RESEARCH

Machine Learning to Assess Language Impairments in Older Adults

Mah Parsa[1*], Muhammad Raisul Alam[1,2,3] and Alex Mihailidis[2,3,4]

*Correspondence:
mahparsa@cs.toronto.edu
[1]Department of Computer Science, University of Toronto, Toronto, Canada
Full list of author information is available at the end of the article

Abstract

Objectives: The main objective of this paper is to propose a methodology based on machine learning classifiers for assessing language impairments associated with dementia in older adults. To do so, we compare the impact of different types of language tasks, features, and recording media on our ML-based methodology's efficiency.

Methodology: The methodology encompasses the following steps: 1) Extracting linguistic and acoustic features from subjects' speeches which have been collected from subjects with dementia (N=9) and subjects without dementia (N=13); 2) Employing feature selection methods to rank informative features; 3) Training ML classifiers using extracted features to recognize subjects with dementia from subjects without dementia; 4) Evaluating the classifiers; 5) Selecting the most accurate classifiers to develop the languages assessment tools.

Results: Our results indicate that 1) we can find more predictive linguistic markers to distinguish language impairment associated with dementia from participants' speech produced during the picture description language task than the story recall task. 2) a phone-based recording interface provides a more high-quality language dataset than the web-based recording systems; 3) classifiers trained with selected features from acoustic features or linguistic features show higher performance than the classifiers trained with pure features.

Conclusion: Our results show that the tree-based classifiers that have been trained using the PD dataset can be used to develop an ML-based language assessment tool that can detect language impairment associated with dementia as quickly as possible.

Keywords: Alzheimer's Disease; Acoustic Features; Dementia; Language Impairments; Linguistic Features; Machine Learning; Mild Cognitive Impairment

1 Introduction

More than 50 million people worldwide are living with different types of dementia [1] including *Alzheimer's Disease* (AD). Thus, neurodegenerative dementias including AD, *Vascular Dementia* (VaD), *Lewy Body Dementia* (LBD), and *Frontotemporal Lobar Dementia* (FTD) [2] are one of the leading global neurodegenerative diseases. and have notable economic impacts on individuals and societies [3]. To mitigate the impact of neurodegenerative dementias on older adults and help older adults plan for the future and find external sources of support [4], early detection of dementia is necessary. It would help older adults at the early stages of the disease seek out different intervention programs [5], including psycho-social interventions (e.g., walking programs and art therapy) [6], non-pharmaceutical intervention programs

(e.g., music interventions [7]) as well as clinical interventions. It would help them to maintain their quality of life [8] and slow down disease progression.

We can detect dementia patients using *language assessment* (LA) tools, which have been recognized as low-cost and effective tools with high specificity and sensitivity of diagnosis [9] at the earliest stage of their disease. These tools can recognize language impairments, which are signs of the first cognitive manifestations of any types of dementia, specifically the onset of AD [10] and *mild cognitive impairment* (MCI) [1]. Using LA tools, we can also identify the different types of language impairment including difficulties with finding a relative expression, naming, and word comprehension and various level of language impairments [10]. Thus, it would be possible to detect different types of dementia and various stages of AD using the LA tools. For example, the LA tools can identify 1) lexical-semantic language problems such as naming the things or being vague in what they want to say [10], which are patients' language problems at the mild stage of AD; 2) signs of empty speech (e.g., "The thing is over there, you know"), which are patients' language problems at the moderate stage of AD; 3) phonological, morphological, syntactical problems and the lack of verbal fluency, which are signs of language impairments at the severe stage of AD.

In this paper, we propose a methodology to develop ML-based language assessment tools to detect dementia. Unlike the previous research papers [12, 13, 14, 15, 16], we have not just focus on examining classifiers to distinguish subjects with dementia from subjects without dementia, rather we have defined different experiments to understand the impact of the language tasks, types of features and recording media on the efficiency of the tool. More specifically, we seek to find out the impact of 1) different language tasks, e.g., the picture description and the story recall tasks, 2) recording media e.g., phone vs web-base, and 3) linguistic and acoustic features on the efficiency ML-based language assessment tools.

The main contribution of this paper is to propose a methodology for developing an efficient ML-based language assessment tool. Another contribution of the paper is to introduce four metrics to measure incoherence and tangential speech in elderly individuals. To the best of our knowledge, no study has investigated the efficiency of language tasks and recording media on ML-based language assessment tools' efficiency and the development of metrics to detect incoherent and tangentially in speech of the elderly individuals.

The remainder of the paper is organized as follows. Section 2 reviews various types of assessment tools for detecting dementia and compare them from two perspectives, cost-effective point and user-friendly; it also describes their advantages and limitations. Such a review highlights that ML-based language assessment tools are useful and quick tools to detect dementia in elderly individuals. Section 3 describes the methodology to develop ML-based LA tools. Section 4 presents our results. Section 5 discusses the data limitation, feature selection, validity, reliability, fairness, and explainability aspects of ML-based LA tools. Finally, Section 6 concludes the article highlighting its main contributions and our future direction.

[1] MCI refers to the condition where an older adult experiences cognitive impairment especially in tasks related to orientation and judgment [11]

2 Background

There is no single test to diagnose dementia. Thus, clinicians run different tests, including cognitive and neuropsychological, neurological, brain-imaging, laboratory tests, and psychiatric evaluation, to detect cognitive impairment associated with dementia in older adults. Thus, in this section, we review numerous assessment tools and describes how quickly, accurately, and cost-effectively they can distinguish patients with dementia. This section aims to provide useful information for ML developers who might have innovative ideas to produce ML-based assessment tools to detect dementia.

2.1 Cognitive and Neuropsychological Assessments

Different tests have been proposed to evaluate cognitive function including memory, orientation, attention, reasoning and judgment, language skills, and attention [17, 18, 19]. Examples of such tools include the *Eight-item Informant Interview to Differentiate Aging and Dementia* (AD8), *Annual Wellness Visit* (AWV), *General Practitioner Assessment of Cognition* (GPCOG) and *Health Risk Assessment* (HRA), *Memory Impairment Screen* (MIS) (i.e., it is for testing verbal memory capability), *Short Informant Questionnaire on Cognitive Decline in the Elderly* (IQCODE), *Montreal Cognitive Assessment* (MoCA) (i.e., it consists of a 30-points scales [20] to identify subjects with MCI) [21], *Addenbrooke's Cognitive Assessment* (ACE) [22, 23], and the *Alzheimer's Disease Assessment Scale–Cognitive Subscale ADAS-Cog* [24, 25] (i.e., it examines attention, orientation, memory, language, visual perception, and visuospatial skills), the test can detect cognitive impairment that is related to AD and fronto-temporal dementia. Other examples include the *Cambridge Assessment of Memory and Cognition*, the *Mini-Mental State Examination* (MMSE) (it can assess impairments related to five cognitive functions such as orientation, attention, memory, language and visual-spatial skills by just asking subjects 11 questions [17]) test (its sensitivity is around 0.79, while its specificity is around 0.95) [26], the *Saint Louis University Mental Status* (SLUMS) (i.e., it encompasses a 30-point screening questionnaire to examine not only executive function but also orientation, memory, and attention) [27] and *Cognitive Disorders Examination* (Codex) (an ML-based assessment tool that combines a decision tree with the MMSE test and the clock drawing test to diagnose dementia. Codex has shown high sensitivity and specificity in detecting dementia) [28].

The above assessment tools can detect the cognitive impairment associated with dementia by asking some determined questions. They are comfortable and available tools and can provide a quick assessment of a person's cognitive status. But each of them presents various psychometric features. Moreover, the validity of the above tests is not similar. For instance, the MoCA test provides more accurate results than the MMSE test in detecting MCI and dementia. In addition to the above testing methods, online testing has also been proposed as an appropriate alternative for traditional assessment tools. These online methods are beneficial for elderly individuals that have difficulty reaching out to clinical services. However, the main issue with these online assessment tools is that they might provide an inaccurate diagnosis that could have a negative impact on the mental health of individuals if the results indicate they might be suffering from AD. Another issue is that personal

information of individuals [29] might be revealed. In general, the benefit of using the cognitive assessment tests, as mentioned earlier, is that they are quick and cost-effective. Consequently, they are widely used by clinical services and psychiatrists [30, 31]. These tests' drawback is their inability to diagnose people with dementia with high sensitivity and specificity [26] This can be especially problematic since the false results of such tests can significantly impact health insurance and some social rights of individuals [32]. Thus, it has been suggested that such tests be considered the beginning steps of early detection programs and be combined with behavioral and LA tools [33, 26] to detect dementia quickly and accurately as possible in older adults.

2.2 Behavioural Assessment Tools

Behavioral assessment tools evaluate responsive behaviors, as well as behavioral and psychological signs to diagnose AD/MCI. One of the quickest examination to detect *Behavioural and Psychological Symptoms in Dementia* (BPSD) is *Neuropsychiatric Inventory* (NPI), [34, 35] that considers both the frequency and severity of some behaviors such as delusions, agitation, depression, and behavioral irritability [26]. Moreover, it has been stated that *Behavioral and Psychological Symptoms* (BPS) might accompany MCI that makes it useful to distinguish AD/MCI [36]. Another example of such a tool is the *Behavioral Dyscontrol Scale* (BDS), which can assess behavioral regulation in elderly adults. It is a useful discriminative tool for people with AD and MCI [37].

The main advantage of using behavioral assessment tools such as the BDS is their abilities to estimate functional independence and measure executive function skills in older adults [38]. The disadvantage of these tools is that they are not suitable tests for low executive function skills. One resolution is to use an electronic version of BDS, known as *Behavioral Dyscontrol Scale-Electronic Version* (BDS-EV). It combines the *Push-Turn-Taptap* (PTT) to evaluate action planning, action learning, and motor control speed and accuracy [39] and *motor programming* (MP) [40]. Another solution involves combining behavioral and cognitive assessment tools to provide a better way to assess AD [36]. A good example of such a tool is the *Relevant Outcome Scale for Alzheimer's Disease* (ROSA), which has 16 items with 21 points to assess the subject's cognitive function and behavioral symptoms. ROSA's main characteristic is its ability to evaluate the severity of a patient's AD by changing the assessment scenario. Another example is a tool called ABC Dementia Scale, a mix of behavioral and cognitive assessments to detect AD signs and track changes in its symptoms over time [41]. The authors of [42] have proposed a tool named as the *History-based Artificial Intelligent Clinical Dementia Diagnostic System* (HAICDDS), which combines different assessment tools such as the *Instrumental Activities of Daily Living* (IADL) Scale; *Cognitive Abilities Screening Instrument* (CASI) and MoCA. The main aim of HAICDDS is to distinguish individuals with *Visuospatial Dysfunction* (VSD) from AD and *Dementia with Lewy Bodies* (DLB).

In general, behavioral assessment tools are useful for assessing the progress of AD; however, they might not be beneficial to detect early signs of AD, mainly because most of these tests consist of symptoms rated by patient family members or caregivers.

2.3 Language Assessment Tools

These tools can detect language disorders, incoherent speech, tangentiality and grammatical error, lexical retrieval difficulties, auditory comprehension difficulties, grammatical and spelling failures [43, 44, 45] in the subjects. These signs are associated to *Language and Communication Impairment* (LCI) in patients with dementia. In particular, LA tools that analyze spontaneous speech, produced during the completion of cognitive tasks [46] can recognize linguistic features associated with language performance deficits in elderly individuals [47]. Therefore, they are efficient methods to diagnose AD/MCI in elderly adults [48]. The main advantage of using LA tools is their cost-effectiveness and user-friendliness. It is beneficial to patients and clinicians alike to use them.

2.4 Machine Learning

Discovering linguistic markers using machine learning has captured the attention of researchers in the field of neurodegenerative disease [12, 13, 14, 15, 16]. In more details, ML analyzes language produced by individuals (e.g., patients and healthy subjects) to distinguish healthy subjects from patients with dementia. So far, various ML algorithms (such as *k-Nearest Neighbor* (kNN) [13], *Support Vector Machine* (SVM), *Decision Trees* (DTs) [13, 14] and *Random Forest* (RF) classifiers [49] as well as *Deep Learning* (DL) architectures [16] have been examined to distinguish patients from healthy subjects. In more details, the ML algorithms have been trained by linguistic features to identify language performance deficits in elderly individuals [47, 13, 14, 49]. One of the earliest studies to develop such an ML algorithm was proposed using the SVM to detect voice impairments in patients with AD [12]. In another work [50], an SVM classifier was trained by language features extracted from the *DementiaBank* (DB) dataset and could achieve 80% accuracy in predicting probable AD. In [51], an SVM classifier was trained on a dataset combining datasets from DB with Talk2ME (i.e, encompass 167 patients with AD and 187 health controls), and achieved 70% accuracy. Another excellent results obtained from employing an SVM classifier on a dataset that combined DB and CCC with 15 healthy controls and 26 patients with AD. They showed that the SVM can distinguish patients and healthy controls with 75% accuracy [52, 13]. In [52, 13], the authors showed that k-NN can distinguish patients with MCI from healthy subjects with 63% accuracy, they also showed that employing Bayesian Network on the CCC dataset, we can achieve 66% accuracy.

To develop a typical ML-based language assessment tool, we generally combine the following steps: 1) Collecting language datasets or getting access to available language datasets (e.g., DB); 2) Engineering Features: 2.1) Extracting linguistic and acoustic features; 2.2) Employing various feature selection methods to select informative features; 3) Training different classifiers using multiple sets of features; 4) Selecting ML algorithms with the highest performance to be the assessment tool's basis.

2.4.1 Language Datasets

We need labelled language datasets of patients with dementia to develop supervised ML algorithms to detect language impairments in patients. So far different language

datasets, such as *Carolina Conversations Collections* (CCC) [53] and the DB [2] set [54] have been introduced. The datasets were obtained using various language tests (see Table 1) such as the *Boston Naming Test* (BNT) is one standard test to assess language performance in participants with aphasia or dementia. Deficits in naming production appear in the first stages of Alzheimer's disease and boost with time. Thus, BNT is one of the tests that can be used to detect the disease and follow its course. Moreover, it is useful in discriminating healthy elderly persons and those with dementia [55]. Another example is the *Letter Fluency Task* (LFT), which is a part of the verbal fluency test, including another test named category fluency and aims to test of verbal functioning. The above tests aim to collect language data to assess various aspects of language impairment in subjects. For example, the *Picture Description* (PD) task is usually used to evaluate the semantic knowledge in subjects [56]. Using the Cookie Theft picture for the PD task, we can assess the structural language skills [57] of patients and amplify signs of language impairment. While the *Story Recall* (SR) task[3] can help assess impairment in episodic and semantic memory and also global cognition [58].

Table 1: Different Tests/Tasks To Assess Language and Cognitive Deficit

Test/Task Name	Cognitive function and Language Deficit	Reference
MMSE	Cognitive Impairment	[59]
MoCA	Cognitive Impairment	[21]
Boston Naming Test	Lexical retrieval deficit	[60]
Letter Fluency Task	Lexical retrieval deficit	[60]
Picture Description (the Cookie Theft) Task	Lexical retrieval deficit	[60, 61]
Countdown	-	[52]
Semantic Fluency	Executive Function (Frontal Lobe) and Language Deficit (Temporal Lobe) Function	[52, 61, 62]
Sentence repeating	-	[52]
Story Recall	-	[61]
Image naming	Impaired semantic knowledge	[61, 63]
Vocabulary	-	[61]
Winograd schemas	Pragmatic deficits	[61]
Word-colour Stroop	Cognitive impairment	[61]
General disposition	-	[61]

2.4.2 Feature Engineering

One of the main steps to develop ML-based LA tools is to extract linguistic (e.g., lexical, for more examples see Tables 2) and acoustic (e.g., pitch, see Table 3 for more examples) features from raw text and audio files. These extracted features can directly be used to train ML algorithms or a set of informative features selected from them (i.e., using the feature selection methods such as ANOVA) can be prepared to develop ML algorithms.

This section aims to provide an overview of various assessment tools that are used to identify language impairments associated with Alzheimer's disease and mild cognitive impairment. Based on the overview, we believe that ML-based language assessment tools can be considered as cost-effective, user-friendly, reliable and valid

[2]the DB dataset has been collected by recording the voice of patients with AD (N=167) and healthy control (N=97) while completing a PD task.

[3]During the story recall task, participants are shown a short passage with one of the following options 1) My Grandfather, 2) Rainbow or 3) Limpy that are three well-known passage to assess memory capacity of participants.

Table 2: List of Linguistic Features reported in the Literature within AD Domain. Different linguistic features, including lexical, syntactic, semantic and pragmatic features can be associated with different types of language deficits in patients with AD and MCI [64].

Name	Type	Cognitive Function	References
Coordinated sentences	Syntactic	Syntactic processing	[50]
Subordinated sentences	Syntactic	Syntactic processing	[50]
Reduced sentences	Syntactic	Syntactic processing	[50]
Number of predicates	Syntactic	Syntactic processing	[50]
Average number of predicates	Syntactic	Syntactic processing	[50]
Dependency distance	Syntactic	Syntactic processing	[50]
Number of dependencies	Syntactic	Syntactic Processing	[50]
Average dependencies per sentence	Syntactic	Syntactic processing	[50]
Production rules	Syntactic	Syntactic processing	[50]
Noun Rate (NR)	Syntactic	Cognitive strength	[65, 13]
Pronoun Rate (PR)	Syntactic	Cognitive strength	[65, 13]
Adjective Rate (AR)	Syntactic	Cognitive strength	[65, 13]
Verbal Rate (VR)	Syntactic	Cognitive strength	[65, 13]
Utterances	Lexical	the linguistic strength	[50]
Function words	Lexical	—	[50]
Word count	Lexical	—	[50]
Character length	Lexical	—	[50]
Total sentences	Lexical	—	[50]
Unique words	Lexical	language processing	[50]
Repetitions	Lexical	—	[50]
Revisions	Lexical	—	[50]
Morphemes	Lexical	—	[50]
Trailing off indicator	Lexical	—	[50]
Word replacement	Lexical	—	[50]
Incomplete words	Lexical	—	[50]
Filler words	Lexical	—	[50]
Type token ration (TTR)	Semantic	Vocabulary Richness	[65, 13]
Brunet's index (BI)	Semantic	Vocabulary Richness	[65, 13]
Honore's statistics (HS)	Semantic	Vocabulary Richness	[65, 13]
Fillers	Pragmatic	Cognitive Lapse	[65, 13]
GoAhead utterances	Pragmatic	Cognitive functionality	[65, 13]
Repetitions	Pragmatic	Cognitive Lapse	[65, 13]
Incomplete words	Pragmatic	Cognitive lapse	[65, 13]
Syllables Per Minute	Pragmatic	Cognitive impairment	[65, 13]

tools to detect language impairment in the older adults. Therefore, in the next section, we describe our methodology to develop an accurate and quick ML-based language assessment tool.

3 Methodology

This section describes our methodology for developing an ML-based LA tool. We consider the sequential following steps: 1) Extracting linguistic and acoustic features; 2) Employing feature selection methods such as *Variance Threshold* (VT) or *Minimal Redundancy Maximal Relevance Criterion* (mRMR) to select informative features. 3) Training various classifiers such as SVM, DTs, Extra Trees (ETs) by features.

3.1 Dataset

The datasets of this paper is a small audio and text language dataset extracted from our database (Language data recorded using a web or phone interface), named Talk2Me [4]. The datasets contain language data of patients (N=9) with various types of dementia (patients have been diagnosed by physician from three hospitals in Toronto) as well as healthy controls (N=13). In more details, textual and audio

[4] each subject has signed a consent form that has been provided approved by the Research Ethics Board protocol 31127 of the University of Toronto

Table 3: List of Acoustic Features reported in the Literature within AD Domain.

Type	Name	References
Cepstral Coefficients	Mean of MFCCs	[66, 67]
	Kurtosis of MFCCs	[66, 67, 68]
	Skewness of MFCCs	[66, 67, 68]
Pauses and fillers	Total duration of pauses	[66, 69, 52]
	Mean duration of pauses	[66, 69, 52, 70]
	Median duration of pauses	[52, 52]
	SD of the duration of pauses	[52]
	Long and short pause counts	[66, 69]
	Pause to word ratio	[66, 67, 69, 52, 70]
	Percentage of voiceless segments	[71]
	Fillers (um, uh)	[66, 13, 69]
Pitch and Formants	Mean of F0, F1, F2, F3	[66]
	Variance of F0, F1, F2, F3	[66]
	Mean, SD, Max and Min of F0	[71]
Aperiodicity	Jitter	[66, 71]
	Shimmer	[66, 71]
	Recurrence rate	[66]
	Recurrence period density entropy	[66]
	Determinism	[66]
	Length of diagonal structures	[66]
	Laminarity	[66]
Temporal aspects of the speech sample	Total duration	[66, 71]
	Phonation time	[71]
	Speech rate, syllable/s	[71]
	Articulation rate, syllable/s	[71]
Others	Zero-crossing rate	[66]
	Autocorrelation	[66]
	Linear prediction coefficients	[66]
	Transitivity	[66]

responses were collected from participants using a variety of language tasks such as the PD (e.g., the Cookie Theft (see Figure 1) or the Picnic Scene (see Figure 2)) and SR tasks.

3.2 Extracting Linguistic and Acoustic Features

3.2.1 Linguistic Features

We extract different linguistic features (e.g., the lexical diversity) from textual data using the *Natural Language Toolkit* (NLTK) [73]. The linguistic features of this paper is can be divided into three categories: 1) Lexical features (e.g.,lexical richness); 2) Syntactic features (e.g., part-of-speech (POS)); 3) Semantic features.

Lexical Features Since dementia can influence on the lexical richness aspect of patients' language, different studies have proposed different types of lexical features as markers of language impairment in patients with AD/MCI. For example, in [50], utterances, word count, character length, total sentences, unique words, repetitions, revisions, morphemes, incomplete words, filler words, trailing off indicator and word replacement extracted as lexical features. However, in our study, we have extracted multiple features such as *Brunet' s index* (BI) (see Equation 1) and *Honor's Statistic* (HS) with Equation 2 [74] to measure the lexical richness. In Equations 1 and 2, w and u are the total number of word tokens and the total number of unique word types, respectively. There are five readability scores namely the *Flesch-Kincaid* (F_K) (see Equation 3), the *Flesch Reading-Ease* (FRES) Test (see Equation 4) [75], *Gunning fog index* (GFI)[76] (see Equation 5), SMOG grading [77] (see Equation 6)

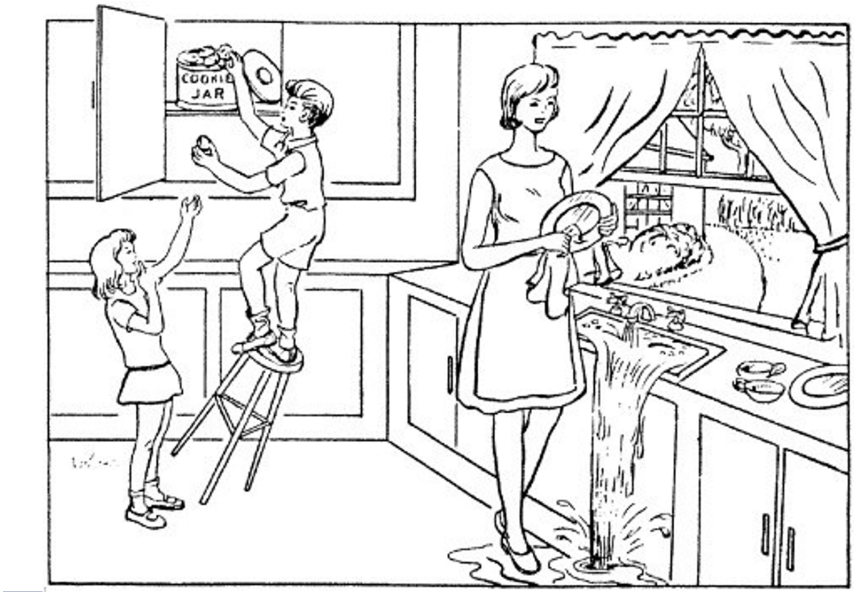


Figure 1: The Cookie Theft Picture from the Boston Diagnostic Aphasia Examination. For the PD task, the examiner asks subjects to describe the picture by saying, "Tell me everything you see going on in this picture". Then subjects might say, "there is a mother who is drying dishes next to the sink in the kitchen. She is not paying attention and has left the tap on. As a result, water is overflowing from the sink. Meanwhile, two children are attempting to make cookies from a jar when their mother is not looking. One of the children, a boy, has climbed onto a stool to get up to the cupboard where the cookie jar is stored. The stool is rocking precariously. The other child, a girl, is standing next to the stool and has her hand outstretched ready to be given cookies"[57]

and the Dale–Chall readability formula [78] (see Equation 7) to test the readability of the transcripts. Here, s and syl indicate the total number of sentences and the total number of syllables, respectively. In Equations 5, 6 and 7, c_w, P_syl and d_w indicate numbers of complex words, polysyllables and difficult words, respectively. In this paper, we have just considered two readability scores, Flesch-Kincaid (F_K) and FRES test.

$$BI = w^{(u^{-0.165})} \tag{1}$$

$$HS = \frac{100 \log w}{1 - \frac{w}{u}} \tag{2}$$

$$F_K = 0.39 \left(\frac{w}{s}\right) + 11.8 \left(\frac{SYL}{w}\right) - 15.59 \tag{3}$$

$$FRES = 206.835 - 1.015 \left(\frac{w}{s}\right) - 84.6 \left(\frac{SYL}{w}\right) \tag{4}$$

Figure 2: The the picnic scene from The Arizona Alzheimer's Disease Center (ADC) [72]. A normal subject might say: A family outing at a lake shore showed people doing sev- eral things. Mom and Dad sat on a blanket while dad read a book. Dad was over comfortable without his shoes, while mom listened to the radio and poured herself a cup of coffee. Junior was having fun flying his kite, and the family dog was interested in what all was going on. An- other of the family was spending quiet time and fisher- man, and another was playing in the shallow water. Other friends waved to them as they sailed by. It was a perfect day with just enough wind to move the flag and provide lift for the kite. It must have been comfortable sitting un- der the shade tree

$$GFI = 0.4\left[\left(\frac{w}{s}\right) + 100\left(\frac{c\_w}{w}\right)\right] \tag{5}$$

$$SMOG = 1.0430\sqrt{w \times \frac{30}{\mathrm{s}}} + 3.1291 \tag{6}$$

$$Dale\text{–}Chall = 0.1579\left(\frac{d\_w}{w} \times 100\right) + 0.0496\left(\frac{w}{s}\right) \tag{7}$$

Syntactic Features We have also extracted syntactic features such as part-of-speech (POS) ratios: 1) third pronouns (3rd-pron-pers) to proper nouns (prop); 2) first pronouns (1st-pron-pers) to pronouns (1st-pron-pers) [5]; 3) nouns to verbs; and 4) *subordinate* (SUB) to *coordinate* (CO) [61] to calculate syntactical error in speech, which is indicative of *frontotemporal dementia* (FTD) [80], and also propositional and content density equations 8 and 9 to quantify the syntax complexity. Here, NN, VB, JJ, RB, IN, and CC are the number of nouns, verbs, adjectives, adverbs, prepositions, and conjunctions respectively.

$$density_p = \frac{VB + JJ + RB + IN + CC}{N} \tag{8}$$

$$density_c = \frac{NN + VB + JJ + RB}{N} \tag{9}$$

[5] People with dementia may use first person singular pronouns than physicians perhaps as a way of focusing attention on their perspective [79]

Semantic-based Features Patients with dementia cannot easily retrieve semantic knowledge, reflecting a semantic decline in their language [52]. To develop a tool that can detect semantic decline and also incoherent speech, tangentiality, we suggest training ML algorithms using extracted semantic-based features, which are referred to as incoherent and tangential metrics in this paper. The incoherence metrics are extracting by calculation themilarity (Equation 10) between sentence embeddings: v_{s_j}. Various sentence embeddings such as simple average (SA)[6](see Equation 12), or smooth inverse frequency (SIF) embeddings [7] [81] (see Equation 13) and *term frequency-inverse document frequency* (tf-IDF) (see Equation 14). We have also calculated a tangential metric employing *latent dirichlet allocation* (LDA) [82] [83] (see Equation 15). Using the tangential metric, we can measure tangentiality in speech of patients with dementia.

$$Similarity_{\mathbf{SA}}(v_{s_i}, v_{s_j}) = \frac{v_{s_i} \cdot v_{s_j}}{\|v_{s_i}\|\|v_{s_j}\|} \tag{10}$$

$$Similarity_{\mathbf{SIF}}(v_{s_i}, v_{s_j}) = 1 - \frac{v_{s_i} \cdot v_{s_j}}{\|v_{s_i}\|\|v_{s_j}\|} \tag{11}$$

$$Incoherence_{\mathbf{SA}} = \min_i \max_j \frac{Similarity_{\mathbf{SA}}(v_{s_i}, v_{s_j})}{abs(i-j)+1} \tag{12}$$

$$Incoherence_{\mathbf{SIF}} = \min_i \sum_j \frac{Similarity_{\mathbf{SIF}}(v_{s_i}, v_{s_j})}{abs(i-j)+1} \tag{13}$$

$$Incoherence_{\mathbf{TFIDF}} = \min_i \sum_j \frac{Similarity_{\mathbf{TFIDF}}(v_{s_i}, v_{s_j})}{abs(i-j)+1} \tag{14}$$

We measured tangential speech using Equation 15. Here, N_{topic}, is the optimal number of topics for a corpus made of interview of subjects.

$$Tangentiality = 1 - \frac{N_{topic}}{\sum_j N_{topic}} \tag{15}$$

3.2.2 Acoustic Features

We extract the acoustic features using the *COre Variable Feature Extraction Feature Extractor* (COVFEFE) tool [61]. We consider 37 acoustic features and their mean, *standard deviation* (std), skewness (skew) (lack of symmetry of a data distribution) and kurtosis (kurt) (measure of peakedness around the mean of a data distribution) which results in a total of 148 features. We also include the deltas of these 148 features. Therefore, our feature selection methods consider 296 features in total. For example, we consider mean, std, skew and kurt of an MFCC feature (described later) and its deltas. Thus from a single acoustic feature, we extract 8 additional features.

[6]SA provides sentence embedding by averaging generated word embeddings from text files.

[7]SIF provides sentence embedding by calculating the weighted average of word embeddings and removing their first principal component

Table 4: List of Acoustic Features that are Considered in this Research

Type	Name	Functional	# of Features
Spectral Features	MFCCs 0 - 14	mean, kurt, skew, std	60
	Δ MFCCs 0 - 14	mean, kurt, skew, std	60
	log Mel freq 0 - 7	mean, kurt, skew, std	32
	Δ log Mel freq 0 - 14	mean, kurt, skew, std	32
	LSP freq 0 - 7	mean, kurt, skew, std	32
	Δ LSP freq 0 - 7	mean, kurt, skew, std	32
Phonation and Voice Quality Features	F0	mean, kurt, skew, std	4
	Δ F0	mean, kurt, skew, std	4
	Jitter local	mean, kurt, skew, std	4
	Δ Jitter local	mean, kurt, skew, std	4
	Jitter DDP	mean, kurt, skew, std	4
	Δ Jitter DDP	mean, kurt, skew, std	4
	Shimmer	mean, kurt, skew, std	4
	Δ Shimmer	mean, kurt, skew, std	4
	Loudness	mean, kurt, skew, std	4
	Δ Loudness	mean, kurt, skew, std	4
Speech Features	Voicing prob.	mean, kurt, skew, std	4
	Δ Voicing prob.	mean, kurt, skew, std	4

We follow the same procedure to extract all 296 features. We divide our features in 3 groups: 1) Spectral Features, 2) Phonation and Voice Quality Features, and 3) Speech Features. Table 4 shows the list of features that we consider in the research. In this section, we only describe the features that are identified as meaningful by our feature selection methods.

Spectral Features We consider the features derived from the *Mel Frequency Cepstrum* (MFC) and the *Line Spectral Pairs* (LSPs) to develop our ML classifiers. MFC uses the Mel scale to represent short-term power spectrum of a sound. *Mel Frequency Cepstral Coefficients* (MFCCs) represent energy variations between frequency bands of a speech signal and are effectively used for speech recognition and speaker verification. MFCCs aim at accurately representing the phonemes articulated by speech organs (tongue, lips, jaws, etc.). Delta MFCCs are the trajectories of the MFCCs over time. The logarithm of Mel filter banks are calculated as an intermediate step of computing MFCCs and we consider the *Log Mel Frequency Bands* and the *Delta Log Mel Frequency Bands* as spectral features. Previous research identified the MFCCs as one of the most relevant acoustic features to distinguish patients with different types of dementia [66, 67, 68]. Our analysis also confirm this claim (see Tables 6 and 9).

LSPs are strongly related to underlying speech features and are thus useful in speech coding [84]. They are correlated to unvoiced speech, pause and silence which are reportedly effective in identifying linguistic impairments [85]. The delta of LSPs represents the change of LSPs over time. Our feature selection methods confirm the importance of LSPs and their deltas (see Tables 6 and 9).

Phonation and Voice Quality Features This feature group includes *fundamental frequency* (F0), shimmer, jitter, loudness and the deltas of these features. The F0 feature is defined as the rate of oscillation of the vocal folds [86]. F0 is nearly periodic in speech of the healthy people but less periodic in patients [87]. Jitter describes frequency instability and shimmer is a measure of amplitude fluctuations. Loudness affects the amplitude of vibrations and it is correlated to the emotional states of the

speaker [88]. Previous studies reported that phonation and voice quality features are correlated with MCI and AD [89, 90] and our findings also support these claims (see Tables 6 and 9).

Speech Features We consider the voicing probability and the delta of voicing probability as relevant acoustic features. A voicing probability shows the percentage of unvoiced and voiced energy in a speech signal. A delta voicing probability indicates the rate of change over time. Our feature selection methods identified that mean, std and kurt of both features are discriminative features to identify older adults living with dementia (see Table 9).

4 Results

This section describes the results obtained from training different ML algorithms on various language features extracted from subjects' speech during two assessment tasks: the PD and SR tasks. Our results show a comparison of the performance of the classifiers that are trained using the SR and PD task datasets and the datasets were collected using phone-based and web-based interfaces. The results obtained from these two experiments allows us to extend our analyses further and verify the language task's impact and recording media on the classifiers. Note that, we have trained the classifiers separately with linguistic and acoustic features, and in this section, we compare the performance of the classifiers developed with these two groups of elements.

4.1 Language Tasks

This subsection investigates the impact of the two language tasks: the PD and SR tasks on the performance of classifiers at the subject level. These two tasks assess different cognitive characteristics of patients with dementia, and thus it is worth investigating and comparing the effectiveness of these two language tasks.

Table 5: Statistics about our textual datasets

DATA	Ave Sentence	Std Sentence	Ave Word	Std Word
The PD Task	9.0	4.4	153.5	97.92
The SR Task	6.79	4.00	57.07	26.91
Recording Media (Phone)	3.5	4.66	74.0	44.90
Recording Media (Web)	2.27	1.25	65.59	31.11

4.1.1 The PD Task

We investigate the efficacy of linguistic (see Figure 3) and acoustic features (see Table 6), which have been extracted from the speech of the subjects without dementia (N=3) and subjects with dementia (N=5) during completing the PD task to develop an ML-based assessment tool for dementia. To do so, we train different ML algorithms with a various set of features.

Classifiers with Linguistic Features This section presents our results obtained from training various ML algorithms using lexical, semantic and syntactic features. The features have been extracted from the textual datasets. Our results show (see Figure

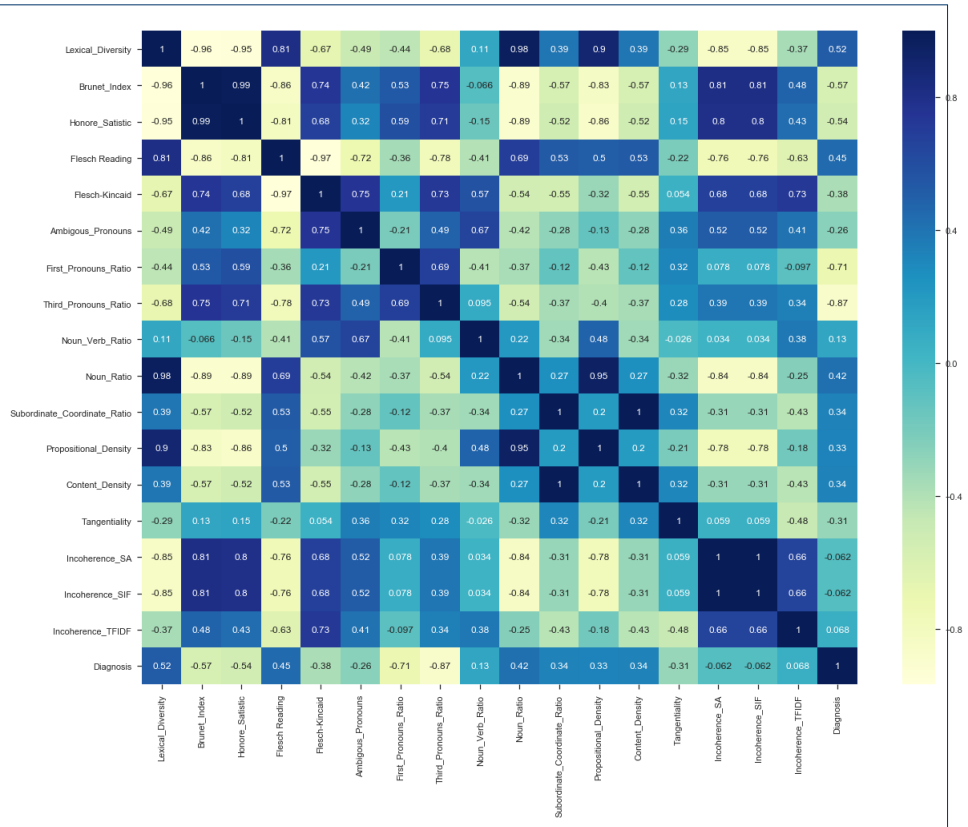
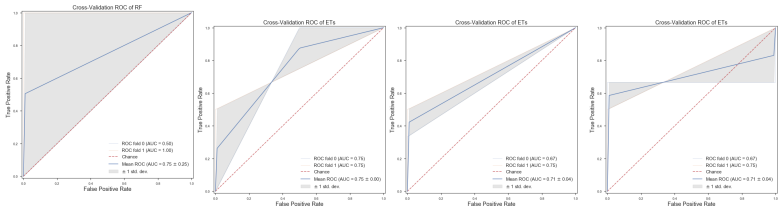


Figure 3: Correlation heat-map between 17 linguistic features.

4) that if we train the ETs algorithm with a set of lexical features, it can achieve more accurate classification results than other ML algorithms on classifying subjects with dementia from those without dementia. Training ML algorithms using the set of lexical, semantic and syntactic features decreases the accuracy of classifiers. By training the ML classifiers using 8 Syntactic features, we observed the ETs algorithm could classify the classes with an accuracy of 63.0% (+/-7%). By training various ML classifiers using 4 semantic features, we observed ETs provide more accurate results than others and could classify the classes with an accuracy of 63.0% (+/-7%) (see Figure 4.(c)).

Training ML algorithms with 3 principle components (see Figures 10.(a) and 10.(b)) extracted from 17 features, we observed that the SVM algorithm with the linear kernel could classify with 63.0% (+/- 7%)accuracy. Furthermore, among lexical features, two Flesch-Kincaid (CV[8]=23.17%, p_value=0.25) and Flesch-Reading-Ease (CV=15.15%, p_value=0.35) can provide better discrimination between these two groups of subjects, while the number of the third pronouns (the effect size equals to 1.319) and the first pronouns (the effect size equals to 2.198) among subjects without dementia has higher value than subjects with dementia. Thus, these two syntactic features can be considered as markers to detect subjects with mild cognitive impairments. Another interesting result is that measuring tangentiality

[8]coefficient of variation



(a) ROC curves of RF trained by 17 linguistic features

(b) ROC curves of ETs by lexical features

(c) ROC curves of ETs trained by syntactic features

(d) ROC curves of ETs trained by semantic features

Figure 4: ROC curves of ETs trained by different sets of linguistic features

(see Figure 9) (with the effect size of 0.020) in speech can provide a better understanding to determine subjects with dementia from healthy subjects.

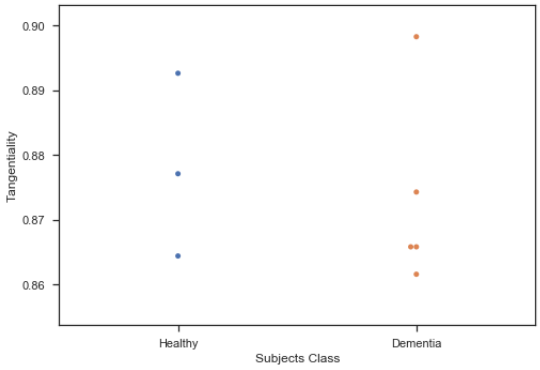
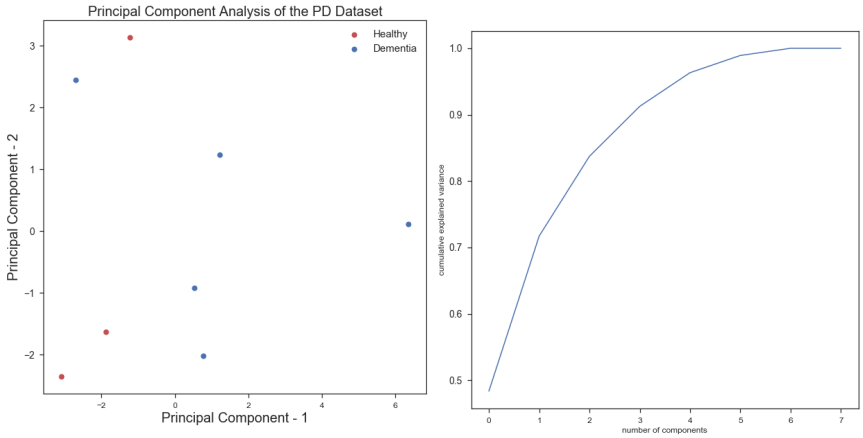


Figure 5: A comparison between the tangentiality measure for subjects with and without dementia.

Classifiers with Acoustic Features Table 8 presents the classification results obtained by applying ML tools on the extracted features from the audio files. The ML classifiers are trained using the spectral (e.g., MFCC, LSP), speech (e.g., voicing probability), phonation (e.g., F0) and voice quality (e.g., jitter, shimmer) feature as described in Section 3.2.2. We rank all these features using ANOVA, RF and mRMR methods and use the top 8 features identified by each of these methods to train the ML classifiers. Table 6 shows the top common acoustic features ranked by the above mentioned feature selection methods. We found that scikit-learn's [91] default configurations work fine for the considered ML classifiers. Therefore, we use the default configurations for all classifiers. The F1 micro scores in Table 8 are obtained using the 3-fold cross-validation method. Our results show that the tree-based classifiers, e.g., RF, ET, and DT provide better performance than others.

4.1.2 The SR Task

This section presents the results obtained by training different ML classifier using linguistic and acoustic features extracted from language data produced subjects without dementia (N=10) and subjects with dementia (N=4) during the SR task.



(a) It shows that subjects with dementia and healthy controls cannot be linearly separated using 2 principle components.

(b) It presents the values of cumulative explained variance for different number of principle components.

Figure 6

Table 6: Common acoustic features obtained by applying ANOVA, RF and mRMR feature selection methods on the recorded audio files of the PD and SR tasks

PD Task	SR Task
MFCC 13 (mean)	MFCC 12,13 (skew)
MFCC 12,13 (kurt)	Δ MFCC 3,4,6,13 (mean)
MFCC 10,13 (skew)	Δ MFCC 4 (skew)
Δ MFCC 2,11 (mean)	Δ LSP freq 7 (mean)
Δ MFCC 2,3,6,7 (kurt)	Δ LSP freq 3,6 (kurt)
Δ MFCC 6,11,13 (skew)	Loudness (kurt, skew)
Δ LSP freq 3,5 (mean)	Δ Loudness (kurt, skew)
Δ LSP freq 2,6 (kurt)	F0 (kurt)
Δ LSP freq 1 (skew)	Δ F0 (mean)

Classifiers with Linguistic Features This section examines the efficiency of using different linguistic features to train ML classifiers. Using 5 lexical features to train classifiers, the SVM (with the rbf kernel and C=0.01) and RF ($n\_estimators$=2 and max_depth=2), can classify subjects with dementia and healthy subjects accurately with 71% accuracy. We can get the same results using 8 syntactic features to train the SVM (with the rbf kernel and C=0.01) and RF($n\_estimators$=2 and max_depth=2). If we train classifiers (3-fold Cross-Validation) with 17 lexical, semantic, and syntactic features, the SVM (with the rbf kernel and C=0.01) can classify subjects with dementia and healthy subjects accurately with 72% accuracy (see Figure 7.(a)). Training ML algorithms with 3 principle components (see Figures 10.(a) and 10.(b)) extracted from 17 features, we observed that the SVM algorithm with the rbf kernel could classify with 71% accuracy.

Classifiers with Acoustic Features In this section, we present the classification results obtained by applying ML algorithms on the features extracted from the audio files. We use audio data collected from subjects without dementia ($N = 10$) and subjects with dementia ($N = 4$). We follow the same approach that we use in Section 4.1.1 to rank the acoustic features and use the top 15 features to develop the

Table 7: F1 (micro) scores obtained by applying ML algorithms on linguistic features

Features	Algorithms	PD Task	SR Task	Web	Phone
Lexical	DT	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.42 (+/- 0.17)	0.92 (+/- 0.17)
	ET	**0.73 (+/- 0.13)**	0.57 (+/- 0.57)	0.83 (+/- 0.00)	0.92 (+/- 0.17)
	kNN	0.52 (+/- 0.29)	0.42 (+/- 0.00)	0.45 (+/- 0.00)	0.45 (+/- 0.00)
	LDA	0.63 (+/- 0.07)	0.63 (+/- 0.07)	0.75 (+/- 0.17)	0.92 (+/- 0.17)
	R_SVM	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	L_SVM	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	1.00 (+/- 0.00)
	LR	0.63 (+/- 0.07)	0.63 (+/- 0.07)	0.83 (+/- 0.00)	1.00 (+/- 0.00)
	RF	0.47 (+/- 0.27)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.92 (+/- 0.17)
Syntactic	DT	0.73 (+/- 0.13)	0.57 (+/- 0.00)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	ET	0.80 (+/- 0.40)	0.64 (+/- 0.14)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	kNN	0.69 (+/- 0.63)	0.53 (+/- 0.23)	0.45 (+/- 0.00)	0.45 (+/- 0.00)
	LDA	0.37 (+/- 0.07)	0.50 (+/- 0.43)	0.75 (+/- 0.17)	0.75 (+/- 0.50)
	R_SVM	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	L_SVM	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.67 (+/- 0.33)
	LR	0.80 (+/- 0.40)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.75 (+/- 0.17)
	RF	0.47 (+/- 0.27)	0.57 (+/- 0.00)	0.75 (+/- 0.17)	0.92 (+/- 0.17)
Semantic	DT	0.53 (+/- 0.27)	0.64 (+/- 0.14)	0.83 (+/- 0.00)	0.83 (+/- 0.33)
	ET	0.57 (+/- 0.47)	0.71 (+/- 0.29)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	kNN	0.69 (+/- 0.63)	0.53 (+/- 0.23)	0.45 (+/- 0.00)	0.45 (+/- 0.00)
	LDA	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.58 (+/- 0.50)
	R_SVM	0.63 (+/- 0.07)	**0.71 (+/- 0.00)**	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	L_SVM	0.63 (+/- 0.07)	0.71 (+/- 0.00)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	LR	0.63 (+/- 0.07)	0.50 (+/- 0.43)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	RF	**0.73 (+/- 0.13)**	0.57 (+/- 0.00)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
All	DT	**0.73 (+/- 0.13)**	0.64 (+/- 0.14)	0.75 (+/- 0.17)	1.00 (+/- 0.00
	ET	0.63 (+/- 0.07)	**0.79 (+/- 0.14)**	0.83 (+/- 0.00)	0.75 (+/- 0.50)
	kNN	0.52 (+/- 0.29)	0.39 (+/- 0.05)	0.45 (+/- 0.00)	0.45 (+/- 0.00)
	LDA	0.63 (+/- 0.07)	0.64 (+/- 0.14)	0.75 (+/- 0.17)	0.75 (+/- 0.50
	R_SVM	0.63 (+/- 0.07)	0.71 (+/- 0.00)	0.83 (+/- 0.00)	0.83 (+/- 0.00)
	L_SVM	0.63 (+/- 0.07)	0.64 (+/- 0.14)	0.75 (+/- 0.17)	1.00 (+/- 0.00)
	LR	0.70 (+/- 0.60)	0.64 (+/- 0.14)	0.83 (+/- 0.00	0.75 (+/- 0.17)
	RF	0.63 (+/- 0.07)	0.71 (+/- 0.00)	0.83 (+/- 0.00)	1.00 (+/- 0.00)

classifiers. Table 6 shows the top common acoustic features provided by ANOVA, RF and mRMR feature selection methods. We use scikit-learn's default configurations for all classifiers of this sub-section. The F1 micro scores in Table 8 are obtained using the 3-fold cross-validation method. Our results show that the ET classifiers outperform others.

4.1.3 Comparison between the PD and the SR Tasks

We evaluate the impact of language tasks on detecting AD and MCI. We use audio recordings and transcribed textual datasets to extract linguistic and acoustic features from PD and SR tasks. Our datasets are imbalanced and therefore micro F1 scores are more appropriate to report the performance of the ML classifiers. To asses the efficiency of PD and SR tasks, we calculate a range of F1 scores using different feature sets and classifiers as shown in Tables 7 and 8. We use lexical, syntactic, semantic and combination of all these 3 feature groups as linguistic features. For acoustic features, we use ANOVA, RF and mRMR feature selection methods. We also use the common features in these 3 feature selection methods as another set of acoustic features. Finally, we apply DT, ET, Linear SVM, RBF SVM, LDA, LR, kNN and RF algorithms to compute the F1 scores. Figure 11(a) shows the distributions of F1 scores for PD and SR tasks. A one-way ANOVA test performed on the F1 scores of the PD and SR tasks shows that the means are significantly different ($F(1,126) = 8.27$, $p = 0.005$). A Tukey's post-hoc test shows that the mean

Table 8: F1 (micro) scores obtained by applying ML algorithms on acoustic features

Features	Algorithms	PD Task	SR Task	Web	Phone
ANOVA	DT	0.83 (+/- 0.24)	0.50 (+/- 0.24)	**0.89 (+/- 0.16)**	0.81 (+/- 0.02)
	ET	0.98 (+/- 0.03)	**0.86 (+/- 0.09)**	0.83 (+/- 0.24)	0.93 (+/- 0.09)
	kNN	0.83 (+/- 0.24)	0.78 (+/- 0.02)	**0.89 (+/- 0.16)**	0.93 (+/- 0.09)
	LDA	0.89 (+/- 0.16)	0.70 (+/- 0.14)	**0.89 (+/- 0.16)**	**1.00 (+/- 0.00)**
	R_SVM	0.72 (+/- 0.21)	0.78 (+/- 0.02)	**0.89 (+/- 0.16)**	0.76 (+/- 0.06)
	L_SVM	0.83 (+/- 0.24)	0.70 (+/- 0.14)	0.72 (+/- 0.21)	**1.00 (+/- 0.00)**
	LR	0.83 (+/- 0.24)	0.78 (+/- 0.02)	0.72 (+/- 0.21)	0.93 (+/- 0.09)
	RF	**0.99 (+/- 0.02)**	0.83 (+/- 0.06)	0.83 (+/- 0.24)	0.93 (+/- 0.09)
RF	DT	0.72 (+/- 0.21)	0.57 (+/- 0.17)	**1.00 (+/- 0.00)**	0.87 (+/- 0.09)
	ET	0.99 (+/- 0.02)	0.80 (+/- 0.04)	**1.00 (+/- 0.00)**	0.99 (+/- 0.02)
	kNN	0.89 (+/- 0.16)	0.78 (+/- 0.02)	0.89 (+/- 0.16)	0.93 (+/- 0.09)
	LDA	**1.00 (+/- 0.00)**	0.57 (+/- 0.17)	0.89 (+/- 0.16)	0.93 (+/- 0.09)
	R_SVM	0.61 (+/- 0.08)	0.78 (+/- 0.02)	0.61 (+/- 0.08)	0.76 (+/- 0.06)
	L_SVM	0.89 (+/- 0.16)	0.78 (+/- 0.02)	0.72 (+/- 0.21)	**1.00 (+/- 0.00)**
	LR	0.89 (+/- 0.16)	**0.87 (+/- 0.09)**	0.72 (+/- 0.21)	0.93 (+/- 0.09)
	RF	**1.00 (+/- 0.00)**	0.78 (+/- 0.02)	0.90 (+/- 0.14)	**1.00 (+/- 0.00)**
mRMR	DT	**1.00 (+/- 0.00)**	0.70 (+/- 0.14)	0.83 (+/- 0.24)	0.87 (+/- 0.09)
	ET	**1.00 (+/- 0.00)**	**0.81 (+/- 0.05)**	0.97 (+/- 0.05)	**1.00 (+/- 0.00)**
	kNN	0.50 (+/- 0.14)	0.78 (+/- 0.02)	**1.00 (+/- 0.00)**	0.81 (+/- 0.16)
	LDA	**1.00 (+/- 0.00)**	0.77 (+/- 0.21)	**1.00 (+/- 0.00)**	**1.00 (+/- 0.00)**
	R_SVM	0.61 (+/- 0.08)	0.78 (+/- 0.02)	0.72 (+/- 0.21)	0.76 (+/- 0.06)
	L_SVM	0.78 (+/- 0.31)	0.50 (+/- 0.08)	**1.00 (+/- 0.00)**	0.87 (+/- 0.19)
	LR	0.78 (+/- 0.31)	0.78 (+/- 0.02)	**1.00 (+/- 0.00)**	0.87 (+/- 0.19)
	RF	0.99 (+/- 0.02)	0.78 (+/- 0.02)	0.88 (+/- 0.16)	**1.00 (+/- 0.00)**
Common	DT	**1.00 (+/- 0.00)**	0.52 (+/- 0.37)	**1.00 (+/- 0.00)**	0.80 (+/- 0.28)
	ET	**1.00 (+/- 0.00)**	**0.84 (+/- 0.11)**	0.74 (+/- 0.21)	0.94 (+/- 0.08)
	kNN	0.83 (+/- 0.24)	0.70 (+/- 0.14)	0.89 (+/- 0.16)	**1.00 (+/- 0.00)**
	LDA	0.78 (+/- 0.16)	0.80 (+/- 0.16)	0.89 (+/- 0.16)	0.87 (+/- 0.19)
	R_SVM	0.72 (+/- 0.21)	0.78 (+/- 0.02)	0.78 (+/- 0.16)	0.76 (+/- 0.06)
	L_SVM	0.83 (+/- 0.24)	0.77 (+/- 0.21)	0.72 (+/- 0.21)	**1.00 (+/- 0.00)**
	LR	0.83 (+/- 0.24)	0.70 (+/- 0.14)	0.72 (+/- 0.21)	**1.00 (+/- 0.00)**
	RF	0.98 (+/- 0.02)	0.78 (+/- 0.02)	0.81 (+/- 0.20)	0.95 (+/- 0.07)

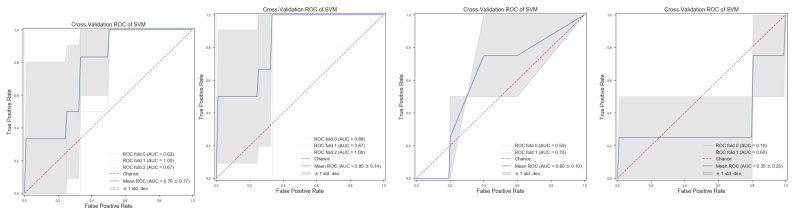
F1 scores of the PD task are higher than the SR task ($p = 0.005$), i.e., the ML classifiers developed by using PD stimuli perform better than the SR stumuli.

4.2 Recording Media

In this section, we are interested in figuring out the direct effect of using web-interface or phone-interface on the quality of recorded language data that indirectly affect on the accuracy of ML-based Language Assessment Tools.

Classifiers with Linguistic We trained various ML classifiers using 15 linguistic features extracted from recorded language data (12 samples, 10 samples related to subjects without dementia and 2 samples related to subjects with dementia) using the phone-interface. Using 5 lexical features, the SVM (with the linear kernel) classifier and the logistic regression (LR) can classify samples with 99.9% accuracy. However, using 8 syntactic features, we drop all ML classifiers' performance, including the SVM (with the linear kernel); thus, the SVM can determine subjects with dementia with 83% accuracy. Similarly, if we use 4 semantic features to train classifiers, they can provide better performance than using 8 syntactic features.

Classifiers with Acoustic Features We develop the classifiers using the acoustic features extracted from the the audio files. We use 16 phone-based recordings from 3 healthy adults and 1 dementia patient (each participant attended 4 sessions). Similarly, we consider 8 web-based recordings from subjects with dementia ($N = 3$) and



(c) ROC curves of SVM trained by 17 features

(d) ROC curves of SVM trained by lexical features

(e) ROC curves of SVM trained by syntactic features

(f) ROC curves of SVM trained by 4 semantic features

Figure 7: ROC curves of SVM trained by different sets of linguistic features

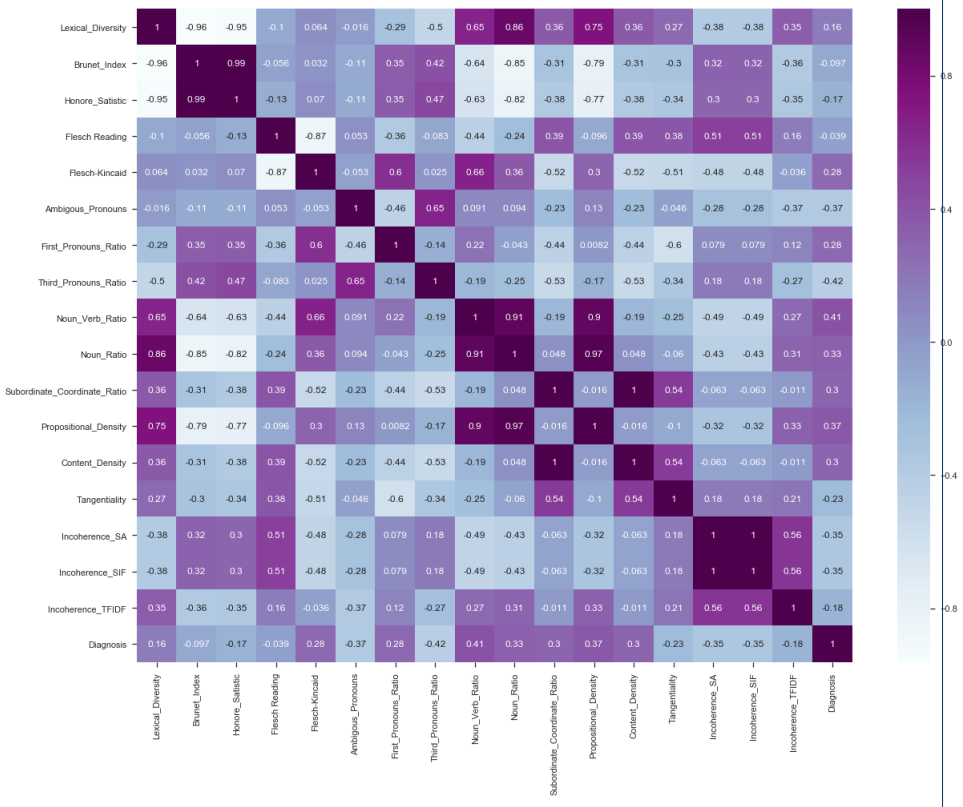


Figure 8: The correlation between different features

subjects with dementia ($N = 5$) (only 1 session each). We follow the same approach to rank the acoustic futures as we described before for the acoustic features. Table 9 shows the common features ranked by ANOVA, RF and mRMR methods. We use the top 15 features to train the classifiers. Table 8 shows the F1 scores obtained from the DT, ET, Linear SVM, RBF SVM, LDA, LR, kNN and RF algorithms. We use scikit-learn's default configurations and the 3-fold cross-validation method to calculate the F1 scores. We found that DT perform better for web-based recordings and linear the SVM showed better performance for phone-based recordings.

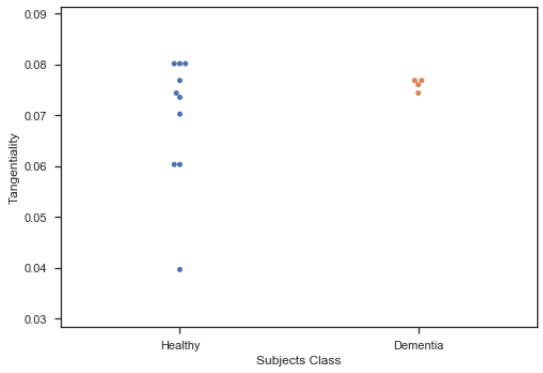
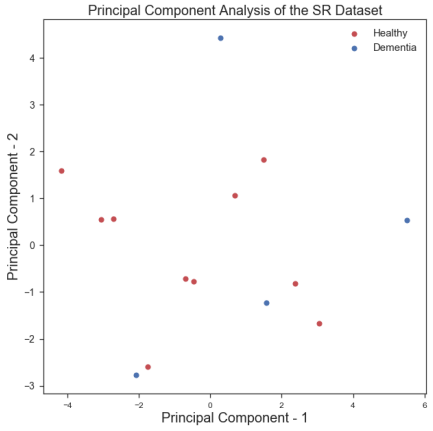
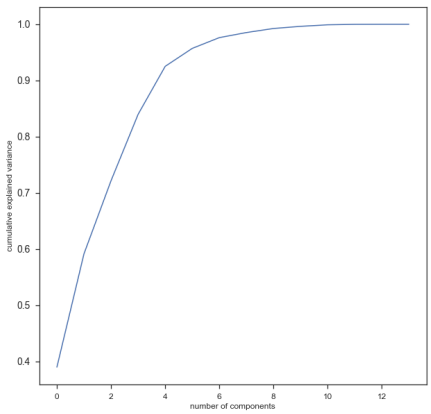


Figure 9: A comparison between the tangentiality measure for subjects with dementia and healthy subjects.



(a) It shows that subjects with dementia and healthy controls cannot be linearly separated using 2 principle components.

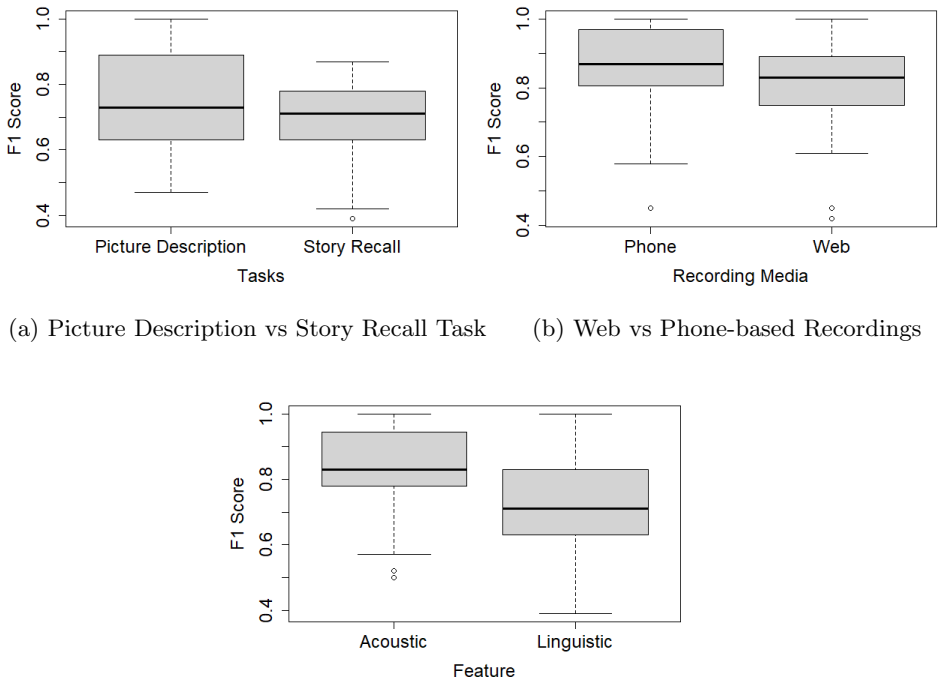


(b) It presents the values of cumulative explained variance for different number of principle components.

Figure 10

Table 9: Common acoustic features obtained by applying ANOVA, RF and mRMR feature selection methods on phone and web-based recordings

Web	Phone
MFCC 5,11,12 (mean)	MFCC 6, 9 (std)
Δ MFCC 11, 13 (mean)	MFCC 3 (skew)
Δ MFCC 0,3,6,9,10 (skew)	MFCC 3, 5 (kurt)
Δ log Mel freq 0,5,6 (skew)	Δ MFCC 0 (std)
Voicing prob. (kurt, std)	LSP freq 7 (mean)
Δ Voicing prob. (kurt, mean, std)	LSP freq 2, 3, 4 (skew)
LSP freq 0 (kurt)	LSP freq 1 (kurt)
F0 (skew)	Δ LSP freq 3 (mean)
Jitter local (kurt, skew)	Δ LSP freq 5 (skew)
Δ Jitter local (kurt)	log Mel freq 2 (skew)
Δ Jitter DDP (kurt)	Δ log Mel freq 1, 2, 3 (std)
Δ Shimmer local (kurt)	Voicing prob. (kurt, std)
	Loudness (kurt)



(a) Picture Description vs Story Recall Task (b) Web vs Phone-based Recordings

(c) Acoustic vs Linguistic Features

Figure 11: Boxplots showing the F1 scores obtained from different classifiers: a) distributions of the F1 scores in the picture description and story recall tasks, b) distributions of the F1 scores for web and phone-based recordings, and c) distributions of the F1 scores in the linguistic and acoustic features. These boxplots show that the picture description task, phone-based recordings and acoustic features provide better performance than the story recall task, web-based recorings and linguistic features.

4.2.1 Comparison between the phone-based and the web-based interfaces

We perform a one-way ANOVA test on the F1 scores of the phone and web-based recordings as shown in Tables 7 and 8. Our analysis shows that the means of these 2 groups are significantly different ($F(1,126) = 4.26$, $p = 0.04$). Figure 11(b) shows the distributions of F1 scores of these 2 groups. A Tukey's post-hoc test shows that the mean F1 scores of the classifiers developed by the extracted features from the phone-based recordings are higher than web-based recordings ($p = 0.04$), i.e., the ML classifiers trained with the phone-based recordings perform better than the web-based recordings.

4.3 Linguistic and Acoustic Features

Tables 7 and 8 show the results obtained by using different linguistic features and acoustic features. We consider all F1 scores (total 256) to compare the performance between the classifiers trained with linguistic and acoustic features. Figure 11(c) shows the distributions of F1 scores of these 2 groups. A one-way ANOVA test

performed on the F1 scores shows that the means are significantly different ($F(1, 256) = 62.43$, $p \approx 0$). A Tukey's test for post-hoc analysis shows that the mean F1 scores of the classifiers trained with the acoustic features are higher than the classifiers trained with the linguistic features ($p = 0$). That is, the ML classifiers trained with the acoustic features perform better than the classifiers trained with the linguistic features.

5 Discussion

We aim to use our proposed methodology to recognize different dimensions of language impairments in subjects with dementia. This section discusses it from different perspectives, such as the data limitation, validity, reliability, fairness, and explainability.

5.1 Data Limitation

How many samples are enough to develop an ML-based LA tool? No doubt having a lot of data samples, ML algorithms can learn better [92] to map linguistic features to the group of subjects (i.e., with dementia or without dementia). In other words, determining the optimal sample size for developing an efficient ML-based LA tool assures an adequate power to detect statistical significance [93]. However, for our problem, collecting language data from too many subjects is expensive and needs a lot of time. Thus, even it is necessary to estimate what is the sufficient size of samples for achieving acceptable classification results and then start to develop an ML-based assessment tool, but our results have shown that we could achieve good performance even with using the language data of less than 10 subjects (see Figure 12).

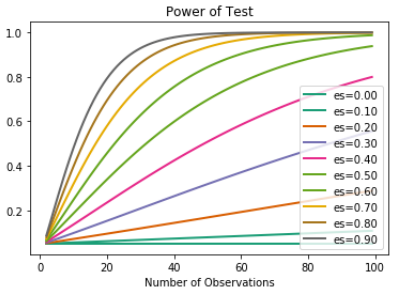


Figure 12: A description of using a power analysis to estimate the minimum sample size is required for achieving a desired effect size; It shows the impact of different effect sizes (es) and various sizes of the data sample on the statistical power

5.2 End-to-end Learning Approach

To develop an ML-based language assessment tool, we can use an end-to-end learning approach. Table 10 presents the obtained results employing *Convolutional Neural Networks* (ConvNet/CNN) on classifying the textual datasets. It shows that we can obtain approximately same results using CNN to classify our dataset. It might be a good idea to employ deep learning algorithm if we ave sufficient samples.

Table 10: Results from CNN

Experiment	Model Structure	Accuracy
PD task	2-channel CNN, Filter Size= 4, Kernel Size=2	66.7%
SR task	2-channel CNN, Filter Size= 4, Kernel Size=2	71.4%
Recording Media (Web)	2-channel CNN, Filter Size= 4, Kernel Size=2	80.0%
Recording Media (PHONE)	2-channel CNN, Filter Size= 4, Kernel Size=2	40.0%

5.3 Generalization - Selecting Meaningful Features

One of the problems we have faced with the acoustic features is that when we have applied ANOVA, RF, and mRMR feature selection methods on different datasets (i.e., obtained from various recording media or language tasks), each time we have received different sets of features (see Tables 6 and 9). Therefore, we are interested in combining all features so that we get almost consistent performance with all datasets. For this purpose, we have used PCA to combine a group of features, as shown in Table 11. We have considered MFCCs (0 to 14th order coefficients), the deltas of these MFCCs, and the deltas of LSP frequency bands (0 to 7) in the PCA because these groups of features appear more frequently in our rankings (see Tables 6 and 9). We found that the first two principal components (PCs) can retain, on average 75% of the variance, and hence we have considered only the first 2 PCs to train the classifiers. Table 12 shows how these PCA features perform with 4 different sets of data. Our results show that we achieved almost consistent performance with the tree-based classifies ranging from 78% to 93% F1 scores with these generalized set of features.

Table 11: Generalization - Combine the acoustic features using PCA

Feature Name	Functional	Principle Component (PC)
MFCC 0 - 14	mean	1st PC from the means of 15 MFCCs
		2nd PC from the means of 15 MFCCs
	kurt	1st PC from the kurt of 15 MFCCs
		2nd PC from the kurt of 15 MFCCs
	skew	1st PC from the skew of 15 MFCCs
		2nd PC from the skew of 15 MFCCs
Δ MFCC 0 - 14	mean	1st PC from the means of 15 Δ MFCCs
		2nd PC from the means of 15 Δ MFCCs
	kurt	1st PC from the kurt of 15 Δ MFCCs
		2nd PC from the kurt of 15 Δ MFCCs
	skew	1st PC from the skew of 15 Δ MFCCs
		2nd PC from the skew of 15 Δ MFCCs
Δ LSP freq 0 - 7	mean	1st PC from the means of 8 Δ LSP freq
		2nd PC from the means of 8 Δ LSP freq
	kurt	1st PC from the kurt of 8 Δ LSP freq
		2nd PC from the kurt of 8 Δ LSP freq
	skew	1st PC from the skew of 8 Δ LSP freq
		2nd PC from the skew of 8 Δ LSP freq

5.4 Is an ML-based LA tool Valid and Reliable?

We refer to reliability as the measure to trust the classification results [94]. We can assess the reliability and validity of an ML-based LA tool measuring the *intra-class correlation coefficient* (ICC) [95] and the *Pearson correlation coefficient* (PCC) [95] (i.e., if PCC returns a value close to 1, then the ML tool provides valid results; however, if the value is below, 0.5 indicates less correlation and validation).

Table 12: Results obtained by applying ML algorithms on PCA-based acoustic features that are extracted from all datasets

Classifier	PD	SR	Web	Phone
DT	**0.78 (+/- 0.16)**	**0.78 (+/- 0.02)**	0.72 (+/- 0.21)	0.80 (+/- 0.16)
ET	0.61 (+/- 0.28)	0.61 (+/- 0.28)	0.68 (+/- 0.22)	0.92 (+/- 0.10)
kNN	0.61 (+/- 0.08)	**0.78 (+/- 0.02)**	0.61 (+/- 0.08)	0.80 (+/- 0.28)
LDA	0.50 (+/- 0.14)	0.50 (+/- 0.14)	0.50 (+/- 0.14)	0.87 (+/- 0.09)
R_SVM	0.61 (+/- 0.08)	0.65 (+/- 0.18)	0.72 (+/- 0.21)	0.76 (+/- 0.17)
L_SVM	0.50 (+/- 0.14)	0.35 (+/- 0.33)	**0.89 (+/- 0.16)**	0.67 (+/- 0.25)
LR	0.72 (+/- 0.21)	0.22 (+/- 0.16)	**0.89 (+/- 0.16)**	0.60 (+/- 0.28)
RF	0.54 (+/- 0.15)	**0.78 (+/- 0.02**	0.79 (+/- 0.23)	**0.93 (+/- 0.07)**

5.5 Is an ML-based LA tool Fair?

ML-based LA tools are supervised classifiers, and are therefore prone to producing unfair results. In our work, we tried not to consider sensitive attributes such as gender, race as features [96]. However, we are working on different verbal tasks that might slightly be influenced by gender differences [97]. Another issue is that this type of assessment tool compares the user's language against similar users who are assumed to have AD or MCI [98]. Another essential attribute that might affect the fairness of ML-based LA tools is the level of education. It has been shown that some LA tools cannot provide accurate diagnostic when there are subjects with low levels of education among the population of study [19]. ML-based LA tools require a set of mechanisms to ensure that end-users trust in their performances and know how the system provides output. It is essential to motivate people to adopt not only the methods but also to share their data. Fairness and explainability are essential concerns, especially as ML-based assessment tools are being deployed more broadly in detecting other types of mental health problems. Fairness, in the end, comes down to robustness aspect. When we create ML-based assessment tools, we want them to be fair, and this means robust when deployed in different geographic settings and populations.

5.6 Is an ML-based LA tool Explainable?

An ML-based LA tool should be accurate and explainable to be adopted by psychiatrists during their assessment procedures. Thus, it is essential to choose an ML algorithm to develop a reliable ML-based LA tool that can describe its purpose, rationale, and decision-making process that can be understood by both clinicians and patients; it can foster the confidence of mental health professionals in employing it to detect subjects with dementia quickly.

5.7 What are other Application of ML-based LA tools?

Automated analysis of spontaneous connected speech can be useful for assessing and monitoring the progress of AD in patients. For example, integrating the ML-based LA tools to a conversational robot that can record patients' speech provides us, clinicians, an automated approach to follow the progress of the diseases [99]. In more detail, we can ask elderly individuals to describe the cookie theft picture to engage them to attend in a conversation. By extracting linguistic and acoustic features from the speech produced by them, we can identify if they are suffering from the linguistic disorder associated with AD or MCI [99]. The ML-based LA tools are the core part of any smartphone application that aims to support elderly

individuals with limited access to clinical services to receive real-time, cost-effective health care services. It decreases the burden on the caregivers.

6 Conclusion

This paper suggested a methodology to develop an efficient ML-based language assessment tool to detect language impairment in the older adults. We showed that the assessment tool is fitted with traditional ML classifiers, which can be trained with various sets of linguistic and acoustic features. Our results showed that the classifiers that have been trained using the PD dataset perform better than the SR dataset. We also found that the dataset obtained using phone-based recordings could increase ML classifiers' performance compared to the web-based dataset. Finally, we revealed that the classifiers trained only with the selected features using feature selection methods had higher performance than classifiers trained with pure features.

In the future, we will be working in the following directions: 1) Developing a cascade classifier that will be trained using both linguistic and acoustic features; 2) Using other types of data, such as eye-tracking; 2) Using few-shot ML algorithms and transfer learning techniques; 3) Considering pragmatic features such as fillers, GoAhead utterances, repetitions, incomplete words, and also contextual features using BERT (Bidirectional Encoder Representations from Transformers); 4) Using text data augmentation techniques such as EDA: Easy Data Augmentation techniques to augment data samples. 5) Classifying data.

7 Declarations

7.1 Ethics approval and consent to participate

The consent form has been approved by the Research Ethics Board protocol 31127 of the University of Toronto and has been signed by each subject has signed a consent form that has been.

7.2 Data

The datasets and codes would be accessible upon sending requests.

7.3 Funding

The first author would like to thank The Michael J. Fox Foundation for Parkinson's Research for funding her postdoctoral research projects. The second and third authors would like to thank AGE-WELL NCE for funding this research.

Author's contributions

Dr. Parsa and Dr. Alam worked with the linguistic and acoustic features, respectively, and they contributed equally to the methodology and results in sections. Dr. Mihailidis reviewed the manuscript and provided valuable feedbacks on the manuscript.

Acknowledgements

The first and second authors would like to thank Dr. Frank Rudzicz for providing access to the Talk2Me Database and valuable academic suggestions. The first author would like to thank Marina Tawfik for extracting datasets from the Talk2Me DB and reviewing the paper.

Author details
[1]Department of Computer Science, University of Toronto, Toronto, Canada. [2]Department Occupational Science and Occupational Therapy, University of Toronto, Toronto, Canada. [3]Toronto Rehab Institute, University Health Network, , Toronto, Canada. [4]Institute of Biomedical Engineering, University of Toronto, , Toronto, Canada.

References

1. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia (2019)
2. Ripich, D.N., Horner, J.: The neurodegenerative dementias: Diagnoses and interventions. The ASHA Leader **9**(8), 4–15 (2004)
3. Nichols, E., Szoeke, C.E., Vollset, S.E., Abbasi, N., Abd-Allah, F., Abdela, J., Aichour, M.T.E., Akinyemi, R.O., Alahdab, F., Asgedom, S.W., *et al.*: Global, regional, and national burden of alztteimer's disease and other dementias, 1990–2016: a systematic analysis for the global burden of disease study 2016. The Lancet Neurology **18**(1), 88–106 (2019)
4. SantaCruz, K., Swagerty Jr, D.L.: Early diagnosis of dementia. American Family Physician **63**(4), 703 (2001)
5. Green, R., Clarke, V., Thompson, N., Woodard, J., Letz, R.: Early detection of alzheimer disease: methods, markers, and misgivings. Alzheimer disease and associated disorders **11**(5), 1 (1997)
6. Duan, Y., Lu, L., Chen, J., Wu, C., Liang, J., Zheng, Y., Wu, J., Rong, P., Tang, C.: Psychosocial interventions for alzheimer's disease cognitive symptoms: a bayesian network meta-analysis. BMC geriatrics **18**(1), 175 (2018)
7. Fischer, C.E.: Music intervention approaches for alzheimer's disease: a review of the literature. Frontiers in Neuroscience **13**, 132 (2019)
8. Logsdon, R.G., McCurry, S.M., Teri, L.: Evidence-based interventions to improve quality of life for individuals with dementia. Alzheimer's care today **8**(4), 309 (2007)
9. El-Gamal, F.E., Elmogy, M.M., Ghazal, M., Atwan, A., Casanova, M.F., Barnes, G.N., Keynton, R., El-Baz, A.S., Khalil, A.: A novel early diagnosis system for mild cognitive impairment based on local region analysis: A pilot study. Frontiers in human neuroscience **11**, 643 (2018)
10. Klimova, B., Maresova, P., Valis, M., Hort, J., Kuca, K.: Alzheimer's disease and language impairments: social intervention and medical treatment. Clinical interventions in aging **10**, 1401 (2015)
11. Chiu, P.-Y., Tang, H., Wei, C.-Y., Zhang, C., Hung, G.-U., Zhou, W.: Nmd-12: A new machine-learning derived screening instrument to detect mild cognitive impairment and dementia. PloS one **14**(3) (2019)
12. Godino-Llorente, J.I., Gómez-Vilda, P., Sáenz-Lechón, N., Blanco-Velasco, M., Cruz-Roldán, F., Ferrer-Ballester, M.A.: Support vector machines applied to the detection of voice disorders. In: International Conference on Nonlinear Analyses and Algorithms for Speech Processing, pp. 219–230 (2005). Springer
13. Guinn, C.I., Habash, A.: Language analysis of speakers with dementia of the alzheimer's type. In: 2012 AAAI Fall Symposium Series (2012)
14. Orimaye, S.O., Wong, J.S.-M., Golden, K.J.: Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 78–87. Association for Computational Linguistics, Baltimore, Maryland, USA (2014). doi:10.3115/v1/W14-3210. https://www.aclweb.org/anthology/W14-3210
15. Asgari, M., Kaye, J., Dodge, H.: Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's & Dementia: Translational Research & Clinical Interventions **3**(2), 219–228 (2017)
16. Karlekar, S., Niu, T., Bansal, M.: Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. arXiv preprint arXiv:1804.06440 (2018)
17. Tsoi, K.K.F., Chan, J.Y.C., Hirai, H.W., Wong, S.Y.S., Kwok, T.C.Y.: Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. JAMA Internal Medicine **175**(9), 1450–1458 (2015). doi:10.1001/jamainternmed.2015.2152. https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2301149/ioi150045.pdf
18. Creavin, S.T., Wisniewski, S., Noel-Storr, A.H., Trevelyan, C.M., Hampton, T., Rayment, D., Thom, V.M., Nash, K.J., Elhamoui, H., Milligan, R., et al.: Mini-mental state examination (mmse) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. Cochrane Database of Systematic Reviews (1) (2016)
19. Tavares-Júnior, J.W.L., de Souza, A.C.C., Alves, G.S., Bonfadini, J.d.C., Siqueira-Neto, J.I., Braga-Neto, P.: Cognitive assessment tools for screening older adults with low levels of education: A critical review. Frontiers in Psychiatry **10**, 878 (2019). doi:10.3389/fpsyt.2019.00878
20. Kalish, V.B., Lerner, B.: Mini-mental state examination for the detection of dementia in older patients. American family physician **94**(11), 880–881 (2016)
21. Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H.: The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society **53**(4), 695–699 (2005)
22. Mathuranath, P., Nestor, P., Berrios, G., Rakowicz, W., Hodges, J.: A brief cognitive test battery to differentiate alzheimer's disease and frontotemporal dementia. Neurology **55**(11), 1613–1620 (2000)
23. Bruno, D., Vignaga, S.S.: Addenbrooke's cognitive examination iii in the diagnosis of dementia: a critical review. Neuropsychiatric disease and treatment **15**, 441 (2019)
24. Rosen, W.G., Mohs, R.C., Davis, K.L.: A new rating scale for alzheimer's disease. The American journal of psychiatry (1984)
25. Kueper, J.K., Speechley, M., Montero-Odasso, M.: The alzheimer's disease assessment scale–cognitive subscale (adas-cog): modifications and responsiveness in pre-dementia populations. a narrative review. Journal of Alzheimer's Disease **63**(2), 423–444 (2018)
26. Sheehan, B.: Assessment scales in dementia. Therapeutic advances in neurological disorders **5**(6), 349–358 (2012)
27. Tariq, S.H., Tumosa, N., Chibnall, J.T., Perry III, M.H., Morley, J.E.: Comparison of the saint louis university

mental status examination and the mini-mental state examination for detecting dementia and mild neurocognitive disorder—a pilot study. The American journal of geriatric psychiatry **14**(11), 900–910 (2006)
28. Ziso, B., Larner, A.J.: Codex (cognitive disorders examination) decision tree modified for the detection of dementia and mci. Diagnostics **9**(2), 58 (2019)
29. Robillard, J.M., Illes, J., Arcand, M., Beattie, B.L., Hayden, S., Lawrence, P., McGrenere, J., Reiner, P.B., Wittenberg, D., Jacova, C.: Scientific and ethical features of english-language online tests for alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **1**(3), 281–288 (2015)
30. O'Keeffe, S.T., Mulkerrin, E.C., Nayeem, K., Varughese, M., Pillay, I.: Use of serial mini-mental state examinations to diagnose and monitor delirium in elderly hospital patients. Journal of the American Geriatrics Society **53**(5), 867–870 (2005)
31. Vertesi, A., Lever, J.A., Molloy, D.W., Sanderson, B., Tuttle, I., Pokoradi, L., Principi, E.: Standardized mini-mental state examination. use and interpretation. Canadian Family Physician **47**(10), 2018–2023 (2001)
32. Chambers, L.W., Sivananthan, S., Brayne, C.: Is dementia screening of apparently healthy individuals justified? Advances in preventive medicine **2017** (2017)
33. Chaves, M.L., Godinho, C.C., Porto, C.S., Mansur, L., Carthery-Goulart, M.T., Yassuda, M.S., Beato, R.: Cognitive, functional and behavioral assessment: Alzheimer's disease. Dementia & neuropsychologia (2011)
34. Cummings, J.L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D.A., Gornbein, J.: The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia. Neurology **44**(12), 2308–2308 (1994)
35. Lai, C.K.: The merits and problems of neuropsychiatric inventory as an assessment tool in people with dementia and other neurological disorders. Clinical interventions in aging **9**, 1051 (2014)
36. Rubio, M.M., Antonietti, J., Donati, A., Rossier, J., Von Gunten, A.: Personality traits and behavioural and psychological symptoms in patients with mild cognitive impairment. Dementia and geriatric cognitive disorders **35**(1-2), 87–97 (2013)
37. Belanger, H.G., Wilder-Willis, K., Malloy, P., Salloway, S., Hamman, R.F., Grigsby, J.: Assessing motor and cognitive regulation in ad, mci, and controls using the behavioral dyscontrol scale. Archives of clinical neuropsychology **20**(2), 183–189 (2005)
38. Kraybill, M.L., Suchy, Y.: Executive functioning, motor programming, and functional independence: Accounting for variance, people, and time. The Clinical Neuropsychologist **25**(2), 210–223 (2011)
39. Niermeyer, M.A., Franchow, E.I., Suchy, Y.: Reported expressive suppression in daily life is associated with slower action planning. Journal of the International Neuropsychological Society **22**(6), 671–681 (2016)
40. Suchy, Y., Derbidge, C., Cope, C.: Behavioral dyscontrol scale-electronic version: First examination of reliability, validity, and incremental utility. The Clinical Neuropsychologist **19**(1), 4–26 (2005)
41. Mori, T., Kikuchi, T., Umeda-Kameyama, Y., Wada-Isoe, K., Kojima, S., Kagimura, T., Kudoh, C., Uchikado, H., Ueki, A., Yamashita, M., *et al.*: Abc dementia scale: a quick assessment tool for determining alzheimer's disease severity. Dementia and geriatric cognitive disorders extra **8**(1), 85–97 (2018)
42. Wang, C.-T., Hung, G.-U., Wei, C.-Y., Tzeng, R.-C., Chiu, P.-Y.: An informant-based simple questionnaire for visuospatial dysfunction assessment in dementia. Frontiers in Neuroscience **14**, 44 (2020)
43. Krein, L., Jeon, Y.-H., Miller Amberber, A.: Development of a new tool for the early identification of communication-support needs in people living with dementia: An australian face-validation study. Health & Social Care in the Community **28**(2), 544–554 (2020). doi:10.1111/hsc.12887. https://onlinelibrary.wiley.com/doi/pdf/10.1111/hsc.12887
44. Green, S., Reivonen, S., Rutter, L.-M., Nouzova, E., Duncan, N., Clarke, C., MacLullich, A.M., Tieges, Z.: Investigating speech and language impairments in delirium: a preliminary case-control study. PloS one **13**(11) (2018)
45. 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. Frontiers in aging neuroscience **7**, 195 (2015)
46. Ferreira, L.K., Busatto, G.F.: Neuroimaging in alzheimer's disease: current role in clinical practice and potential future applications. Clinics **66**, 19–24 (2011)
47. McCullough, K.C., Bayles, K.A., Bouldin, E.D.: Language performance of individuals at risk for mild cognitive impairment. Journal of Speech, Language, and Hearing Research **62**(3), 706–722 (2019)
48. Vestal, L., Smith-Olinde, L., Hicks, G., Hutton, T., Hart Jr, J.: Efficacy of language assessment in alzheimer's disease: comparing in-person examination and telemedicine. Clinical interventions in aging **1**(4), 467 (2006)
49. Yancheva, M., Rudzicz, F.: Vector-space topic models for detecting Alzheimer's disease. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2337–2346. Association for Computational Linguistics, Berlin, Germany (2016). doi:10.18653/v1/P16-1221. https://www.aclweb.org/anthology/P16-1221
50. Orimaye, S.O., Wong, J.S., Golden, K.J., Wong, C.P., Soyiri, I.N.: Predicting probable alzheimer's disease using linguistic deficits and biomarkers. BMC bioinformatics **18**(1), 34 (2017)
51. Noorian, Z., Pou-Prom, C., Rudzicz, F.: On the importance of normative data in speech-based assessment. arXiv preprint arXiv:1712.00069 (2017)
52. König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., *et al.*: Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **1**(1), 112–124 (2015)
53. Pope, C., Davis, B.H.: Finding a balance: The carolinas conversation collection. Corpus Linguistics and Linguistic Theory **7**(1), 143–161 (2011)
54. Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J., McGonigle, K.L.: The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology **51**(6), 585–594 (1994)
55. Calero, M.D., Arnedo, M.L., Navarro, E., Ruiz-Pedrosa, M., Carnero, C.: Usefulness of a 15-item version of the boston naming test in neuropsychological assessment of low-educational elders with dementia. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences **57**(2), 187–191 (2002)

56. Slegers, A., Filiou, R.-P., Montembeault, M., Brambati, S.M.: Connected speech features from picture description in alzheimer's disease: A systematic review. Journal of Alzheimer's Disease (Preprint), 1–26 (2018)
57. Cummings, L.: Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. Pragmatics and Society **10**(2), 153–176 (2019)
58. Coutinho, G., Drummond, C., de Oliveira-Souza, R., Moll, J., Tovar-Moll, F., Mattos, P.: Immediate story recall in elderly individuals with memory complaints: how much does it contribute to memory assessment? International psychogeriatrics **27**(10), 1679–1686 (2015)
59. Kurlowicz, L., Wallace, M., *et al.*: The mini-mental state examination (mmse). Journal of gerontological nursing **25**(5), 8–9 (1999)
60. Pekkala, S., Wiener, D., Himali, J.J., Beiser, A.S., Obler, L.K., Liu, Y., McKee, A., Auerbach, S., Seshadri, S., Wolf, P.A., Au, R.: Lexical retrieval in discourse: An early indicator of alzheimer's dementia. Clinical Linguistics & Phonetics **27**(12), 905–921 (2013). doi:10.3109/02699206.2013.815278. PMID: 23985011
61. Komeili, M., Pou-Prom, C., Liaqat, D., Fraser, K.C., Yancheva, M., Rudzicz, F.: Talk2me: Automated linguistic data collection for personal assessment. PloS one **14**(3) (2019)
62. Weakley, A., Schmitter-Edgecombe, M.: Analysis of verbal fluency ability in alzheimer's disease: the role of clustering, switching and semantic proximities. Archives of Clinical Neuropsychology **29**(3), 256–268 (2014)
63. Melrose, R.J., Campa, O.M., Harwood, D.G., Osato, S., Mandelkern, M.A., Sultzer, D.L.: The neural correlates of naming and fluency deficits in alzheimer's disease: an fdg-pet study. International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences **24**(8), 885–893 (2009)
64. Araújo, A.M.G.D.d., Lima, D.O., Nascimento, I.d.P., Almeida, A.A.F.d., Rosa, M.R.D.d.: Linguagem em idosos com doença de alzheimer: uma revisão sistemática. Revista CEFAC **17**(5), 1657–1663 (2015)
65. Habash, A.: Language Analysis of Speakers with Dementia of the Alzheimer's Type
66. Yancheva, M., Fraser, K.C., Rudzicz, F.: Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias. In: Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies, pp. 134–139 (2015)
67. Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's Disease **49**(2), 407–422 (2016)
68. Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., Roche-Bergua, A.: Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **10**, 260–268 (2018)
69. Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, J.: Spoken language derived measures for detecting mild cognitive impairment. IEEE transactions on audio, speech, and language processing **19**(7), 2081–2090 (2011)
70. Ahmed, S., Haigh, A.-M.F., de Jager, C.A., Garrard, P.: Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. Brain **136**(12), 3727–3737 (2013)
71. Meilán, J.J., Martínez-Sánchez, F., Carro, J., Sánchez, J.A., Pérez, E.: Acoustic markers associated with impairment in language processing in alzheimer's disease. The Spanish journal of psychology **15**(2), 487–494 (2012)
72. Weissenbacher, D., Johnson, T.A., Wojtulewicz, L., Dueck, A., Locke, D., Caselli, R., Gonzalez, G.: Automatic prediction of linguistic decline in writings of subjects with degenerative dementia. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1198–1207 (2016)
73. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics (2002)
74. Malvern, D., Richards, B., Chipere, N., Durán, P.: Lexical Diversity and Language Development. Springer
75. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)
76. Zhou, S., Jeong, H., Green, P.A.: How consistent are the best-known readability equations in estimating the readability of design standards? IEEE Transactions on Professional Communication **60**(1), 97–111 (2017)
77. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading **12**(8), 639–646 (1969)
78. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin, 37–54 (1948)
79. Sakai, E.Y., Carpenter, B.D.: Linguistic features of power dynamics in triadic dementia diagnostic conversations. Patient education and counseling **85**(2), 295–298 (2011)
80. Peelle, J.E., Cooke, A., Moore, P., Vesely, L., Grossman, M.: Syntactic and thematic components of sentence processing in progressive nonfluent aphasia and nonaphasic frontotemporal dementia. Journal of Neurolinguistics **20**(6), 482–494 (2007)
81. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings (2016)
82. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
83. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes **25**(2-3), 259–284 (1998)
84. McLoughlin, I.V.: Line spectral pairs. Signal Processing **88**(3), 448–467 (2008)
85. McLoughlin, I.V., Thambipillai, S.: Lsp parameter interpretation for speech classification. In: ICECS'99. Proceedings of ICECS'99. 6th IEEE International Conference on Electronics, Circuits and Systems (Cat. No. 99EX357), vol. 1, pp. 419–422 (1999). IEEE
86. De Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America **111**(4), 1917–1930 (2002)
87. Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O.: Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity. Journal of the royal society interface **8**(59), 842–855 (2011)

88. Yanushevskaya, I., Gobl, C., Ní Chasaide, A.: Voice quality in affect cueing: does loudness matter? Frontiers in psychology **4**, 335 (2013)
89. Meilán, J.J.G., Martínez-Sánchez, F., Carro, J., López, D.E., Millian-Morell, L., Arana, J.M.: Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? Dementia and Geriatric Cognitive Disorders **37**(5-6), 327–334 (2014)
90. Lopez-de-Ipina, K., Alonso, J.B., Travieso, C.M., Egiraun, H., Ecay, M., Ezeiza, A., Barroso, N., Martinez-Lage, P.: Automatic analysis of emotional response based on non-linear speech modeling oriented to alzheimer disease diagnosis. In: 2013 IEEE 17th International Conference on Intelligent Engineering Systems (INES), pp. 61–64 (2013). IEEE
91. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
92. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM **55**(10), 78–87 (2012)
93. Suresh, K., Chandrashekara, S.: Sample size estimation and power analysis for clinical research studies. Journal of human reproductive sciences **5**(1), 7 (2012)
94. Kukar, M., Kononenko, I.: Reliable classifications with machine learning. In: European Conference on Machine Learning, pp. 219–231 (2002). Springer
95. Molodynski, A., Linden, M., Juckel, G., Yeeles, K., Anderson, C., Vazquez-Montes, M., Burns, T.: The reliability, validity, and applicability of an english language version of the mini-icf-app. Social psychiatry and psychiatric epidemiology **48**(8), 1347–1354 (2013)
96. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning
97. Scheuringer, A., Wittig, R., Pletzer, B.: Sex differences in verbal fluency: the role of strategies and instructions. Cognitive processing **18**(4), 407–417 (2017)
98. Burr, C., Morley, J., Taddeo, M., Floridi, L.: Digital psychiatry: Risks and opportunities for public health and wellbeing. IEEE Transactions on Technology and Society **1**(1), 21–33 (2020)
99. Pou-Prom, C., Raimondo, S., Rudzicz, F.: A conversational robot for older adults with alzheimer's disease. ACM Transactions on Human-Robot Interaction (THRI) **9**(3), 1–25 (2020)

Additional Files