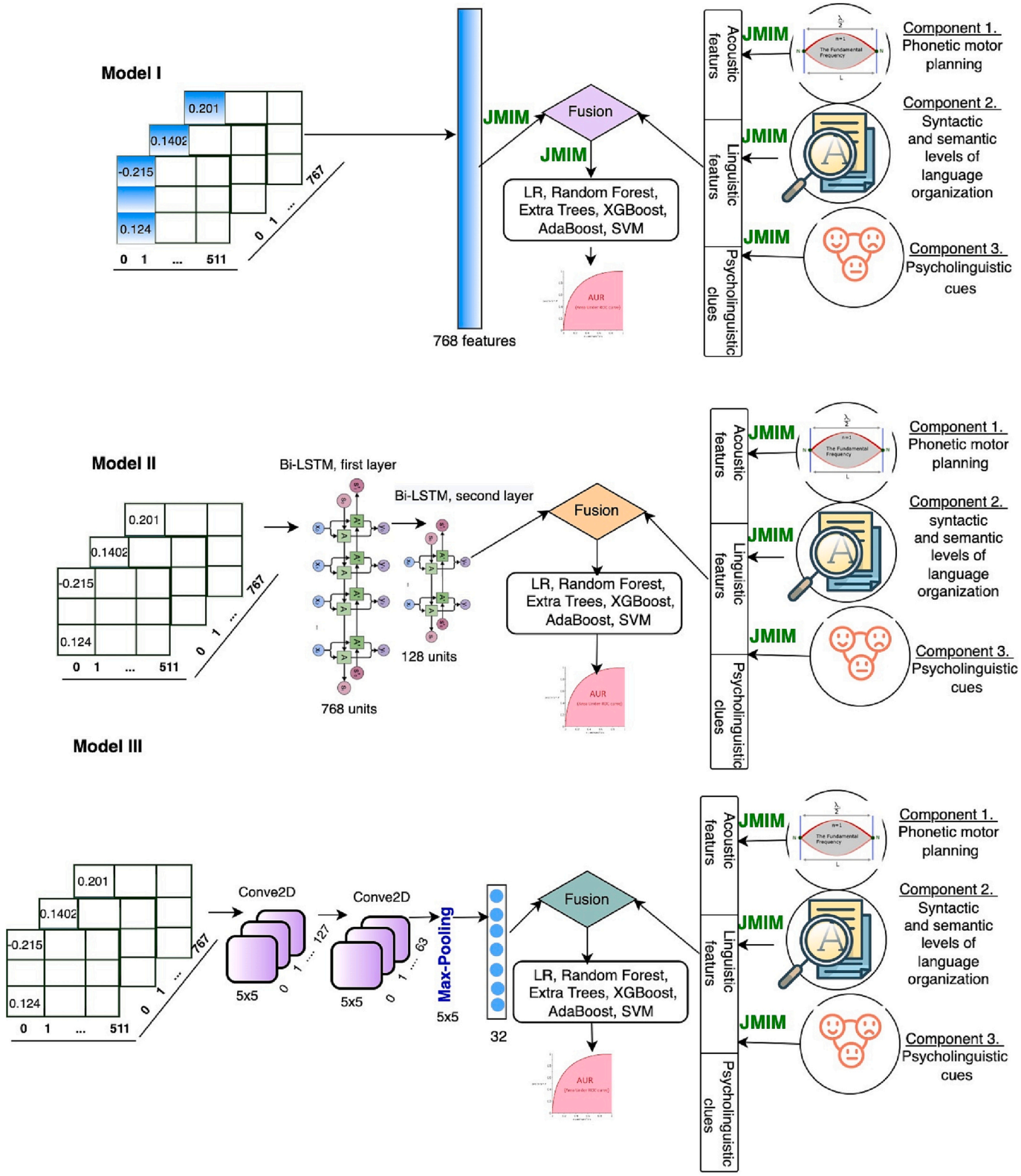


Fig. 3. A schematic view of machine learning architectures used to infuse acoustic and linguistic parameters of phonetic, linguistic, and emotion components with the word embeddings generated by the DistilBERT language model.

The information gain method is the simplest feature selection method built on MI. It assumes features are conditionally independent of one another. To tackle this problem, the JMIM method attempts to take into account the potential dependency among feature set F = {f_1,f_2, ……,f_N} by selecting a subset of feature S = {s_1,s_2, ……,s_k} with the dimension of K where $K(number\ of\ features\ in\ S) < N(number\ of\ features\ in\ F)$ and $S \subseteq F$, while minimizing the redundancy of information among selected features and maximizing the joint mutual information among the subset of S and outcome class label Y. Mathematically,

$$f_{JMIM} = arg\ \max_{f_i \in F-S}\left(\min_{f_s \in S}(I(f_i,f_s;Y))\right)$$

The most important advantage of the JMIM method compared to feature selection methods from the class of wrapper and embedding is its generalizability of selected features for unseen datasets, which can improve the stability and generalizability of the ML models on an unseen dataset.

3.6.3. *In processing phase: machine learning (ML) architecture*

We used three different ML architectures to model the relationship between speech parameters (predictors variables) and the presence of ADRD (outcome variable). Fig. 2 shows a schematic view of these three architectures. In the first architecture (Fig. 3. Model I), we first trained the DistilBERT model on the training dataset to extract the contextual word embeddings from the last layer of the model. In this way, we obtained a single 768-dimensional feature vector for each patient's description for the Cookie-Theft test. This approach was used by Pompili et al. [39] and had a promising result for identifying patients with ADRD. Next, we applied the JMIM method to the 768-dimensional feature vector to extract the important features associated with ADRD for combination with other linguistic and acoustic feature sets.

To select important acoustic, linguistic, and psycholinguistic features from modeling the phonetic motor planning (phonetic component), semantic and syntactic levels of language organization (linguistic component) and vocal and sematic psycholinguistic cues (psycholinguistic component), we independently applied the JMIM method to the acoustic/linguist features computed for each component. Next, we fused selected features with important contextual features obtained from the DistilBERT. For the classification task, we tested the performance of different ML classifiers including, Logistic Regression (as a baseline), Random Forest and Extremely Randomized Trees [134] (Extra Trees), two popular algorithms from the family of bootstrap aggregation (bagging) ensemble decision tree algorithms, Adaptive Boosting [135] (AdaBoost) and Extreme Gradient Boosting (XGBoost) [136], two popular algorithms from the family of gradient boosting ensemble decision tree algorithms, and Support Vector Machine (SVM) from the family of the general category of kernel methods [137]. See more information about these algorithms in Appendix H. Fig. 3. Model I provides a schematic view of the architecture of Model I that was applied to the test dataset.

The second architecture (Fig. 3. Model II) is composed of two Bi-Long Short-Term Memory (LSTM) that were trained on contextual word embeddings vectors computed using DistilBERT. Bi-LSTM is a particular type of recurrent neural network (RNN) through which the relationships between the longer input and output variables are modeled. In a Bi-LSTM network, the given input variables are utilized twice for training (i.e., first from left to right, and next from right to left). We used Bi-LSTM rather than the LSTM network because previous studies showed that it has a higher performance in modeling sequential data [138]. The output of the second Bi-LSTM layer was passed to a fully connected (FC) layer. The outcome of the FC layer was a feature set for fusing with acoustic and linguistic parameters. To determine the number of Bi-LSTM layers and the size of the FC layer, we computed the overall performance of this architecture on the training dataset with different number of layers (1,2,3,4 layers) for Bi-LSTM and different number of neurons for the FC layer (120, 64, and 32). This architecture was followed by a SoftMax layer for training. Two Bi-LSTM layers with one FC layer with 32 neurons had the highest performance. Next, we fused the outcome of the FC layer with important acoustic/linguist features computed for each component (phonetic, linguistic, psycholinguistic) using the JMIM method that were passed to classification algorithms (LR, Random Forest, Extra Trees, XGBoost, AdaBoost, and SVM) for separating patients with ADRD from the participants without ADRD. Fig. 3. Model II provides a schematic view of the architecture of Model II that was applied to the test dataset.

For the third architecture (Fig. 3. Model III), we investigated the performance of the Convolutional Neural Network (CNN) in modeling the relationship between embedding vectors of DistilBERT and the output variable. CNN is a type of deep learning model that has a grid pattern (e.g., two dimensional matrices) and "is designed to automatically and adaptively learn special to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns" [139]. A typical architecture of CNN is composed of the stack of several convolution layers and a pooling layer linked to one or more FC layers. The convolutional layer is a specialized type of linear operation used for feature extraction, and the pooling layer uses a down-sampling operation to reduce the dimensionality of the extracted feature [139]. In the second architecture, we used two CNN layers, a max pooling layer, and one FC layer (with the same functionality of the FC layer in the ML Model II). To determine the number of convolution layers, a function for pooling operation, and the number of neurons for the FC layer, we computed the overall performance of this architecture on the training dataset with different size (1,2,3 layers) for the convolution layer, average and max functions for the pooling operation, and a different number of neurons for the FC layer (120, 64, and 32). This architecture was followed by a SoftMax layer for training. A model with two convolution layers, max function for the pooling layer, and 32 neurons for the FC layer had the highest performance on the training dataset. Similar to Model II, we fused the outcome of the FC layer with important acoustic/linguist features that were passed to classification algorithms for separating patients with ADRD from the participants without ADRD. Fig. 3. Model III provides a schematic view of the architecture of Model III that was applied to the test dataset.

3.6.4. *Post-processing phase: training and evaluating performance of machine learning architecture*

All the processes for selecting important acoustic and linguist parameters (for phonetic, linguistic, and psycholinguistic components) were conducted using the JMIM method. We also used the training sample for fine tuning the parameters of machine learning algorithms used in ML Models I, II, and III. To do this, the training sample was partitioned into five equal subsets ("folds"), with the random partitioning stratified by ADRD patients to ensure that their distribution was approximately the same in all the folds. Since the algorithms (LR, Random Forest, Extra Trees, XGBoost, AdaBoost, and SVM) require choosing tuning parameters for optimal performance, a grid search was implemented within the five-fold cross-validation to select the best parameters. See Appendix I for parameters tuned for each classification algorithm. Specifically, for each combination of parameters defined over a grid, the algorithms were trained with those parameters using five folds and assessed the performance of the model on the 5th or hold-out fold. This procedure was repeated five times for each parameter set in the grid, until each fold has been used for testing.

Then, the optimal parameters for each algorithm were selected based on the Area Under Curve-Receiver Operating Characteristics (AUC-ROC) over the five repeated runs. For each algorithm, the performance of the final model was obtained by retraining the algorithm on the entire training dataset using the optimal parameters. We then used the test dataset to evaluate the performance of the final model on the validation cohort.

3.6.5. ML models testing

Goodness of fit of a classifier is evaluated using standard (general) performance metrics, including AUC-ROC, Cumulative Gains curve, Gini Score, Sensitivity, Specificity, Positive Predictive Value (PPV), and F-score (a harmonic mean between Precision and Recall). AUC-ROC is the tradeoff between the True Positive Rate (TPR or Sensitivity) and the False Positive Rate (FPR or 1-Specificity) and has the advantage of being invariant to the class distribution. For each ML model, we reported the standard deviation of the performance metrics over the five-fold cross-validations. For the best performing ML model, we computed the optimal sensitivity and specificity using the geometric mean (G-mean) metric and visualized it on the AUC-ROC curve.

4. Results

4.1. Most informative acoustic and linguistic features for screening patients with ADRD

The results of applying JMIM for extracting the most informative acoustic and linguistic features from each component are presented in Figs. 4, 5, and 6. Fig. 4 presents the top 20 informative acoustic features for modeling impairment in phonological motor planning. The acoustic parameters of all five domains, speech fluency, frequency and spectral parameters, voice intensity, the rhythmic structure of the voice, and voice quality are associated with the risk of ADRD. Acoustic parameters indicating alternation in frequency/spectral parameters (line spectrum pair [LSP] frequencies, MFCC, long-term average spectrum, F0) and speech fluency (pause rate, voice segments, voice probability, unvoiced segments) have higher presentation among the top 20 acoustic parameters than other domains. Parameters presenting voice intensity including: the rising slope of loudness, loudness peak, the energy ratio of the spectral harmonic peak of F0 and F1, and the energy ratio of the spectral harmonic peak of F0 and F3 are more informative than other parameters in this domain for screening patients with ADRD. Alternation in jitter and shimmer as indicators of alternation in the rhythmic structure of the voice are also among the most informative acoustic parameters. Harmonic to noise ratio [HNR] as an indicator of voice quality is also an important parameter; however, it is less informative compared to frequency/spectral and speech fluency parameters (see Fig. 4).

Fig. 5 presents the importance of 20 linguistic features. Features of lexical richness (root and corrected of type-token ratio, Honor's and Brunet's Index) are among the top important features for modeling the patient's verbal language. Features extracted using POS tagging (e.g., pronouns, part-of-speech rate) are also informative indicators of the ADRD risk, as expected. Features presenting pausing behavior (e.g., total average silence duration per word within clauses, total average silence duration in initial clauses) and repetitiveness (e.g., proportion of clauses with a similarity score of zero) are also among the top important informative features in detecting patients with ADRD. We did not combine the word embedding features extracted from DistilBERT with the linguistic domain features for presentation in Fig. 5 as they are not explainable.

Fig. 6A presents the top 20 informative features for vocal psycholinguistic cues extracted using GeMAPS's acoustic features. Features from four domains of GeMAPS, including frequency domain (std [standard deviation] of F2 bandwidth), spectral domain (mean spectral slope 0–500 Hz, mean spectral slope 500–1500 Hz), voice quality domain (HNR), and energy/amplitude domain (SD [standard deviation] of rising slope of loudness) are shown among the top five features. Other top 15 features are mostly related to the frequency and energy/amplitude domains as expected. Fig. 6B presents the top 20 informative linguistic features for semantic psycholinguistic cues extracted using LIWC. Features from linguistic dimension are among the most informative features for expression of semantic psycholinguistic cues in patients with ADRD. As expected, features from psychological process are also informative for modeling psycholinguistic cues in those patients. Other domains, relativity, personal concerns, and spoken language were not presented as important features for modeling the psycholinguistic cues in patients with ADRD.

4.2. Performance of ML models

We evaluated the performance of three ML Models (I, II, III) for screening patients with ADRD. These three ML Models are different in terms of processing the contextual word embedding vectors obtained from the DistilBERT language model. In Model I, we extracted a 768-dimensional feature vector from the last contextual word embedding vectors of DistilBERT and fused it with the informative features extracted from the three components of phonetic, linguistic, and

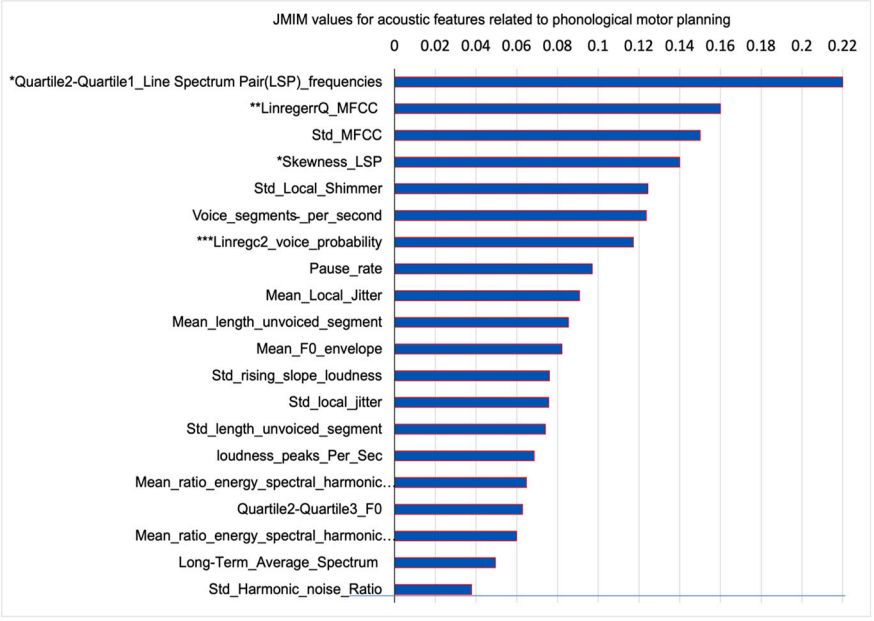


Fig. 4. JMIM values of acoustic features used for modeling the impairment in phonological motor planning.

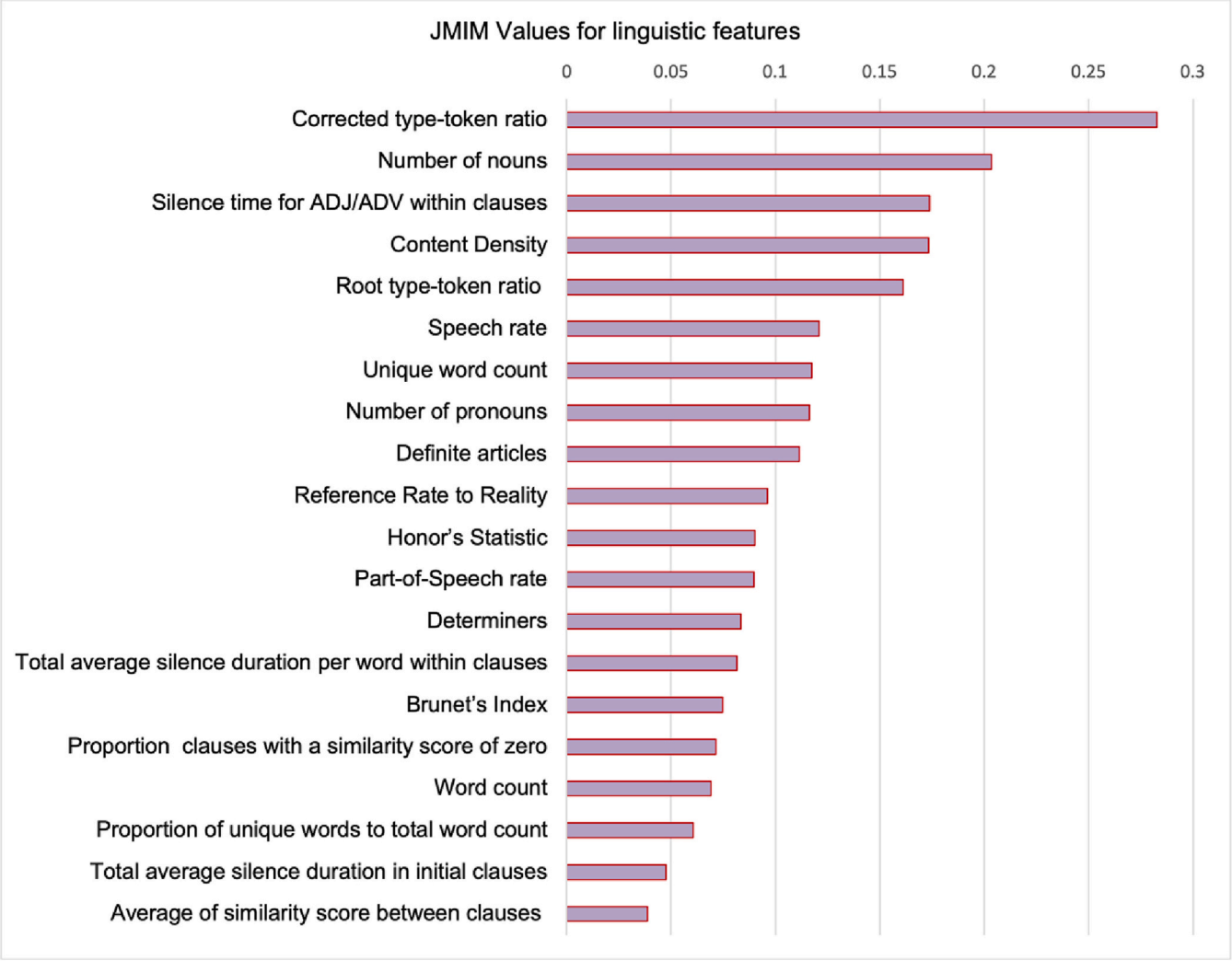


Fig. 5. JMIM values for linguistic features used for modeling semantic and syntactic levels of language organization.

psycholinguistic components. This model outperformed ML Model II and III in which the Bi-LSTM and CNN models were trained on the word embedding vectors (512 × 768) of DistilBERT, and their outcomes were fused with informative features of the three components (see Fig. 3 for the architecture of these three models). Table 3 shows the performance of ML model I for different ML classifiers for both development and test datasets. SVM with RBF (Radial Basis Function) Kernel had the highest performance with AUC-ROC = 92.53 ± 3.34, accuracy = 85.04 ± 3.41, and F1-score = 84.64 ± 3.58 for the training dataset and AUC-ROC = 93.89, accuracy = 90.14, and F1-score = 89.55 for the test dataset. Ensemble decision trees had almost the same performance for AUC-ROC, F-score, and accuracy on the training dataset. However, the performance of gradient boosting trees (AdaBoost and XGBoost) on the test dataset was less than bagging trees (Random Forest and Extra Trees). This might indicate that the boosting trees overfitted the training dataset. Logistic Regression had the lowest performance for both training and test dataset. This might indicate that the relationship between input predictors and the outcome variable is not linear.

The performance of ML model II on the training dataset using SVM algorithm was AUC-ROC = 89.03 ± 5.68, accuracy = 80.76 ± 7.15, and F1-score = 82.97 ± 6.56; while the performance of ML III (using SVM algorithm) was AUC-ROC = 85.77 ± 6.75, accuracy = 77.73 ± 5.48, and F1-score = 78.65 ± 5.77 (see more results in Appendix J, Tables 5 and 6 for ML Models II and III). These results were less than the ML Model I, indicating that modeling the high-dimensional word embedding vectors using the Bi-LSTM and the CNN network on the small training sample of this study does not improve the prediction power of the SVM algorithm. Additionally, Models II and III had higher standard deviations compared to ML Model I (measured using the five-fold cross validation on the training data), indicating the lower stability of the Models. The performance of Model II on the test data was AUC-ROC = 94.44, accuracy = 87.32, and F1-score = 87.67; while the performance of Model III was AUC-ROC = 92.06, accuracy = 88.73, and F1-score = 88.57. See Table 4 for the performance of Model II and Model III on the test dataset.

The accuracy and F1-score of both models were less than Model I, but the AUC-ROC of Model II was slightly better than Model I. We decided to choose the Model I as our final model for building the screening algorithm because of its higher stability and lower computational cost of this model. Table 5 shows the most informative features used for building Model I.

Tables 6 and 7 provide information about the prediction power of the component of phonetic motor planning (phonetic component) and syntactic and semantic levels of language organization (linguistic component) in identifying patients with ADRD, measured on the test dataset. The highest computed accuracy for phonetic component and linguistic component in detecting patients with ADRD was 78.87 and 83.09, respectively. This result indicates that cognitive impairment has a negative impact on both acoustic and linguistic parts of speech; thus, as shown in Table 1, combining these two components can improve the

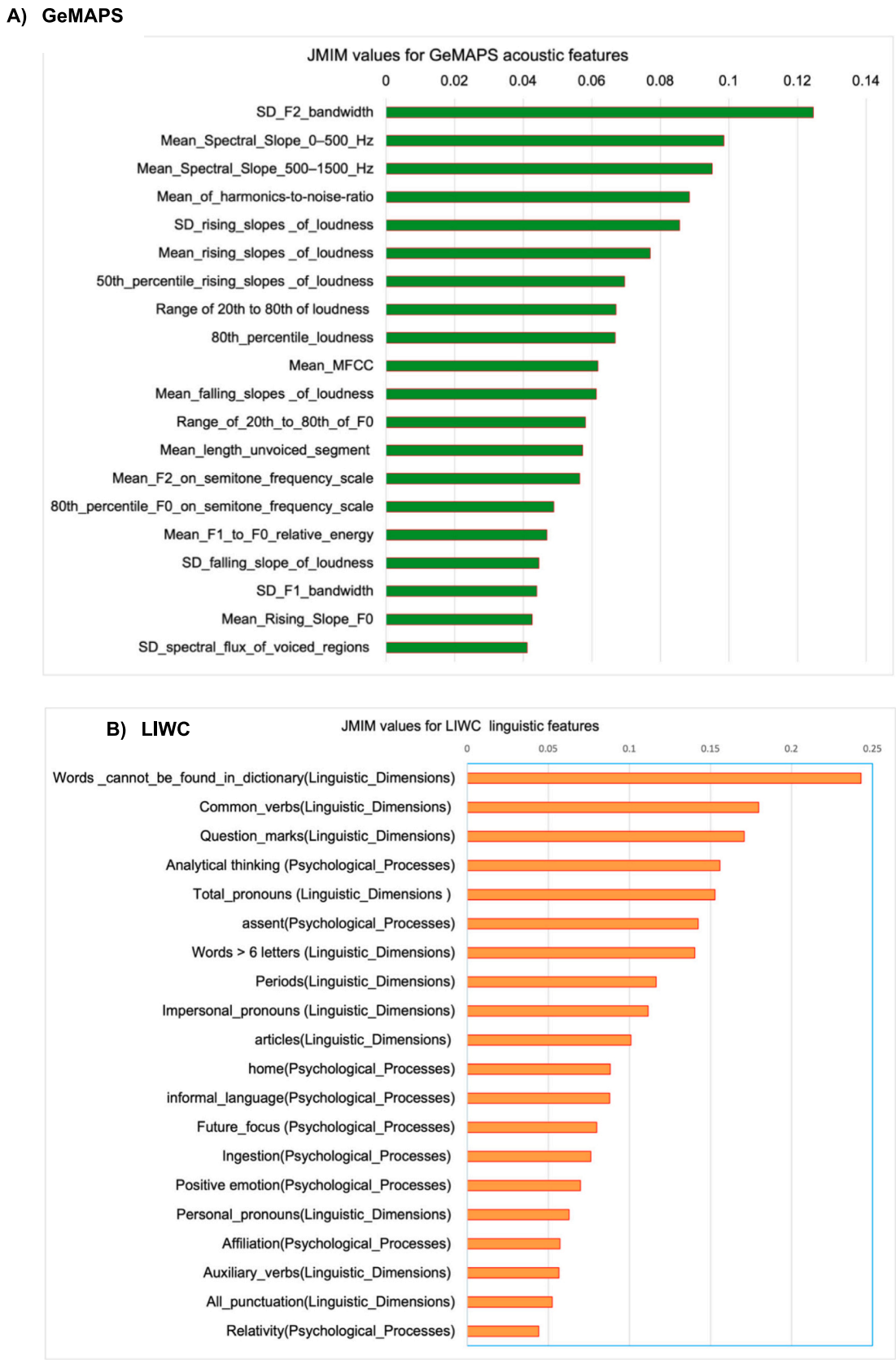


Fig. 6. GeMAPS (A) and LIWC (B) for extracting vocal and semantic linguistic cues in patients with ADRD.

Table 3
Performance of ML model I with different classification algorithms for training and test datasets.

Algorithms	Precision	Recall	F1-score	AUC-ROC	Accuracy
5-Fold cross validation performance of classification algorithms on the training dataset					
SVM	86.92 ± 5.18	82.68 ± 3.94	84.64 ± 3.58	92.53 ± 3.34	85.04 ± 3.41
Extra Trees	85.77 ± 6.05	83.58 ± 7.19	84.26 ± 3.62	92.82 ± 3.64	84.52 ± 3.21
Random Forest	85.14 ± 6.11	83.45 ± 7.72	83.92 ± 4.47	92.08 ± 4.24	84.16 ± 4.04
AdaBoost	88.06 ± 7.36	80.52 ± 9.79	83.63 ± 6.36	93.69 ± 3.61	84.47 ± 5.63
XGBoost	82.99 ± 6.23	82.81 ± 7.04	83.00 ± 5.33	92.39 ± 4.58	82.74 ± 5.18
LR	82.80 ± 5.21	79.21 ± 12.17	80.65 ± 8.36	90.01 ± 4.06	81.55 ± 6.93
Performance of classification algorithms on the test dataset					
SVM	93.75	85.71	89.55	93.89	90.14
Random Forest	88.78	79.54	83.84	91.60	84.94
Extra Trees	89.94	76.79	82.80	91.29	84.31
AdaBoost	77.42	68.57	72.72	84.84	74.65
XGBoost	83.87	74.28	78.78	88.17	80.28
LR	87.10	77.14	81.81	88.01	83.09

Table 4
Performance of ML model II and ML model III with different classification algorithms for training and test datasets.

Algorithms	Precision	Recall	F1-score	AUC-ROC	Accuracy
Performance of ML model II with different classification algorithms on the test dataset					
SVM	**84.21**	**91.42**	**87.67**	**94.44**	**87.32**
Random Forest	82.08	78.57	80.28	91.31	80.98
Extra Trees	82.21	80.18	81.14	91.85	81.67
AdaBoost	81.84	77.35	79.51	80.34	80.38
XGBoost	82.85	82.85	82.85	83.09	83.09
LR	85.29	82.85	84.05	92.06	84.50
Performance of ML model III with different classification algorithms on the test dataset					
SVM	86.11	88.57	87.32	93.57	87.32
Random Forest	84.37	82.01	83.16	91.37	83.62
Extra Trees	83.29	84.09	83.67	91.21	83.83
AdaBoost	80.14	80.13	79.77	79.90	79.90
XGBoost	72.50	82.85	77.33	76.15	76.05
LR	**88.57**	**88.57**	**88.57**	**92.06**	**88.73**

The best precision, recall, F1-score, AUC-ROC, and accuracy performance for the best performing algorithm are highlighted if bold font.

performance of the screening algorithms for detecting patients with ADRD (accuracy = 90.14).

4.3. Added value of speech components in screening patients with ADRD

Fig. 7A demonstrates the added values of each speech component for screening patients with ADRD for AUC-ROC based on the result we obtained from the ML Model I on the training dataset. As this figure shows, combining all three components of speech (phonetic, linguistic, and psycholinguistic) can substantially improve AUC-ROC (AUC-ROC = 92.53) compared to using only the acoustic features of phonetic component (AUC-ROC = 78.49) and the combination of phonetic and linguistic components (AUC-ROC = 90.02). We can see this increase in performance for the Precision/Recall curve (Fig. 7B), Positive Predictive Value (Fig. 7D), and Sensitivity (Fig. 7E). Fig. 7C presents information gain of the speech components in identifying patients with ADRD with respect to the percentage of the study's sample. The gain curve shows

Table 5
Features selected via JMIM for building the highest-performing machine learning algorithm (Model I), ranked by importance.
Component (Com)- 1: phonetic motor planning; 2: semantic and syntactic levels of language organization; 3: psycholinguistic cues.

Linguistic and acoustic features	Com	Linguistic and acoustic features	Com
linregc2 of voice probability	2	Part-of-speech rate	3
Common verbs	4	linregerrQ of simple moving average of LSP frequency	2
Quartile 1 of MFCC	2	Functional words	3
Words cannot be found in Dictionary in LIWC	4	Textual lexical diversity	3
Article	4	Silence time for VERB within clauses	3
Std of LSP frequency	2	Content words	3
Quartile2-Quartile1 LSP frequency	2	Silence time for ADJ/ADV within clauses	2
Voiced segments per second	2	Average of similarity score between clauses without stop word	3
Pause rate	2	Brunet's Index	3
Total average silence duration in initial clauses	2	Indefinites articles	3
LinregerrQ of MFCC	2	Rate of negative adverbs	2
Words that are longer than six letters.	4	Root type-token ratio	3
Definite articles	2	Interquartile range of the 3rd MFCC coefficient	2
Content density	3	Analytical thinking (summary variables in LIWC that measure cognitive language style)	4
Lexical frequency	3	Corrected type-token ratio	3
Std of MFCC	2	Linregc2 of simple moving average of LSP frequency	2
Average length of unvoiced segments	2	linregc1 of perceived loudness	4
Proportion of clauses with a similarity score of zero with stop word	3	Honor's statistic	3
Cognitive processes	4	pitch	4
Skewness of LSP frequency	2	MaxPos of simple moving average (sma) of LSP	2
Std of local Shimmer	2	Normalized standard deviation of simple moving average of F2	4
Silence time for NOUNs within clauses	3	Normalized Std of simple moving average of the amplitude of F1 relative to F0	4
Mean of local jitter	2	Determiners	3
Standard deviation of similarity score between clauses with stop word	3	80th percentile of frequency of 27.5 Hz	2
Relative pronouns rate	3	Ratio of standardized mean amplitude of F3 and F0	4
Mean F0 envelope	2	Std of length of unvoiced segment	2
Total average silence duration per word within clauses	3	80th percentile of loudness	2
Pronouns	3	Reference rate to reality	3
Std of rising slope of loudness	2	Average of similarity score between clauses with stop word	3
std local jitter	2	Std of harmonic noise ratio	4
Mean ratio energy spectral harmonic	2	Unique word count	3
Speech rate	2	Proportion of clauses with a similarity score of zero without stop word	3
Nouns	3	Hypergeometric distribution diversity	3
Quartile2-Quartile3 of F0	2	Word count	4
Long term average spectrum	2	Consecutive repeated clauses	3
Normalized standard deviation of the amplitude of F2 to F0	2	Type-token ratio	3

Table 6
Performance of the component of phonetic motor planning in identifying patients with ADRD.

Algorithms	Precision	Recall	F1-score	AUC-ROC	Accuracy
XGBoost	79.41	77.14	78.26	80.32	78.87
Random Forest	78.47	77.84	78.13	83.07	78.53
ExtraTrees	76.58	77.77	77.15	83.34	77.29
AdaBoost	73.73	74.285	74.01	77.22	74.27
SVM	70.27	74.28	72.22	82.31	71.83
Logistic Regression	68.42	74.28	71.23	75.55	70.42

Table 7
Performance of the component of syntactic and semantic levels of language organization in identifying patients with ADRD.

Algorithms	Precision	Recall	F1-score	AUC-ROC	Accuracy
SVM	84.84	80	82.35	89.05	83.09
Logistic Regression	84.84	80	82.35	89.60	83.09
AdaBoost	85.71	68.57	76.19	82.77	78.87
ExtraTrees	79.73	72.66	75.98	86.01	77.41
Random Forest	77.90	72.34	74.96	85.18	76.25
XGBoost	76.66	65.71	70.77	84.28	73.23

that if we selected the top 40 % of the entire population, representing 66 patients (out of 166), the sample would contain approximately 80 % of patients with ADRD. However, the gain value for using only acoustic features for phonetic component is about 60 % for the sample size (N = 66).

Additionally, as Fig. 7A, B, C, D, and E demonstrates, modeling the patient's psycholinguistic cues can improve the performance of the screening algorithm. For example, as shown in Fig. 7A, adding the psycholinguistic cues to the combination of phonetic and linguistic components improved the AUC-ROC by 2.79 %. Finally, Fig. 7F demonstrates the density plot for the SVN model. As this figure shows, this model has good separability for separating patients with ADRD (class 1) from patients without ADRD (class 0).

Fig. 8A demonstrates the added values of each speech component for screening patients with ADRD based on the result we obtained from the ML Model I on the test dataset. As this figure shows, ML Model I has good generalizability on the unseen data (test dataset) with AUC-ROC = 93.89 for the combination of features of all three components (phonetic, linguistic, and psycholinguistic), AUC-ROC = 93.02 for the combination of phonetic and linguistic components, and AUC-ROC = 82.31 for phonetic component. This improvement can be seen in other metrics, Precision/Recall curve (Fig. 8B), Positive Predictive Value (Fig. 8D), and Sensitivity (Fig. 8E). The information gain curve also shows this improvement (Fig. 8C). By selecting the top 40 % of the entire population (representing 66 patients out of 166), the sample would contain approximately 84 % of patients with ADRD.

5. Discussion

In this study, we provided a detailed perspective on how the patient's verbal response (spontaneous speech) for the Cookie-Theft test can be used for modeling three speech components: (1) the individual's ability in phonetic motor planning, (2) the semantic and syntactic level of language organization, and (3) vocal and semantic psycholinguistic cues. Modeling these three components generated a list of informative features with high discrimination power for building ML models for the proactive identification of patients with ADRD. Each component was composed of domain-related features that can provide insight into the underlying factors associated with the development of ADRD, such as the individual's ability to control their vocal cords, to demonstrate their recall ability and to construct the semantic and syntactic structure of sentences.

Neuropsychological assessment tools, such as the Mini-Mental State Examination (MMSE) [140], the Montreal Cognitive Assessment [141], and the Memory Impairment Screen [142], exhibit acceptable sensitivity and specificity in detecting patients with ADRD [143]. However, their application in clinical settings is often limited due to the patient's difficulty recognizing early symptoms [10] and the clinicians'

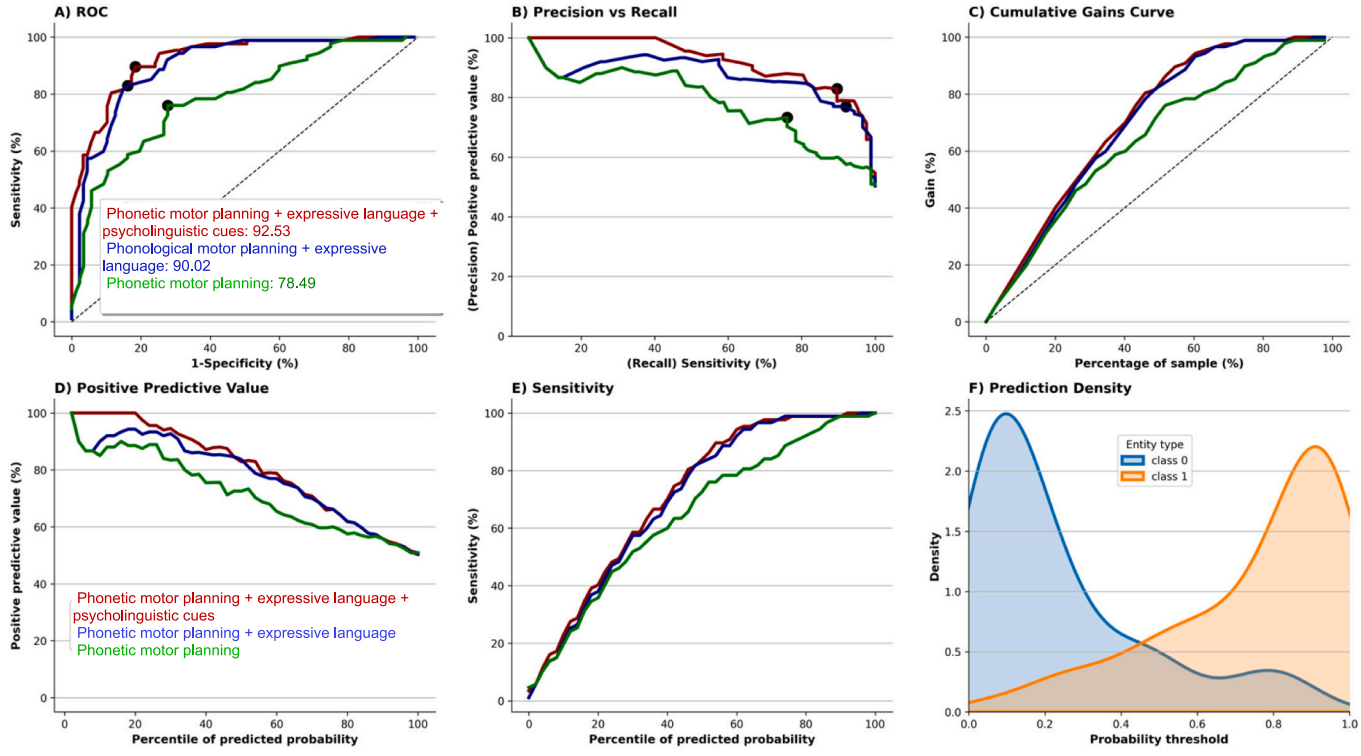


Fig. 7. 5-fold cross validation performance of ML model I and the added value of each speech component for screening patients with ADRD.

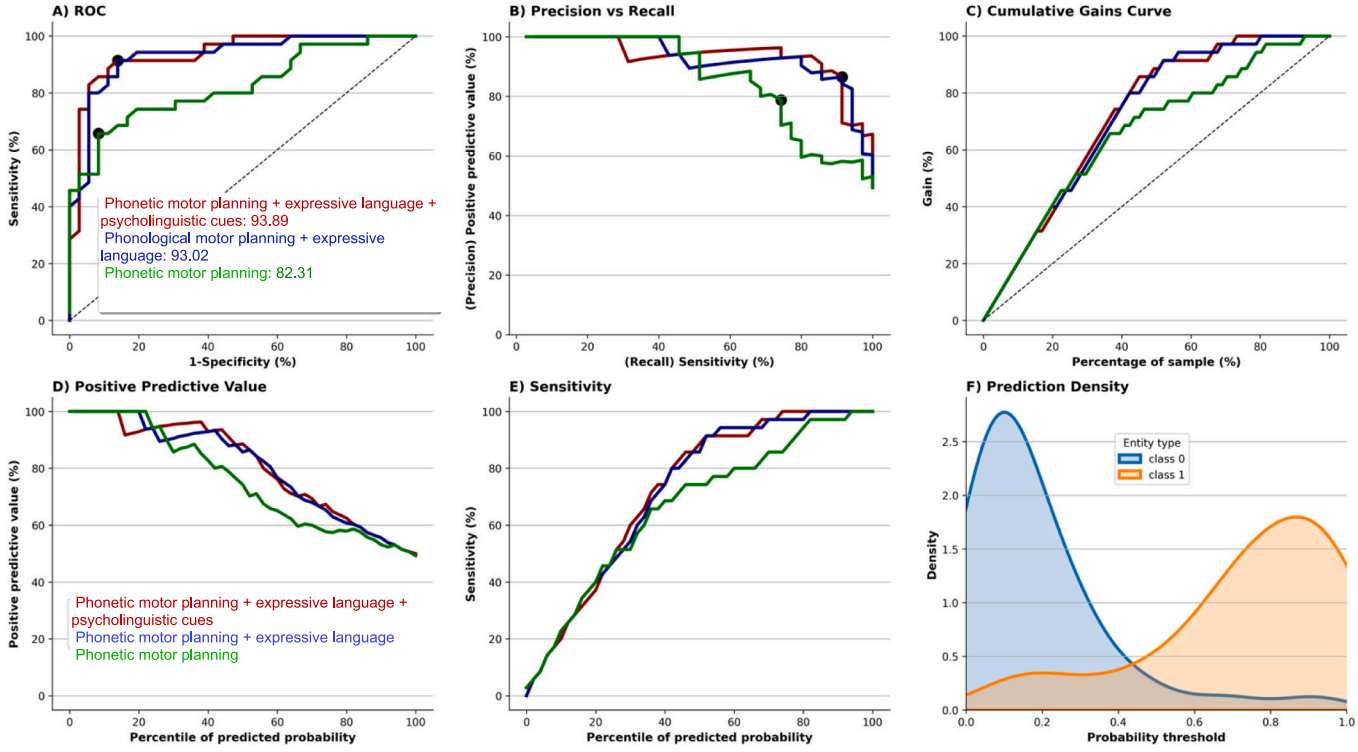


Fig. 8. Performance of ML model I and the added value of each speech component for screening patients with ADRD computed on the test dataset.

insufficient time to assess cognitive impairment [13]. Incorporating AI-based ADRD screening algorithms, such as speech-processing algorithms, into clinical workflows can streamline the patient screening process for ADRD diagnosis by alerting clinicians to patients' cognitive status. This may enable clinicians to implement appropriate interventions, including lifestyle modifications, comorbidity management, or referrals to behavioral health specialists [144]. Consequently, early detection may lead to improvements in the quality of life for patients and their caregivers while reducing overall healthcare utilization and costs [145].

Language corpora in speech production in ADRD has inspired several studies for the automatic assessment of individuals at risk of ADRD. The DementiaBank is the only relatively large, publicly available corpora that include individual verbal responses for the Cookie-Theft test. Several studies were published on the development of automatic screening algorithms for the identification of patients with ADRD using DementiaBank. To model the acoustic part of patients' speech, studies used two major approaches: (1) open-access repositories of acoustic assessment algorithms and (2) transformer-based pretrained speech processing models. For the first approach, AVEC-2013 [26], EMO_Large [27], ComParE-2013 [28], COVAREP [34], and IS10-Paralinguistics feature-set [54] are examples of repositories of acoustic algorithms used by Shah et al. [25], Chen et al. [31], Syed et al. [53], and Rohanian et al. [33] The highest accuracy obtained using these repositories was for ComParE-2013 with accuracy = 71.69 [31] for identifying patients with ADRD. This relatively low accuracy is mostly due to the low generalizability of the acoustic feature sets available in these repositories for screening patients with ADRD. For example, ComParE-2013 includes 6373 generic acoustic descriptors that were mostly developed for music information retrieval and general sound analysis. The development of a repository of acoustic parameters that are specifically associated with cognitive impairment diseases (e.g., dementia, Alzheimer's disease, mild cognitive impairment) can improve the assessment's accuracy of the acoustic component of speech in detecting patients with ADRD. By including significant acoustics parameters associated with ADRD in the component of phonetic motor planning, we were able to achieve an accuracy = 78.87 in this study.

Transformer-based pretrained speech processing models are the second approach that was used by previous studies to process the acoustic part of speech [38,40,42,50,53]. VGGish [51] had the highest accuracy = 72.92 (reported by Koo et al [50]) in detecting patients with ADRD patients compared to MobileNet [43], YAMNet [44], x-vectors [38], i-vectors [40], and Speech BERT [45]. However, the accuracy for VGGish was 64.58 in another study reported by Syed et al. [53] MobileNet, YAMNet, and VGGish were trained on annotated YouTube audio data, and Speech BERT was trained on the LibriSpeech dataset [48]. None of the training audio data is a good representative of the speech data collected from patients with cognitive impairment, resulting in the models' relatively low performance in detecting patients with ADRD. In this study, we also investigated the performance of YAMNet (accuracy = 64.78), but due to reducing the overall performance of ADscreen in detecting patients with ADRD, we decided to exclude it from the model.

To model the linguistic part of patients' speech, studies used both domain-related linguistic features and transformer-based pretrained language models. POS tagging, TF-IDF, n-grams, grammatical dependency, and filled/non-filled pauses are examples of domain-related features that have been used to quantify syntactic and semantic parameters of the patient's langue [24,25,29,50]. For transformer-based pretrained language models, BERT and its extended versions (e.g., XLNet, DistilRoBERTa) were used. Previous studies reported different accuracy for BERT in detecting patients with ADRD deepening on the type of the BERT model used for generating conceptual embedding vectors from the patients' descriptions and the analysis method used to process the embedding vectors. For example, Pompili et al. [39] an accuracy = 72.92 for BERT base-cased and Logistic Regression algorithm for classification. While Zhu et al. [42] reported an accuracy = 82.08 for the longformer language model. Overall, BERT and its extended versions had relatively higher performance than only linguistic related domains. This is mostly because BERT language models were trained on a very large dataset from Wikipedia, Google News, or Biomedical literature

that included textual information similar to patients' descriptions for the Cookie-Theft test. Additionally, BERT models have the capability of modeling the disfluency in patient language, which is a very important factor in detecting patients with ADRD (see Appendix C for details). In this study, we evaluated the performance of BERT base-cased and its extended versions. The DistilBERT model had the highest accuracy in detecting patients with ADRD in our initial analysis (see Appendix D for details). We combined the DistilBERT's word embeddings with linguistic and acoustic domain-related features, which resulted in an accuracy = 83.09.

In summary, ADscreen is an example of a screening algorithm that can provide insight into three major components of speech for modeling phonetic motor planning, levels of language organization, and psycholinguistic cues. Additionally, the analytic pipeline for modeling these three components is generalizable to other speech datasets generated through individuals' spontaneous speech for other neurological assessment tests (e.g., film-recall tasks [8], story-retelling [146] tasks) and speech datasets created through patient-clinician verbal communications in clinical settings. In the next phase of this study, our goal is to upgrade components of the ADscreen by incorporating other domain-related features such as social interaction features that may provide some clues about the risk of ADRD.

5.1. Limitations

- First, although participants in both the case and control groups were selected through extensive physical and neurological examinations, semi-structured psychiatric interviews, and neuropsychological assessments, there remains a risk of misdiagnosing patients with ADRD, primarily due to limited access to biomarkers such as cerebrospinal fluid (CSF) biomarkers. This limitation may affect ADscreen's sensitivity in detecting ADRD patients.
- Second, patients' spontaneous speech for the Cookie-Theft test Pitt corpus was audio-recorded in 1994 using now-outdated technology, resulting in low-quality voice recordings. The low quality of the audio data may impact the accuracy of linguistic and acoustic features extracted from this dataset.
- Third, participants in this dataset are predominantly White. Studies have demonstrated that ML algorithms trained on racially imbalanced data may yield poor predictive performance for minority populations. Therefore, ADscreen's results may not be generalizable to other races and ethnic groups.
- Fourth, although we explored a wide range of acoustic and linguistic features for modeling phonological and language impairment in ADRD patients, other acoustic and linguistic features (such as distinctive grammar patterns) might improve ADscreen's performance in screening for ADRD.
- Fifth, we investigated the performance of three different ML architectures built on both deep-learning methods (CNN and Bi-LSTM) and traditional ML algorithms for developing ADscreen. In the next phase of the study, we plan to explore the performance of other ML architectures (e.g., the combination of CNN and Gated Recurrent Unit [GRU]) in detecting ADRD patients.
- Sixth, the dementia databank does not supply any information regarding the participants' cognitive status or disease stage, which constrains the ML-based screening algorithms developed using this dataset in forecasting disease progression for patients with cognitive impairment.
- Lastly, the databank does not offer evaluation results concerning the participants' emotional status (e.g., anxiety, depression). Changes in emotional status are strong biomarkers for detecting ADRD patients. However, it is not possible to directly model patients' emotions and incorporate them as indicators for ADRD detection. Instead, we extracted vocal and semantic psycholinguistic cues as indicators for detecting ADRD patients using GeMAPS and LIWC in this study.

6. Conclusion

Recent advances in the automated assessment of patients at risk of ADRD should inspire new complex contributions to profiling speech components in pathological cognitive impairment. Both acoustic and linguistic parameters can be very sensitive to changes in the neuropsychological status of the elderly patients; therefore, creating a comprehensive parametric speech profile should be established for (1) assessing the cognitive status of elderly individuals and (2) progression from one clinical stage to another stage (e.g., from cognitively healthy, to mild cognitive impairment, to Alzheimer's disease). This profiling has a potential in monitoring the changes in cognitive status of elderly individuals in order to evaluate the subsequent effectiveness of interventions in stopping or delaying the progress of the disease. In summary, ADscreen has the potential to address the need for an ADRD screening tool, so that patients with these disorders receive appropriate and timely care.

Statement of significance

Alzheimer's disease and related dementias (ADRD) represent a looming public health crisis, affecting roughly 5 million people and 11 % of older adults in the United States [1]. Despite nationwide efforts for timely diagnosis of patients with ADRD, >50 % of them are not diagnosed and unaware of their disease. Missed and delayed diagnosis not only impose more strain on family and caregivers emotionally and financially, but also leads to lost opportunities for treatment and the associated negative outcomes, particularly emergency department visits and hospitalization. Given the projection of 13.2 million ADRD patients by 2050 [14], and the associated cost of more than $1.1 trillion, many organizations, including National Institute of Health and National Science Foundation have recognized the development of a robust diagnostic tool for early identification of elderly patients with ADRD as a critical and urgent research priority by many organizations. Emerging studies showed that patients' spoken language is one of the earliest signs of cognitive impairment, enabling the features of spoken language to act as biomarkers for multiple dimensions of cognitive abilities, including executive functioning, semantic memory, and language. Stablished speech analysis and natural language processing techniques can be utilized for modeling components of spoken language and development of robust acoustic and linguistic metrics for detecting cues of cognitive impairment from the spoken language. In response to the challenges of timely diagnosis of ADRD, we developed ADscreen for proactive automated screening of patients at risk for ADRD. To develop ADscreen, we trained different machine learning algorithms on a combination of a large set of acoustic and linguist parameters and transformer-based methods to detect cues of cognitive impairment in spoken language. We tested performance of the ADscreen on a speech dataset of DementiaBank English Pitt Corpus [60] for ADRD patients. The obtained result not only was promising for identifying patients with ADRD, but it can also provide an insight into the specific type of speech impairments present in these patients in order to adopt appropriate interventions.

Source of funding

K99AG076808- "Development of a Screening Algorithm for Timely Identification of Patients with Mild Cognitive Impairment and Early Dementia in Home Healthcare" from National Institute on Aging.

Declaration of competing interest

We have no financial, commercial, legal, or professional relationship with other organizations, or with the people working with them, that could influence our research.

Appendix A. Cookie-Theft speech description task

Source: Figueiredo, S., & Barfod, V. (2012). Boston Diagnostic Aphasia Examination (BDAE). Chicago [147].

Appendix B. Inclusion and exclusion criteria for recruiting patients for the ADRD study

Patient recruitment: Participants for the ADRD study were enrolled from various clinical settings, including the Benedum Geriatric Center, Multispecialty Outpatient Geriatric Facility at the University of Pittsburgh Medical Center, and local neurologists and psychiatrists [60].

Inclusion criteria: Individuals with a diagnosis for ADRD and symptoms associated with ADRD were eligible for the case group. Individuals with no history of cognitive impairment were eligible for the control group. The research team contacted eligible patients and their caregivers and explained the study's goal, potential risks, and benefits associated with participation. Patients with signed informed consent received extensive physical and neurological examinations, semi-structured psychiatric interviews, and neuropsychological assessments. Also, each participant was interviewed by a psychiatric nurse to assess their physical and cognitive limitations as well as the caregiving burden to their primary caregiver. In addition to the examinations listed, each participant completed various laboratory studies, including blood chemistry, liver and thyroid function tests, and vitamin level tests [60].

Exclusion criteria: Patients with the following symptoms were excluded from this study: Presence of severe manifestation of behavioral and psychiatric symptoms, severe impairment in speech and oral expression of language, significant disease of the central nervous system such as brain tumor, seizure disorder, subdural hematoma, cranial arteritis, the need for emergent care such as uncontrolled pain, wound infection or deterioration, and frequent use of high doses of opioid analgesics.

Appendix C. How does BERT process disfluency?

Tian et al. [32] investigated if and how the BERT language model understands language disfluency using three experiments. For the first experiment, they added a soft-max layer to the medium-sized BERT model (with 12 layers, 12 attention heads, and a total of 110 M parameters). Then, they trained the classifier on a synthetic dataset, including 100 fluent sentences with corresponding disfluent sentences. The finding showed that the BERT language model has an accuracy of 81.3 % in detecting disfluent sentences. The authors suggested that without any fine-tuning on data containing disfluency, BERT already performs fairly well in identifying disfluent data.

With this finding, the authors hypothesized that BERT has an innate understanding of disfluencies. To test this hypothesis, the authors looked inside the Blackbox of the BERT deep learning model to investigate how the embeddings of disfluent sentences change over BERT layers. The authors further hypothesized that if the BERT model can understand language disfluency, the sentence embeddings of the disfluent sentence and its fluent counterpart should be more similar in layers associated with semantic representation than layers associated with surface form and syntactic representation. This is because fluent sentences and their disfluent counterparts are more similar in meaning than form. For this experiment, they used a sample of 900 disfluent utterances from Switchboard corpus [148], including telephone conversations from speakers across the United States. For each disfluent, they created a fluent counterpart by removing filled pauses, interjections, and reparandam. Using two metrics, (1) the raw cosine similarity and (2) the cosine similarity ranking, the authors determined the quality of an embedding in capturing semantic nuances and closeness of a disfluent-fluent pair in the embedding space. The result of the experiment showed that (1) "BERT ranks a disfluent sentence high in similarity compared to all possible fluent counterparts;" (2) "final layer embedding is a relatively good aggregation of sentence meaning;" (3) "In terms of all sentence tokens, the similarity improves steadily in deeper layers, pointing towards increasing semantic selectivity and invariance to disfluencies." These findings confirmed that the BERT language model can understand language disfluency [32]. In the third experiment, the authors analyzed the role of the attention mechanism in identifying semantic disfluency. They found that BERT distinguishes the reparandum and alteration in the sentences by paying less attention to the reparandum in deeper layers.

Overall, the BERT language model processes disfluency by selectively attending to different parts of the disfluency at different intensities using the key attention mechanism. This mechanism allows the BERT language model to differentiate between the word embedding of a disfluent sentence and its fluent counterpart.

Appendix D. Evaluating the performance of different pretrained transformer-based language models in detecting patients with ADRD

BERT [149] is a contextualized word representation model built on a masked language model and pre-trained using bidirectional transformers on large data sets collected from Wikipedia and Google News. The BERT architecture addressed one of the fundamental challenges in language modeling, which is the prediction of a word in a sequence of words (e.g., a sentence). BERT uses a masked language model that predicts randomly masked words in a sequence of words and, therefore, can be used for learning bidirectional representation of a word. This mechanism substantially improved the BERT's performance over traditional language models [150,151] with a mechanism of combining information from two unidirectional language models to improve the accuracy of word prediction.

We hypothesize that the BERT's bidirectional representation mechanism is also critical for modeling the language component of speech in patients with ADRD because it helps to incorporate information related to language disfluency (e.g., repeated words or filler pauses) into word embeddings model. Please see Appendix C for more details about modeling language disfluency using BERT.

We modeled the patient's utterances (for the "Cookie-Theft" test) using the BERT (BERT-base cased, the original version of the BERT) and its extended versions, including DistilBERT, DistilRoBERTa, and XLNet. We selected these four language models because they outperformed other language models in different natural language processing tasks [152,153].

- BERT [149] base-cased comprises 12 transformer blocks, a total of 110 M parameters, 12 attention heads, and a hidden layer size of 768. We used the implementation of the BERT-base case here [154].
- DistilBERT [113]: It is a BERT-based small, fast, and light transformer model that uses 40 % fewer parameters than BERT-base, runs 60 % faster, and keeps >97 % of the BERT's results [113]. DistilBERT comprises the following hyper-parameters: 6 transformer blocks, 65 M parameters, 12 attention heads, and a hidden layer size of 768. We used the implementation of DistilBERT here [114].
- XLNet [52]: It is an extension of the Transformer-XL model, which was trained with an autoregressive method to learn bidirectional contexts. Like BERT base-cased, XLNet comprised the hyperparameters of 12 transformer blocks, a total of 110 M parameters, 12 attention heads, and a hidden layer size of 768. We used the implementation of XLNet here [155].
- DistilRoBERTa-base [156]: DistilRoBERTa-base is a distilled version of the RoBERta-base model. RoBERTa- base model [156] is the extension of the BERT language model. Compared to the BERTmodel, RoBERTa was trained on additional news and stories corpora, and adjusted training strategies were used to improve its performance. DistilRoBERTa-base comprises 6 transformer blocks, a total of 82 M parameters (compared to 125 M parameters for RoBERTa-base), 12 attention heads, and a hidden layer size of 768. On average DistilRoBERTa is twice as fast as RoBERTa-base. We used the implementation of DistilRoBERTa here [157].

We trained each language model independently on the sample of the study (see Section 3.1 Data source), which generated sentence embedding vectors with the size of 768 × 512. The number 768 is the size of the hidden layer, and 512 is the max sequence length (i.e., the maximum number of tokens in a sentence). The embedding vectors for the test dataset were fed into a support vector machine (SVM) classifier algorithm with the RBF kernel. We chose the SVM classifier because it has a better performance in a high-dimensional small sample dataset. Additionally, the SVM classifier has stable performance in different random states. Other classification algorithms, such as ensemble decision trees, provides different performance in different random states. Table 1 shows the performance of the four language models on the test dataset of the study (see Section 3.1 Data source) in detecting patients with ADRD.

Table 1
Performance of pretrained transformer language models in detecting patients with ADRD.

Language model	Precision	Recall	F1-score	Accuracy	AUC-ROC
BERT base-cased	75.67	80	77.77	77.46	83.73
DistilBERT	84.84	80	82.35	83.09	89.05
XLNET	73.68	80	76.71	76.05	83.17
DistilRoBERTa	79.41	77.14	78.26	79.20	78.87

Because Distil BERT had the highest performance compared to other language models, we used this model along with other domain related linguistic features (see Component 3: modeling the patient ability in semantic and syntactic levels of language organization) to model the patient's ability in semantic and syntactic levels of language organization.

Appendix E. Supportive information on the efficacy of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for modeling vocal psycholinguistic cues in spoken language

Vocal psycholinguistic cues conveyed in the voice have been empirically documented by many previous studies in the area of speech and voice analysis [117,158–160]. These studies mostly used established procedures in phonetics and speech sciences to use differentiating parameters of phonation and articulation in speech, which were sensitive to the alternation in the subglottal pressure, transglottal airflow, and vocal fold vibration [161–164]. This includes a large set of acoustic parameters in the "time domain" (e.g., speech rate), the "frequency domain" (e.g., fundamental frequency [F0] or formant frequencies), the "amplitude domain" (e.g., intensity or energy of the voice), and the "spectral energy" domain (e.g., relative energy in different frequency bands). However, a large brute-forced feature set usually results in over-adaptation of classifiers to the training data in machine learning, reducing their generalization capabilities to an unseen dataset. In response to these challenges, the GeMAPS acoustic feature set was introduced to reduce the risk of overfitting and to improve generalization across corpus experiments and ultimately in real-world test scenarios. The GeMAPS contains an acoustic feature set of 18 low-level descriptors (LLD) presented in three acoustic parameter groups: Frequency related parameters, Energy/Amplitude related parameters, and spectral (balance) parameters.

The GeMAPS performance in identifying vocal psycholinguistic cues, specifically Arousal/Valence, was evaluated across several audio-recorded speech datasets labeled for emotion, including (1) FAU AIBO [165], (2) TUM Audiovisual Interest Corpus [165], (3) Berlin Emotion Speech Database, [119] and (4) Geneva Singing Voice Emotion Database [166]. In all these speech databases, the GeMAPS acoustic parameter feature set

surpassed the larger acoustic parameter sets (e.g., ComParE with 6373 acoustic parameters) in detecting psycholinguistic vocal cues, particularly "arousal" and "valence" present in participants' speech.

We conducted a statistical association analysis using *t*-test method to demonstrate the relationship between GeMAPS parameters and ADRD within the study's sample population. The findings can be found in Table 2. According to the findings, out of 88 total acoustic parameters, 57 parameters were significantly associated with ADRD (*P*-value <0.05).

Table 2
Association of GeMAPS features with ADRD.

GeMAPS	Parameter (P-value)
Parameters that are significantly associated with ADRD	loudness_sma3_percentile20.0 (1.80e-29), HNRdBACF_sma3nz_stddevNorm (6.69e-28), mfcc2_sma3_stddevNorm (9.12e-27), slopeV0-500_sma3nz_stddevNorm (4.20e-26), alphaRatioV_sma3nz_stddevNorm (3.23e-24), loudness_sma3_percentile50.0 (2.16e-21), mfcc2V_sma3nz_stddevNorm (1.62e-15), spectralFluxUV_sma3nz_amean (2.34e-13), MeanUnvoicedSegmentLength (7.75e-13), slopeV500-1500_sma3nz_stddevNorm (9.59e-13), spectralFlux_sma3_amean (1.04e-12), StddevUnvoicedSegmentLength (1.49e-12), spectralFluxV_sma3nz_stddevNorm (1.81e-12), MeanVoicedSegmentLengthSec (2.04e-12), loudness_sma3_amean (7.23e-12), mfcc4_sma3_stddevNorm (3.15e-11), loudness_sma3_pctlrange0–2 (1.00e-10), loudness_sma3_percentile80.0 (1.12e-10), spectralFluxV_sma3nz_amean (1.40e-10), StddevVoicedSegmentLengthSec (3.11e-10), loudness_sma3_stddevFallingSlope (8.36e-09), loudness_sma3_stddevRisingSlope (1.12e-08), mfcc3V_sma3nz_stddevNorm (3.55e-08), loudness_sma3_stddevNorm (6.98e-08), mfcc3_sma3_stddevNorm (1.52e-07), spectralFlux_sma3_stddevNorm (3.06e-07), F0semitoneFrom27.5 Hz_sma3nz_pctlrange0–2 (1.28e-06), loudness_sma3_meanFallingSlope (4.70e-06), loudness_sma3_meanRisingSlope (5.03e-06), logRelF0-H1-H2_sma3nz_stddevNorm (6.22e-06), mfcc1V_sma3nz_stddevNorm (1.12e-04), F0semitoneFrom27.5 Hz_sma3nz_stddevFallingSlope (2.04e-04), slopeUV0-500_sma3nz_amean (2.06e-04), mfcc4V_sma3nz_stddevNorm (2.18e-04), HNRdBACF_sma3nz_amean (2.31e-04), jitterLocal_sma3nz_amean (2.39e-04), shimmerLocaldB_sma3nz_stddevNorm (2.77e-04), hammarbergIndexV_sma3nz_stddevNorm (2.96e-04), loudnessPeaksPerSec (7.97e-04), F0semitoneFrom27.5 Hz_sma3nz_percentile20.0 (2.37e-03), slopeV0-500_sma3nz_amean (4.40e-03), logRelF0-H1-A3_sma3nz_stddevNorm (9.82e-03), F0semitoneFrom27.5 Hz_sma3nz_percentile50.0 (9.84e-03), F0semitoneFrom27.5 Hz_sma3nz_stddevNorm (1.20e-02), jitterLocal_sma3nz_stddevNorm (1.25e-02), F1amplitudeLogRelF0_sma3nz_stddevNorm (1.68e-02), F1bandwidth_sma3nz_stddevNorm (1.68e-02), F3frequency_sma3nz_amean (2.04e-02), F2amplitudeLogRelF0_sma3nz_stddevNorm (2.13e-02), F1amplitudeLogRelF0_sma3nz_amean (2.20e-02), F1bandwidth_sma3nz_amean (2.42e-02), mfcc3V_sma3nz_amean (2.47e-02), F3amplitudeLogRelF0_sma3nz_amean (3.08e-02), F3bandwidth_sma3nz_stddevNorm (3.47e-02), F1frequency_sma3nz_amean (3.63e-02), F3amplitudeLogRelF0_sma3nz_stddevNorm (4.35e-02), F2amplitudeLogRelF0_sma3nz_amean (4.48e-02)
Parameters that are not associated with ADRD	hammarbergIndexV_sma3nz_amean (5.64e-02), logRelF0-H1-H2_sma3nz_amean (6.56e-02), F2frequency_sma3nz_stddevNorm (6.61e-02), F2frequency_sma3nz_amean (8.62e-02), F1frequency_sma3nz_stddevNorm (1.02e-01), F0semitoneFrom27.5 Hz_sma3nz_amean (1.14e-01), logRelF0-H1-A3_sma3nz_amean (1.15e-01), F0semitoneFrom27.5 Hz_sma3nz_meanFallingSlope (1.16e-01), F0semitoneFrom27.5 Hz_sma3nz_percentile80.0 (1.31e-01), equivalentSoundLevel_dBp (1.46e-01), slopeUV500-1500_sma3nz_amean (2.19e-01), F3frequency_sma3nz_stddevNorm (2.54e-01), hammarbergIndexUV_sma3nz_amean (2.71e-01), mfcc1V_sma3nz_amean (4.22e-01), mfcc4V_sma3nz_amean (4.86e-01), F3bandwidth_sma3nz_amean (4.91e-01), mfcc3_sma3_amean (5.08e-01), mfcc2_sma3_amean (5.33e-01), mfcc2V_sma3nz_amean (5.79e-01), VoicedSegmentsPerSec (6.12e-01), F0semitoneFrom27.5 Hz_sma3nz_meanRisingSlope (6.83e-01), shimmerLocaldB_sma3nz_amean (7.12e-01), F0semitoneFrom27.5 Hz_sma3nz_stddevRisingSlope (7.19e-01), F2bandwidth_sma3nz_stddevNorm (7.62e-01), mfcc1_sma3_stddevNorm (7.94e-01), alphaRatioUV_sma3nz_amean (8.17e-01), slopeV500-1500_sma3nz_amean (8.33e-01), mfcc1_sma3_amean (8.44e-01), alphaRatioV_sma3nz_amean (8.72e-01), mfcc4_sma3_amean (9.61e-01), F2bandwidth_sma3nz_amean (9.68e-01)

Appendix F. Supportive information on the efficacy of Linguistic Inquiry and Word Count (LIWC) for identifying semantic psycholinguistic cues in spoken language

Developers of LIWC [23] followed a rigorous methodology to improve its internal reliability and external validity for linguist analysis and identifying psychological cues in individual's langague. The methodology consists of seven steps. In step 1, vocabularies for each category were generated from several sources, such as the emotion rating scales (e.g., PANAS), Roget's Thesaurus, and standard English dictionaries. Additionally, 3–6 judges participated in brainstorming sessions to generate more related words for each category. In step 2, each word from the grand list of each category was examined by 4–8 judges and qualitatively rated in terms of "goodness of fit" for each category. In step 3, the developers focused on the analysis of a working version of the dictionary using the Meaning Extraction Helper [167] to determine how frequently the dictionary words were used in various contexts in different language corpora (e.g., Twitter, Facebook). Words that did not occur at least once in these language corpora were removed from the dictionary. In step 4, several language corpora were explored using the MEH tool to identify useful vocabularies that are not available in the dictionary. The candidate vocabularies were evaluated by 4–8 judges for the level of conceptual fitness for a specific category. In step 5, the developers computed the internal consistency of each vocabulary for its related category. Vocabularies that were detrimental to the internal consistency of the category were evaluated by 2–8 judges for retaining or removal from the document. In step 6, the developers repeated steps 1 through 5 to catch any possible mistakes that might have occurred throughout the development process.

By following this rigorous process, the LIWC developers were able to achieve high internal reliability and external validity of this text analysis tool in detecting psycholinguistic cues and language composition elements associated with cognitive and affective processes [95,124–127]. In the area of healthcare, the reliability and validity of LIWC in detecting semantic psycholinguistic cues associated with mental and neurological disorders have been verified in several studies [95,124–127]. For example, Burkhardt et al. [125] found that LIWC features specifically linguistic features (e.g., pronoun use), can distinguish between individuals with and without schizophrenia. The authors concluded that elements of language composition are different in individuals with and without Schizophrenia and LIWC language features are able to catch this difference. LIWC was also useful for detecting patients with cognitive impairment. For example, Asgari et al. found that linguistic markers from domains of "psychological process" and "linguistic structure" were associated with the presence of cognitive impairment [95].

We conducted a statistical association analysis using *t*-test method to demonstrate the relationship between LIWC parameters and ADRD within the study's sample population. The findings can be found in Table 3. According to the findings, out of 93 total linguistic parameters, 79 parameters were significantly associated with ADRD (*P*-value <0.05).

Table 3
Association of LIWC features with ADRD.

LIWC	Parameter (P-value)
Parameters that are significantly associated with ADRD	Other punctuation (4.09e-32), sexual (4.09e-32), filler (8.89e-32), Exclam (1.52e-31), Dash (1.92e-31), swear (8.33e-31), death (1.22e-30), friend (2.19e-30), sad (1.33e-29), relig (4.74e-29), they (2.67e-28), anger (3.86e-28), anx (1.54e-27), risk (7.54e-27), money (1.16e-26), cause (1.43e-25), work (1.32e-23), negemo (1.93e-23), netspeak (4.04e-23), feel (1.26e-21), interrog (2.64e-21), health (3.83e-21), we (3.95e-21), assent (4.84e-21), QMark (5.48e-21), achieve (2.72e-19), focusfuture (1.76e-18), discrep (1.86e-18), WC (4.36e-18), number (7.59e-18), focuspast (9.48e-18), reward (1.44e-17), hear (2.45e-17), family (4.70e-17), you (6.91e-17), posemo (9.92e-17), home (1.18e-16), body (2.17e-16), affiliation (1.39e-15), certain (4.80e-15), compare (1.88e-14), affect (2.70e-14), quant (2.77e-14), informal (3.62e-14), negate (7.36e-14), nonflu (1.25e-13), WPS (2.59e-13), insight (3.78e-13), adj (4.21e-13), i (1.16e-12), tentat (5.62e-12), male (1.62e-11), differ (2.25e-11), Comma (4.02e-11), leisure (1.20e-10), adverb (1.93e-10), power (6.12e-10), time (9.53e-10), see (4.55e-09), motion (6.22e-09), Period (7.86e-08), ipron (3.52e-07), shehe (4.76e-07), conj (1.51e-06), AllPunc (1.63e-06), Authentic (4.06e-06), Clout (4.14e-06), ingest (6.60e-06), cogproc (3.73e-05), bio (4.68e-05), percept (1.23e-04), Apostro (4.77e-04), drives (5.57e-04), female (1.63e-03), Analytic (2.83e-03), ppron (8.14e-03), verb (1.33e-02), function (1.97e-02), pronoun (4.67e-02)
Parameters that are not associated with ADRD	elativ (6.76e-02), space (8.42e-02), social (1.31e-01), Tone (2.55e-01), focuspresent (4.89e-01), article (5.43e-01), prep (6.51e-01), Dic (6.72e-01), Sixltr (6.87e-01), auxverb (7.43e-01), Colon (all values 0 for this variable), SemiC (all values 0 for this variable), Quote (nan), Parenth (all values 0 for this variable).

Appendix G. YAMNet acoustic embedding model

YAMNet is a pretrained deep-learning neural network acoustic embedding model that was trained on a human-labeled YouTube audio dataset [47] for audio events. YAMNet has the same architecture as MobileNet with a Backbone Convolutional Neural Network (CNN). This Backbone takes a 3-dimensional (h, w, 3) matrix of an image as an input, where "h" is height, "w" is width, and 3 is the RGB channel. Next, the Backbone converts this matrix into a three-dimensional matrix with size (h′, w′, 1024) where (h′, w′) is a function of (h, w) and 1024 represents the depth of the backbone CNN. The output matrix (h′, w′, 1024) is then sent to a Global Average Pooling (GAP) layer for reducing the dimension of h' and w' and obtaining a vector of 1024-dimension feature. Then, the GAP output is passed to a Fully Connected (FC) layer with 1000 neurons. The FC is connected to a SoftMax activation layer for producing the classification results. The MobileNet Backbone was trained on image datasets. Compared to MobileNet, YAMNet has a better performance on downstream tasks (e.g., patients' description for the Cookie-Theft test) for processing audio data as it was trained on audio datasets. Table 4 includes the performance of YAMNet for detecting patients with ADRD using the audio-recorded data of patients' descriptions for the Cookie-Theft test.

Table 4
Performance of YAMNet on processing the audio-recorded data (of patients' description for the Cookie-Theft test) for detecting patients with ADRD.

Language model	Precision	Recall	F1-score	Accuracy	AUC-ROC
YAMNet result	61.36	77.14	68.35	64.78	59.60

Appendix H. Detailed descriptions of machine learning classifiers used in machine learning Models I, II, and III to identify patients with ADRD

We used different discriminative machine learning (ML) algorithms including logistic regression as baseline algorithm, bootstrap aggregation [168] (bagging) and Gradient Boosting [169] ensemble decision trees as non-parametric ML methods with the ability of generating a large number of decision trees (weak learners), and support vector machine (SVM) as an parametric kernel based algorithm.

Logistic regression uses a logistic function (an optimization function) to estimate the coefficient of each input variable (predictor) in predicting the outcome variables. Logistic regression performance is often used as a baseline binary ML classification model because of its simplicity and efficiency in implementation [170].

Bagging methods are ensemble decision tree algorithms, in which weak learners are trained independently and in parallel on the entire sample or subsets of the training sample and weighted equally for computing the final outcome [168]. We used two popular algorithms, Random Forest and Extremely Randomized Trees [134] (Extra Trees) from bagging methods. Random Forest generates each weak learner from a bootstrap sample (a resampling technique used to estimate statistics on a population by sampling a dataset with replacement), while the Extra Trees algorithm fits each weak decision tree onto the full original learning sample. Both algorithms use a subset of features randomly selected at each split point (nodes) for growing the weak learner trees. However, unlike Random Forest, which uses a greedy algorithm to select an optimal split point, the Extra Trees algorithm selects the split point completely at random. For both algorithms, outcomes of the weak trees are aggregated using the majority votes in classification problems to yield the final prediction. Since Extra Trees uses explicit randomization of the cut-point and the full original learning sample rather than bootstrap replicas, it has the ability to reduce both variance and bias more strongly than the Random Forest with a weaker randomization scheme [134].

Gradient Boosting methods are ensemble decision tree algorithms in which weak learners are generated in a sequential way, taking into account the error of the previous decision tree algorithm, and they are weighted according to their performance for computing the final outcomes [169]. We used Adaptive Boosting [135] (AdaBoost) and extreme gradient boosting (XGBoost) [136], two popular methods from this boosting category. AdaBoost creates an ensemble of weak learners iteratively by modifying the weights of misclassified data in each iteration. Weak learners with misclassified outcomes receive larger weights, and therefore they have a higher probability of appearing in the training sample for the succeeding weak learners. Unlike AdaBoost, which changes the distribution of a sample distribution for training weak learners, XGBoost uses a gradient descent algorithm (an algorithm for minimizing a loss function) to reduce the number of errors. As weak learners are incrementally on the remaining residual errors of a strong learner, it does not alter the sample distribution. Both algorithms use optimization algorithms for selecting features for growing trees

and split points of the nodes. However, unlike XGBoost, which uses optimization parameters to compute the depth of the weak learners, AdaBoost creates weak learners with a single split, called decision stumps. Additionally, XGBoost uses regularization parameters to reduce the complexity of the regression tree function and bias-variance of the model (the bias-variance tradeoff).

SVMs belong to the general category of kernel methods [137]. A kernel method is an algorithm with a dependence on data only limited through inner products of data vectors. Therefore, a kernel function can replace the inner products, which computes the inner product in high-dimensional data. The advantage of this approach is twofold: first, it has the ability to generate nonlinear boundaries using algorithms for linear classifiers. Second, the use of the kernel function allows the user to apply a classifier to data without fixed-dimensional vector space representation. SVM itself was built on four basic concepts [171]: (1) the separating hyperplane for dividing the data with regard to their outcome label; (2) the maximum-margin hyperplane for maximizing the generalizability of the SVM; (3) the soft margin that uses the regularization parameter that controls the trade-off between maximization of the margin and minimizing the misclassification error; and (4) the kernel function, a function that uses similarity measures for transforming data and reducing the vector dimensions. The kernel function reduces the data complexity and improves the separability of data, which is particularly useful in high dimensional feature space. Studies showed that optimization of SVM parameters can reduce variation and bias in prediction tasks [171,172].

Appendix I. Parameters tuned for machine learning algorithms

ML algorithms	Parameters

Logistic Regression	params = {
	Logistic Regression:
	norm of the penalty: [“l1”, “l2”, “elasticnet”]
	}

Random Forest	params = {
	number of trees in the forest: [10, 100, 500, 1000],
	maxim features at each split: [2, 4, 6, 8],
	function to measure the quality of a split: [“gini”, “entropy”]
	}

Extra Trees	params = {
	number of trees in the forest: [50, 500, 1000, 5000],
	number of features to consider when looking for the best split: {“sqrt”, “log2”, None},
	function to measure the quality of a split: [“gini”, “entropy”]
	}

XGBoost	params = {
	minimum sum of instance weight (hessian) needed in a child: [1, 5, 10],
	minimum loss reduction required to make a further partition on a leaf node of the tree(Gamma): [0.5, 1, 1.5, 2, 5],
	subsample ratio of the training instances: [1.0, 0.8, 0.6],
	subsample ratio of columns when constructing each tree: [0.6, 0.8, 1.0],
	maximum depth: [3, 4, 5]
	}

AdaBoost	params = {
	maximum number of estimators at which boosting is terminated: [500, 1000, 2000, 5000],
	Weight applied to each classifier at each boosting iteration (learning_rate): [0.001,0.01,0.1]
	}

Support Vector Machine	params = {
	kernel: [“poly”, “linear”, “rbf”, “sigmoid”],
	Degree of the polynomial kernel function: [2, 3, 4, 5, 10],
	Kernel coefficient for ‘rbf’, ‘poly’ and ‘sigmoid’: [“scale”, “auto”]
	}

Appendix J. Performance of machine learning Models II and III for detecting patients with ADRD

See details of ML Models II and III in Section 3.6.3 “In processing phase: machine learning (ML) architecture.”

Table 5
Performance of ML Model II with different classification algorithms for training and test datasets.

5-Fold cross validation performance of classification algorithms on the training dataset					
Algorithms	Precision	Recall	F1-score	AUC-ROC	Accuracy
SVM	77.47 ± 6.38	89.67 ± 8.52	**82.97 ± 6.56**	**89.03 ± 5.68**	**80.76 ± 7.15**
Extra Trees	79.21 ± 7.69	82.17 ± 5.98	80.59 ± 6.56	86.91 ± 6.62	79.11 ± 7.26
Random Forest	77.65 ± 7.58	80.81 ± 8.48	79.14 ± 7.78	85.03 ± 6.17	77.67 ± 8.28
AdaBoost	74.58 ± 6.30	70.20 ± 2.64	72.19 ± 3.57	75.87 ± 3.80	71.52 ± 4.78
XGBoost	78.87 ± 7.36	70.13 ± 1.84	74.09 ± 3.47	77.83 ± 7.53	74.09 ± 4.50
LR	79.05 ± 5.79	81.63 ± 9.58	80.09 ± 6.45	84.93 ± 8.15	78.98 ± 6.28

The best F1-score, AUC-ROC, and Accuacy of best performing algorithm are highlighted if bold font.

Table 6
Performance of ML Model III with different classification algorithms for training and test datasets.

5-Fold cross validation performance of classification algorithms on the training dataset					
Algorithms	Precision	Recall	F1-score	AUC-ROC	Accuracy
SVM	78.79 ± 4.67	79.41 ± 10.62	**78.65 ± 5.77**	**85.77 ± 6.75**	**77.73 ± 5.48**
Extra Trees	78.17 ± 6.74	79.38 ± 10.26	78.27 ± 6.23	85.39 ± 8.01	77.16 ± 6.30
Random Forest	77.55 ± 6.91	80.08 ± 12.51	78.33 ± 8.01	85.33 ± 8.34	77.20 ± 8.03
AdaBoost	74.05 ± 5.75	77.10 ± 14.20	74.84 ± 7.96	78.29 ± 9.08	73.47 ± 7.66
XGBoost	75.13 ± 2.94	79.34 ± 13.12	76.61 ± 6.66	83.72 ± 4.69	75.32 ± 5.38
LR	77.47 ± 5.41	77.12 ± 9.47	76.91 ± 5.50	83.52 ± 7.77	75.93 ± 5.56

The best F1-score, AUC-ROC, and Accuacy of best performing algorithm are highlighted if bold font.

References

[1] Association, A, Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. Alzheimers Dement 2013;9:208–45.
[2] Zhu CW, et al. Health-related resource use and costs in elderly adults with and without mild cognitive impairment. J Am Geriatr Soc 2013;61:396–402.
[3] St-Hilaire A, Hudon C, Préville M, Potvin O. Utilization of healthcare services among elderly with cognitive impairment no dementia and influence of depression and anxiety: a longitudinal study. Aging Ment Health 2017;21:810–22.
[4] Rovner BW, Casten RJ. Emergency department visits in African Americans with mild cognitive impairment and diabetes. J Diabetes Complications 2021;35:107905.
[5] Stephens CE, Newcomer R, Blegen M, Miller B, Harrington C. The effects of cognitive impairment on nursing home residents' emergency department visits and hospitalizations. Alzheimers Dement 2014;10:835–43.
[6] Perry W, et al. Population health solutions for assessing cognitive impairment in geriatric patients. Innov Aging 2018;2:igy025.
[7] Boise L, Neal MB, Kaye J. Dementia assessment in primary care: results from a study in three managed care systems. J Gerontol Ser A Biol Sci Med Sci 2004;59:M621–6.
[8] Tóth L, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr Alzheimer Res 2018;15:130–8.
[9] National Institute on Aging. Assessing cognitive impairment in older patients. https://www.nia.nih.gov/health/assessing-cognitive-impairment-older-patients.
[10] Lion KM, et al. Do people with dementia and mild cognitive impairments experience stigma? A cross-cultural investigation between Italy, Poland and the UK. Aging Ment Health 2020;24:947–55.
[11] Van De Pol LA, et al. Magnetic resonance imaging predictors of cognition in mild cognitive impairment. Arch Neurol 2007;64:1023–8.
[12] Zetterberg H, Blennow K. Blood biomarkers: democratizing alzheimer's diagnostics. Neuron 2020;106:881–3.
[13] Judge D, Roberts J, Khandker R, Ambegaonkar B, Black CM. Physician perceptions about the barriers to prompt diagnosis of mild cognitive impairment and Alzheimer's disease. Int J Alzheimer's Dis 2019;2019.
[14] Nichols LO, et al. Impact of the REACH II and REACH VA dementia caregiver interventions on healthcare costs. J Am Geriatr Soc 2017;65:931–6.
[15] National Institute on Aging. The National Institute on Aging: strategic directions for research, 2020–2025. https://www.nia.nih.gov/about/aging-strategic-directions-research/goal-health-interventions#c2; 2021.
[16] Johnson M, Lin F. Communication difficulty and relevant interventions in mild cognitive impairment: implications for neuroplasticity. Top Geriatr Rehabil 2014;30:18.
[17] Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. Front Psychol 2021;12:645.
[18] Tóth L, et al. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Sixteenth annual conference of the International Speech Communication Association; 2015.
[19] Mirzaei S, et al. Two-stage feature selection of voice parameters for early Alzheimer's disease prediction. IRBM 2018;39:430–5.
[20] Han K-H, et al. Impairment of vocal expression of negative emotions in patients with Alzheimer's disease. Front Aging Neurosci 2014;6:101.
[21] Cadieux NL, Greve KW. Emotion processing in Alzheimer's disease. J Int Neuropsychol Soc 1997;3:411–9.
[22] Spazzapan EA, Marino VC de C, Cardoso VM, Berti LC, Fabron EMG. Acoustic characteristics of voice in different cycles of life: an integrative literature review. Rev CEFAC 2019;21.
[23] Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. 2015.
[24] Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. Front Aging Neurosci 2021;13:635945.
[25] Shah Z, et al. Learning language and acoustic models for identifying Alzheimer's dementia from speech. Front Comput Sci 2021;4.
[26] Valstar M, et al. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge; 2013. p. 3–10.
[27] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia; 2010. p. 1459–62.
[28] Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on multimedia; 2013. p. 835–8.
[29] Martinc M, Pollak S. Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia. In: INTERSPEECH; 2020. p. 2157–61.
[30] Eyben F, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 2015;7:190–202.
[31] Chen J, Ye J, Tang F, Zhou J. Automatic detection of Alzheimer's disease using spontaneous speech only. In: Interspeech. vol. 2021. NIH Public Access; 2021. p. 3830.
[32] Tian Y, Nieradzik T, Jalali S, Shiu D. How does BERT process disfluency?. In: Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue; 2021. p. 208–17.
[33] Rohanian M, Hough J, Purver M. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech. arXiv Prepr. arXiv2106.09668. 2021.
[34] Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2014. p. 960–4.
[35] Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–43.
[36] Hough J, Schlangen D. Recurrent neural networks for incremental disfluency detection. 2015.
[37] Pappagari R, Cho J, Moro-Velazquez L, Dehak N. Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. In: INTERSPEECH; 2020. p. 2177–81.
[38] Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-vectors: robust dnn embeddings for speaker recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2018. p. 5329–33.
[39] Pompili A, Rolland T, Abad A. The INESC-ID multi-modal system for the ADReSS 2020 challenge. arXiv Prepr. arXiv2005.14646. 2020.
[40] Saon G, Soltau H, Nahamoo D, Picheny M. Speaker adaptation of neural network acoustic models using i-vectors. In: 2013 IEEE workshop on automatic speech recognition and understanding 55–59. IEEE; 2013.
[41] Nagrani A, Chung JS, Zisserman A. Voxceleb: a large-scale speaker identification dataset. arXiv Prepr. arXiv1706.08612. 2017.
[42] Zhu Y, Liang X, Batsis JA, Roth RM. Exploring deep transfer learning techniques for alzheimer's dementia detection. Front Comput Sci 2021;3.
[43] Howard AG, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv Prepr. arXiv1704.04861. 2017.
[44] Naranjo-Alcazar J, et al. An open-set recognition and few-shot learning dataset for audio event classification in domestic environments. arXiv Prepr. arXiv2002.11561. 2020.
[45] Chuang Y-S, Liu C-L, Lee H-Y, Lee L. Speechbert: an audio-and-text jointly learned language model for end-to-end spoken question answering. arXiv Prepr. arXiv1910.11559. 2019.
[46] Deng J, et al. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009. p. 248–55.
[47] Gemmeke JF, et al. Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2017. p. 776–80.
[48] Pratap V, Xu Q, Sriram A, Synnaeve G, Collobert R. Mls: a large-scale multilingual dataset for speech research. arXiv Prepr. arXiv2012.03411. 2020.
[49] Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv Prepr. arXiv2004.05150. 2020.
[50] Koo J, Lee JH, Pyo J, Jo Y, Lee K. Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. arXiv Prepr. arXiv2009.04070. 2020.

[51] Hershey S, et al. CNN architectures for large-scale audio classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2017. p. 131–5.
[52] Yang Z, et al. Xlnet: generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst 2019;32.
[53] Syed MSS, Syed ZS, Lech M, Pirogova E. Automated screening for Alzheimer's dementia through spontaneous speech. In: Interspeech. 2020; vol. 2020. p. 2222–6.
[54] Schuller B, et al. The INTERSPEECH 2010 paralinguistic challenge. In: Proc. INTERSPEECH 2010, Makuhari, Japan; 2010. p. 2794–7.
[55] 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 Prepr. arXiv2008.01551. 2020.
[56] Kong W. Exploring neural models for predicting dementia from language. 2019.
[57] Bertini F, Allevi D, Lutero G, Calzà L, Montesi D. An automatic Alzheimer's disease classifier based on spontaneous spoken English. Comput Speech Lang 2022;72:101298.
[58] Park DS, et al. Specaugment: a simple data augmentation method for automatic speech recognition. arXiv Prepr. arXiv1904.08779. 2019.
[59] Roshanzamir A, Aghajan H, Soleymani Baghshah M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. BMC Med Inform Decis Mak 2021;21:1–14.
[60] Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Arch Neurol 1994;51:585–94.
[61] Cummings L. Describing the cookie theft picture: sources of breakdown in Alzheimer's dementia. Pragmat Soc 2019;10:153–76.
[62] Slegers A, Filiou R-P, Montembeault M, Brambati SM. Connected speech features from picture description in Alzheimer's disease: a systematic review. J Alzheimers Dis 2018;65:519–42.
[63] RX 8-Great Audio Starts With RX. https://www.izotope.com/en/products/rx.html.
[64] Yang, Q., Wu, P. & Duan, Z. Large-scale analysis of lyrics and melodies in Cantonese pop songs.
[65] Koçer, B. A technical review of white noise in a spotify sample. Porte Akad. Müzik ve Dans Araştırmaları Derg. 7–18.
[66] Meilán JJG, Martínez-Sánchez F, Martínez-Nicolás I, Llorente TE, Carro J. Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. Behav Neurol 2020; 2020.
[67] Duffy JR, Josephs KA. The diagnosis and understanding of apraxia of speech: why including neurodegenerative etiologies may be important. J Speech Lang Hear Res 2012;55:S1518–22.
[68] Ward M, Cecato JF, Aprahamian I, Martinelli JE. Assessment for apraxia in mild cognitive impairment and Alzheimer's disease. Dement Neuropsychol 2015;9: 71–5.
[69] Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology 2000;14:71–91.
[70] Themistocleous C, Eckerström M, Kokkinakis D. Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. PLoS One 2020;15:e0236009.
[71] Huet K, Delvaux V, Piccaluga M, Roland V, Harmegnies B. Inter-syllabic interval as an indicator of fluency in Parkinsonian French speech. In: 11th international seminar on speech production, Tianjin, China; 2017.
[72] Yeldener S. Method of determining the voicing probability of speech signals. Acoust Soc Am J 2002;111:25.
[73] Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the institute of phonetic sciences. vol. 17; 1993. p. 97–110. Amsterdam.
[74] Viegas F, et al. Comparison of fundamental frequency and formants frequency measurements in two speech tasks. Rev CEFAC 2019;21.
[75] Wright R, Nichols D. Measuring vowel formants. Retrieved January 20, 2017. 2015.
[76] Khodabakhsh A, Yesil F, Guner E, Demiroglu C. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. EURASIP J Audio Speech Music Process 2015;2015:9.
[77] Meilan JJG, Martinez-Sanchez F, Carro J, Carcavilla N, Ivanova O. Voice markers of lexical access in mild cognitive impairment and Alzheimer's disease. Curr Alzheimer Res 2018;15:111–9.
[78] Tomas B, Zelenika D. Determination of spectral parameters of speech signal by Goertzel algorithm. In: Speech Technol; 2011.
[79] On CK, Pandiyan PM, Yaacob S, Saudi A. Mel-frequency cepstral coefficient analysis in speech recognition. In: 2006 international conference on computing & informatics. IEEE; 2006. p. 1–5.
[80] Meghanani A, Anoop CS, Ramakrishnan AG. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In: 2021 IEEE spoken language technology workshop (SLT). IEEE; 2021. p. 670–7.
[81] Dessouky MM, Elrashidy MA, Taha TE, Abdelkader HM. Computer-aided diagnosis system for Alzheimer's disease using different discrete transform techniques. Am J Alzheimer's Dis Other Dementias® 2016;31:282–93.
[82] Kong Y-Y, Mullangi A, Marozeau J, Epstein M. Temporal and spectral cues for musical timbre perception in electric hearing. 2011.
[83] Tjaden K, Sussman JE, Liu G, Wilding G. Long-term average spectral (LTAS) measures of dysarthria and their relationship to perceived severity. J Med Speech Lang Pathol 2010;18:125.
[84] Martínez-Nicolás I, Llorente TE, Ivanova O, Martínez-Sánchez F, Meilán JJG. Many changes in speech through aging are actually a consequence of cognitive changes. Int J Environ Res Public Health 2022;19:2137.
[85] Farrús M, Hernando J, Ejarque P. Jitter and shimmer measurements for speaker recognition. In: 8th annual conference of the International Speech Communication Association; 2007 Aug. 27–31; Antwerp (Belgium).[place unknown]: ISCA; 2007. International Speech Communication Association (ISCA); 2007. p. 778–81.
[86] Ivanova O, et al. Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. Comput Speech Lang 2022;73:101341.
[87] Simonyan K, et al. Focal white matter changes in spasmodic dysphonia: a combined diffusion tensor imaging and neuropathological study. Brain 2008;131: 447–59.
[88] David M. The new voice pedagogy. Scarecrow Press; 2008.
[89] Maryn Y, De Bodt M, Roy N. The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders. J Commun Disord 2010;43: 161–74.
[90] Abercrombie D. Elements of general phonetics, Edinburgh Univer. 1966.
[91] Ivanova O, García Meilán JJ, Martínez-Sánchez F, Carro Ramos J. Speech disorders in Alzheimer's disease: preclinical markers of dementia? Psychol Appl Trends Pr C 2018:464–8.
[92] Roark B, Mitchell M, Hosom J-P, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans Audio Speech Lang Process 2011;19:2081–90.
[93] Kim BS, Kim YB, Kim H. Discourse measures to differentiate between mild cognitive impairment and healthy aging. Front Aging Neurosci 2019;11:221.
[94] Aramaki E, Shikata S, Miyabe M, Kinoshita A. Vocabulary size in speech may be an early indicator of cognitive impairment. PLoS One 2016;11:e0155195.
[95] Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's Dement Transl Res Clin Interv 2017; 3:219–28.
[96] Sung JE, Choi S, Eom B, Yoo JK, Jeong JH. Syntactic complexity as a linguistic marker to differentiate mild cognitive impairment from normal aging. J Speech Lang Hear Res 2020;63:1416–29.
[97] Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. J Clin Exp Neuropsychol 2018;40:917–39.
[98] Nicholas M, Obler LK, Albert ML, Helm-Estabrooks N. Empty speech in Alzheimer's disease and fluent aphasia. J Speech Lang Hear Res 1985;28:405–10.
[99] Tomoeda CK, Bayles KA, Trosset MW, Azuma T, McGeagh A. Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task. Alzheimer Dis. Assoc. Disord.; 1996.
[100] Pistono A, et al. What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. Neuropsychologia 2019;124:133–43.
[101] Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci 2015;7:195.
[102] Lofgren M, Hinzen W. Breaking the flow of thought: increase of empty pauses in the connected speech of people with mild and moderate Alzheimer's disease. J Commun Disord 2022;97:106214.
[103] Paganelli F, Vigliocco G, Vinson D, Siri S, Cappa S. An investigation of semantic errors in unimpaired and Alzheimer's speakers of Italian. Cortex 2003;39:419–39.
[104] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis 2016;49:407–22.
[105] Meteyard L, Quain E, Patterson K. Ever decreasing circles: speech production in semantic dementia. Cortex 2014;55:17–29.
[106] Fergadiotis G, Wright HH, Green SB. Psychometric evaluation of lexical diversity indices: assessing length effects. J Speech Lang Hear Res 2015;58:840–52.
[107] Sanborn V, Ostrand R, Ciesla J, Gunstad J. Automated assessment of speech production and prediction of MCI in older adults. Appl Neuropsychol Adult 2020; 1–8.
[108] Ntracha A, et al. Detection of mild cognitive impairment through natural language and touchscreen typing processing. Front Digit Health (Irvine Calif) 2020;2:567158.
[109] Fergadiotis G, Wright HH, West TM. Measuring lexical diversity in narrative discourse of people with aphasia. 2013.
[110] Kapantzoglou M, Fergadiotis G, Auza Buenavides A. Psychometric evaluation of lexical diversity indices in spanish narrative samples from children with and without developmental language disorder. J Speech Lang Hear Res 2019;62: 70–83.
[111] Calzà L, Gagliardi G, Favretti RR, Tamburini F. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Comput Speech Lang 2021;65:101113.
[112] Rocholl JC, et al. Disfluency detection with unlabeled data and small bert models. arXiv Prepr. arXiv2104.10769. 2021.
[113] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv Prepr. arXiv1910.01108. 2019.
[114] DistilBERT. https://huggingface.co/docs/transformers/model_doc/distilbert.
[115] Toffle ME, Quattropani MC. The self in the Alzheimer's patient as revealed through psycholinguistic-story based analysis. Procedia-Social Behav Sci 2015; 205:361–72.

[116] Yu Y, Lu S, Wu Y, Wu Q, Wu J. Dementia and language bilingualism helps ward off Alzheimer's disease. In: Improv qual life dement patients through progress detect treat care; 2017. p. 107–22.
[117] Kamiloğlu RG, Fischer AH, Sauter DA. Good vibrations: a review of vocal expressions of positive emotions. Psychon Bull Rev 2020;27:237–65.
[118] Olowolayemo A, et al. Conversational analysis agents for depression detection: a systematic review. J Integr Adv Eng 2023;3:47–64.
[119] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Interspeech. vol. 5; 2005. p. 1517–20.
[120] Atmaja BT, Akagi M. On the differences between song and speech emotion recognition: effect of feature sets, feature types, and classifiers. In: 2020 IEEE region 10 conference (TENCON). IEEE; 2020. p. 968–72.
[121] Latif S, Rana R, Khalifa S, Jurdak R, Epps J. Direct modelling of speech emotion from raw speech. arXiv Prepr. arXiv1904.03833. 2019.
[122] Bahgat M, Wilson S, Magdy W. LIWC-UD: classifying online slang terms into LIWC categories. In: 14th ACM web science conference 2022; 2022. p. 422–32.
[123] Belz FF, Adair KC, Proulx J, Frankel AS, Sexton JB. The language of healthcare worker emotional exhaustion: a linguistic analysis of longitudinal survey. Front Psychiatry 2022;13:2871.
[124] O'Dea B, et al. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: a longitudinal study. PLoS One 2021;16:e0251787.
[125] Burkhardt HA, et al. Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. J Med Internet Res 2021;23:e28244.
[126] Collins SE, et al. Language-based measures of mindfulness: initial validity and clinical utility. Psychol Addict Behav 2009;23:743.
[127] Glauser T, et al. Identifying epilepsy psychiatric comorbidities with machine learning. Acta Neurol Scand 2020;141:388–96.
[128] Eyben F, Schuller B. openSMILE:) the Munich open-source large-scale multimedia feature extractor. ACM SIGMultimedia Rec 2015;6:4–13.
[129] Praat Vocal Toolkit. http://www.praatvocaltoolkit.com/.
[130] Cummins N, et al. A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. In: Interspeech 2020. ISCA-International Speech Communication Association; 2020. p. 2182–6.
[131] Zolnoori M, et al. Audio recording patient-nurse verbal communications in home health care settings: pilot feasibility and usability study. JMIR Hum Factors 2022;9:e35325.
[132] Zolnoori M, Schilling K, Jones J. Patient-centered decision support for pediatric asthma screening: a web-based interface for parents. 2022.
[133] Schneider S, Baevski A, Collobert R, Auli M. wav2vec: unsupervised pre-training for speech recognition. arXiv Prepr. arXiv1904.05862. 2019.
[134] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63:3–42.
[135] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: icml. vol. 96. Citeseer; 1996. p. 148–56.
[136] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 785–94.
[137] Ben-Hur A, Weston J. A user's guide to support vector machines. In: Data mining techniques for the life sciences. Springer; 2010. p. 223–39.
[138] Siami-Namini S, Tavakoli N, Namin AS. The performance of LSTM and BiLSTM in forecasting time series. In: 2019 IEEE international conference on big data (big data). IEEE; 2019. p. 3285–92.
[139] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611–29.
[140] Galea M, Woodward M. Mini-mental state examination (MMSE). Aust J Physiother 2005;51:198.
[141] All OCNF. Montreal cognitive assessment. Stroke 2015;46:3547–50.
[142] Buschke H, et al. Screening for dementia with the memory impairment screen. Neurology 1999;52:231.
[143] Sheehan B. Assessment scales in dementia. Ther Adv Neurol Disord 2012;5:349–58.
[144] Eating, H. & Blog-Inside, N. I. A. Assessing Cognitive Impairment in Older Patients.
[145] Rasmussen J, Langerman H. Alzheimer's disease–why we need early diagnosis. Degener Neurol Neuromuscul Dis 2019;9:123.
[146] Fraser KC, Rudzicz F, Graham N, Rochon E. Automatic speech recognition in the diagnosis of primary progressive aphasia. In: Proceedings of the fourth workshop on speech and language processing for assistive technologies; 2013. p. 47–54.
[147] Figueiredo S, Barfod V. Boston Diagnostic Aphasia Examination (BDAE). 2012.
[148] Godfrey JJ, Holliman EC, McDaniel J. SWITCHBOARD: telephone speech corpus for research and development. In: Acoustics, speech, and signal processing, IEEE international conference on. vol. 1. IEEE Computer Society; 1992. p. 517–20.
[149] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv Prepr. arXiv1810.04805. 2018.
[150] Luo Y. Recurrent neural networks for classifying relations in clinical notes. J Biomed Inform 2017;72:85–95.
[151] Colón-Ruiz C, Segura-Bedmar I. Protected health information recognition byBiLSTM-CRF. In: Proceedings of the Iberian languages evaluation forum. IberLEF; 2019.
[152] Adoma AF, Henry N-M, Chen W. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In: 2020 17th international computer conference on wavelet active media technology and information processing (ICCWAMTIP). IEEE; 2020. p. 117–21.
[153] Cortiz D. Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. arXiv Prepr. arXiv2104.02041. 2021.
[154] BERT. https://huggingface.co/docs/transformers/model_doc/bert.
[155] XLNet. https://huggingface.co/docs/transformers/model_doc/xlnet.
[156] Liu Y, et al. Roberta: a robustly optimized bert pretraining approach. arXiv Prepr. arXiv1907.11692. 2019.
[157] distilroberta-base · Hugging Face. https://huggingface.co/distilroberta-base.
[158] Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 1996;70:614.
[159] Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull 2003;129:770.
[160] Laukka P, Elfenbein HA. Emotion appraisal dimensions can be inferred from vocal expressions. Soc Psychol Personal Sci 2012;3:529–36.
[161] Moore II E, Clements MA, Peifer JW, Weisser L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Trans Biomed Eng 2007;55:96–107.
[162] Busso C, Lee S, Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans Audio Speech Lang Process 2009;17:582–96.
[163] Sundberg J, Patel S, Bjorkner E, Scherer KR. Interdependencies among voice source parameters in emotional speech. IEEE Trans Affect Comput 2011;2:162–74.
[164] Yap TF. Speech production under cognitive load: effects and classification. 2012.
[165] Steidl S. Automatic classification of emotion related user states in spontaneous children's speech. Germany: Logos-Verlag Berlin; 2009.
[166] Scherer KR, Sundberg J, Tamarit L, Salomão GL. Comparing the acoustic expression of emotion in the speaking and the singing voice. Comput Speech Lang 2015;29:218–35.
[167] Boyd RL. MEH: meaning extraction helper (Version 1.0. 6)[Software]. 2015.
[168] Sun Q, Pfahringer B. Bagging ensemble selection. In: Australasian joint conference on artificial intelligence. Springer; 2011. p. 251–60.
[169] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21.
[170] Cokluk O. Logistic regression: concept and application. Educ Sci Theory Pract 2010;10:1397–407.
[171] Murty MN, Raghava R. Kernel-based SVM. In: Support vector machines and perceptrons. Springer; 2016. p. 57–67.
[172] Wang K, Cheng L, Yong B. Spectral-similarity-based kernel of SVM for hyperspectral image classification. Remote Sens (Basel) 2020;12:2154.