Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge

Saturnino Luz[1], *Fasih Haider*[1], *Sofia de la Fuente*[1], *Davida Fromm*[2], *Brian MacWhinney*[2]

[1]Usher Institute, Edinburgh Medical School, The University of Edinburgh, UK
[2]Department of Psychology, Carnegie Mellon University, USA

{S.Luz, fasih.haider, sofia.delafuente}@ed.ac.uk, {fromm, macw}@andrew.cmu.edu

Abstract

The ADReSS Challenge at INTERSPEECH 2020 defines a shared task through which different approaches to the automated recognition of Alzheimer's dementia based on spontaneous speech can be compared. ADReSS provides researchers with a benchmark speech dataset which has been acoustically pre-processed and balanced in terms of age and gender, defining two cognitive assessment tasks, namely: the Alzheimer's speech classification task and the neuropsychological score regression task. In the Alzheimer's speech classification task, ADReSS challenge participants create models for classifying speech as dementia or healthy control speech. In the the neuropsychological score regression task, participants create models to predict mini-mental state examination scores. This paper describes the ADReSS Challenge in detail and presents a baseline for both tasks, including feature extraction procedures and results for classification and regression models. ADReSS aims to provide the speech and language Alzheimer's research community with a platform for comprehensive methodological comparisons. This will hopefully contribute to addressing the lack of standardisation that currently affects the field and shed light on avenues for future research and clinical applicability.

Index Terms: Cognitive Decline Detection, Affective Computing, computational paralinguistics

1. Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that entails a long-term and usually gradual decrease of cognitive functioning [1]. It is also the most common underlying cause for dementia. The main risk factor for AD is age, and therefore its greatest incidence is amongst the elderly. Given the current demographics in the Western world, where the population aged 65 years or more has been predicted to triple between years 2000 and 2050 [2], institutions are investing considerably on dementia prevention, early detection and disease management. There is a need for cost-effective and scalable methods that are able to identify the most subtle forms of AD, from the preclinical stage of Subjective Cognitive Decline (SCI), to more severe conditions like Mild Cognitive Impairment (MCI) and Alzheimer's Dementia (AD) itself.

Whilst memory is often considered the main symptom of AD, language is also deemed as a valuable source of clinical information. Furthermore, the ubiquity of speech has led to a number of studies investigating speech and language features for the detection of AD, such as [3, 4, 5, 6] to cite some examples. Although these studies propose various signal processing and machine learning methods for this task, the field still lacks balanced and standardised datasets on which these different approaches could be systematically compared.

Consequently, the main objective of the ADReSS Challenge of INTERSPEECH 2020 is to define a shared task through which different approaches to AD detection, based on spontaneous speech, could be compared. This aims to address one of the main problems of this active research field, the lack of standardisation, which hinders its translation into clinical practice. The ADReSS Challenge will therefore: 1) target a difficult automatic prediction problem of societal and medical relevance, namely, the detection of cognitive impairment and Alzheimer's Dementia (AD); 2) to provide a forum for those different research groups to test their existing methods (or develop novel approaches) on a new shared standardized dataset; 3) mitigate common biases often overlooked in evaluations of AD detection methods, including repeated occurrences of speech from the same participant (common in longitudinal datasets), variations in audio quality, and imbalances of gender and age distribution; and 4) focus on AD recognition using spontaneous speech, rather than speech samples that are collected under laboratory conditions.

To the best of our knowledge, this will be the first such shared-task focused on AD. Unlike some tests performed in clinical settings, where short speech samples are collected under controlled conditions, this task focuses on AD recognition using spontaneous speech. While a number of researchers have proposed speech processing and natural language processing approaches to AD recognition through speech, their studies have used different, often unbalanced and acoustically varied datasets, consequently hindering reproducibility, replicability, and comparability of approaches. The ADReSS Challenge will provide a forum for those different research groups to test their existing methods (or develop novel approaches) on a shared dataset which consists of a statistically balanced, acoustically enhanced set of recordings of spontaneous speech sessions along with segmentation and detailed timestamped transcriptions. The use of spontaneous speech also sets the ADReSS Challenge apart from tests performed in clinical settings where short speech samples are collected under controlled conditions which are arguably less suitable for the development of large-scale monitoring technology than spontaneous speech [7].

As data scarcity and heterogeneity have hindered research into the relationship between speech and AD, the ADReSS Challenge provides researchers with the very first available benchmark, acoustically pre-processed and balanced in terms of age and gender. ADReSS defines two different prediction tasks: (a) the *AD recognition task*, which requires researchers to model participants' speech data to perform a binary classification of speech samples into AD and non-AD classes; and (b) the *MMSE prediction task*, which requires researchers to create regression models of the participants' speech in order to predict their scores in the Mini-Mental State Examination (MMSE).

This paper presents baselines for both tasks, including feature extraction procedures and initial results for a classification and a regression model.

2. ADReSS Challenge Dataset

A dataset has been created for this challenge which is matched for age and gender, as shown in Table 1 and Table 2, so as to minimise risk of bias in the prediction tasks. The data consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [8, 9]. Transcripts were annotated using the CHAT coding system [10]. The recorded speech has been segmented for voice activity using a simple voice activity detection algorithm based on signal energy threshold. We set the log energy threshold parameter to 65 dB with a maximum duration of 10 seconds per speech segment. The segmented dataset contains 1,955 speech segments from 78 non-AD subjects and 2,122 speech segments from 78 AD subjects. The average number of speech segments produced by each participant was 24.86 (standard deviation $sd = 12.84$). Recordings were acoustically enhanced with stationary noise removal and audio volume normalisation was applied across all speech segments to control for variation caused by recording conditions such as microphone placement.

Table 1: *ADReSS Training Set: Basic characteristics of the patients in each group (M=male and F=female).*

	AD			non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55)	1	0	30.0 (n/a)	1	0	29.0 (n/a)
[55, 60)	5	4	16.3 (4.9)	5	4	29.0 (1.3)
[60, 65)	3	6	18.3 (6.1)	3	6	29.3 (1.3)
[65, 70)	6	10	16.9 (5.8)	6	10	29.1 (0.9)
[70, 75)	6	8	15.8 (4.5)	6	8	29.1 (0.8)
[75, 80)	3	2	17.2 (5.4)	3	2	28.8 (0.4)
Total	24	30	17.0 (5.5)	24	30	29.1 (1.0)

Table 2: *Characteristics of the ADReSS test set.*

	AD			non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55)	1	0	23.0 (n.a)	1	0	28.0 (n.a)
[55, 60)	2	2	18.7 (1.0)	2	2	28.5 (1.2)
[60, 65)	1	3	14.7 (3.7)	1	3	28.7 (0.9)
[65, 70)	3	4	23.2 (4.0)	3	4	29.4 (0.7)
[70, 75)	3	3	17.3 (6.9)	3	3	28.0 (2.4)
[75, 80)	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	11	13	19.5 (5.3)	11	13	28.8 (1.5)

3. Acoustic and Linguistic Features

Acoustic feature extraction was performed on the speech segments using the openSMILE v2.1 toolkit which is an open-source software suite for automatic extraction of features from speech, widely used for emotion and affect recognition in speech [11], and with in-house software [**?**]. As the purpose of this paper is to describe the prediction tasks and set simple baselines that can be attained without extensive optimisation, we did not perform any feature set reduction procedures. The following is a brief description of the acoustic feature sets used in the experiments described in this paper:

emobase: This feature set contains the mel-frequency cepstral coefficients (MFCC) voice quality, fundamental frequency (F0), F0 envelope, line spectral pairs (LSP) and intensity features with their first and second order derivatives. Several statistical functions are applied to these features, resulting in a total of 988 features for every speech segment [11].

ComParE: The *ComParE 2013* [12] feature set includes energy, spectral, MFCC, and voicing related low-level descriptors (LLDs). LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, bringing the total to 6,373 features.

eGeMAPS: The *eGeMAPS* [13] feature set resulted from an attempt to reduce the somewhat unwieldy feature sets above to a basic set of acoustic features based on their potential to detect physiological changes in voice production, as well as theoretical significance and proven usefulness in previous studies [14]. It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, for a total of 88 features per speech segment.

MRCG functionals: Multi-resolution Cochleagram features (MRCGs) were proposed by Chen et al. [15] and have since been used in speech related applications such as voice activity detection [16], speech separation [15], and more recently for attitude recognition [17]. MRCG features are based on cochleagrams [18]. A cochleagram is generated by applying the gammatone filter to the audio signal, decomposing it in the frequency domain so as to mimic the human auditory filters. MRCG uses the time-frequency representation to encode the multi-resolution power distribution of the audio signal. Four cochleagram features were generated at different levels of resolution. The high resolution level encodes local information while the remaining three lower resolution levels capture spectrotemporal information. A total of 768 features were extracted from each frame: 256 MRCG features (frame length of 20 ms and frame shift of 10 ms), along with 256 Δ MRCG and 256 $\Delta\Delta$ MRCG features. The statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) were applied on the 768 MRCG features for a total of 6,912 features.

Minimal: this feature set consists of basic statistics (mean, standard deviation, median, minimum and maximum) of the duration of vocalisations and pauses and speech rate, and a vocalisation count, similarly to [7].

In sum, we extracted 88 eGeMAPS, 988 emobase, 6,373 ComParE, 6,912 MRCG, and 13 minimal features from 4,077 speech segments. Excepting the minimal feature set, Pearson's correlation test was performed to remove acoustic features that were significantly correlated with duration (when $|R| > 0.2$). Hence, 72 eGeMAPS, 599 emobase, 3,056 ComParE, and 3,253 MRCG features were not correlated with the duration of the speech chunks, and were therefore selected for the machine learning experiments. Examples of features from the ComParE feature set by the above described procedure include L1-norms of segment length functionals smoothed by a moving average filter (including their means, maxima and standard deviations), and the relative spectral transform applied to auditory spectrum (RASTA) functionals (including the percentage of time the signal is above 25%, 50% and 75% of range plus minimum).

In addition, we used the EVAL command in the CLAN program [**?**] to compute a basic set of 34 language outcome measures (e.g., duration, total utterances, MLU, type-token ratio, open-closed class word ratio, percentages of 9 parts of speech) on the CHAT transcripts.

4. AD classification task

The AD classification task consists of creating a binary classification models to distinguish between AD and non-AD patient speech. These models may use speech data, transcribed speech, or both. Any methodological approach may be taken, but participants will work with the same dataset. The evaluation metric for this task are $\text{Accuracy} = \frac{TN+TP}{N}$, precision $\pi = \frac{TP}{TP+FP}$, recall $\rho = \frac{TP}{TP+FN}$, and $\text{F}_1 = 2\frac{\pi \times \rho}{\pi + \rho}$, where N is the number of patients, TP, FP and FN are the number of true positives, false positives and false negatives, respectively.

We performed our baseline classification experiments using five different methods, namely linear discriminant analysis (LDA), decision trees (DT, with leaf size of 20 and the CART algorithm), nearest neighbour (1NN, for KNN with K=1), random forests (RF, with 50 trees and a leaf size of 20) and support vector machines (SVM, with a linear kernel with box constraint of 0.1, and sequential minimal optimisation solver). The classification methods were implemented in MATLAB [19] using the statistics and machine learning toolbox. A leave-one-subject-out (LOSO) cross-validation setting was adopted, where the training data do not contain any information from validation subjects.

Two-step classification experiments were conducted to detect cognitive impairment due to AD (as shown in Figure 1). This consisted of segment-level (SL) classification, where classifiers were trained and tested to predict whether a speech segment was uttered by a non-AD or AD patient, and majority vote (MV) classification, which assigned each subject a class label based on the majority labels of SL classification.

4.1. Results

The classification accuracy is shown in Tables 3 and 4 for LOSO and test settings respectively. These results show that the 1NN (0.574) provides the best accuracy for acoustic features using ComParE set for AD detection, with accuracy above the chance level of 0.50. From the results shown in Table 3, we note that even though 1NN provides the best result (0.574), DT and LDA also exhibit promising performance, being in fact more stable across all feature sets than the other classifiers (the best average accuracy of 0.559 for LDA and 0.570 for DT). We also note that Minimal, ComParE and linguistic also exhibit promising performance, being in fact more stable across all classifiers than the other features (the best average accuracy of 0.552 for Minimal, 0.541 for Compare and 0.713 for linguistic). Based on these findings we have selected the LDA model trained using ComParE as our baseline model for acoustic features.

Table 4 shows that 1NN provides less accurate results on the test set than in LOSO cross validation. However, the results of LDA (0.625) and DT (0.625) improve on the test data for acoustic features. The linguistic features provide an accuracy of 0.75, which is better than automatically extracted acoustic features though it relies on manual transcription. The challenge baseline accuracy for the classification task are therefore 0.625 for acoustic features and 0.75 for linguistic features. The precision, recall and F1 Score are reported in Table 5.

Table 3: *AD classification accuracy on LOSO cross validation.*

Features	LDA	DT	1NN	SVM	RF	mean
emobase	0.500	0.519	0.398	0.491	0.472	0.476
ComParE	**0.565**	0.528	0.574	0.528	0.509	**0.541**
eGeMAPS	0.482	0.500	0.380	0.333	0.482	0.435
MRCG	0.519	0.500	0.482	0.528	0.509	0.507
Minimal	0.519	0.667	0.426	0.565	0.583	0.552
linguistic	**0.768**	0.704	0.740	0.602	0.750	**0.713**
mean	**0.559**	**0.570**	0.500	0.508	0.551	–

Table 4: *AD classification accuracy on test set.*

Features	LDA	DT	1NN	SVM	RF	mean
emobase	0.542	0.688	0.604	0.500	0.729	0.613
ComParE	**0.625**	0.625	0.458	0.500	0.542	0.550
eGeMAPS	0.583	0.542	0.688	0.563	0.604	0.596
MRCG	0.542	0.563	0.417	0.521	0.542	0.517
Minimal	0.604	0.562	0.604	0.667	0.583	0.604
linguistic	**0.750**	0.625	0.667	0.792	0.750	0.717
mean	0.608	0.601	0.573	0.590	0.625	–

Table 5: *Baseline results of AD classification task using the LDA classifier with acoustic and linguistic features.*

	class	Precision	Recall	F1 Score	Accuracy
$LOSO_{Acous}$	non-AD	0.56	0.61	0.58	0.56
	AD	0.57	0.52	0.54	
$TEST_{Acous}$	non-AD	0.67	0.50	0.57	0.62
	AD	0.60	0.75	0.67	
$LOSO_{ling}$	non-AD	0.76	0.78	0.77	0.77
	AD	0.77	0.76	0.77	
$TEST_{ling}$	non-AD	0.70	0.87	0.78	0.75
	AD	0.83	0.62	0.71	

5. MMSE prediction task

The MMSE prediction task consists of generating a regression model for prediction of MMSE scores of individual participants from the AD and non-AD groups. Unlike classification, MMSE prediction is relatively uncommon in the literature, despite MMSE scores often being available. While models may use speech (acoustic) or linguistic data individually or in combination, the baseline described here report results of acoustic and linguistic models built separately.

5.1. Baseline regression

We performed our baseline regression experiments using five different methods, namely decision trees (DT, with leaf size of 20 and CART algorithm), linear regression (LR), gaussian process regression (GPR, with a squared exponential kernel), least-squares boosting (LSBoost, which contains the results of boosting 100 regression trees) and support vector machines (SVM, with a radial basis function kernel with box constraint of 0.1, and sequential minimal optimisation solver). The regression methods are implemented in MATLAB [19] using the statistics and machine learning toolbox. As with classification, the regression experiments were conducted in two steps for acoustic features (Figure 1), with SL regression followed by averaging of predicted MMSE values.

5.2. Results

The regression results are reported as root mean squared error (RMSE) scores in Tables 6 and 7 for LOSOCV and test data. These results show that DT (7.28) provides the best RMSE using MRCG features for MMSE prediction with $r = -0.759$, being more stable across all acoustic feature sets than the other classifiers (the best average RMSE of 6.86 for DT). We also note that Minimal and eGeMaPs also exhibit promising performance, with RMSE of 7.46 and 8.02 respectively across models. Based on this, the DT model trained using the MRCG feature was chosen as the baseline model for the regression task for acoustic features. For linguistic features, we selected the DT

To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimers Disease Detection

Aparna Balagopalan[1], *Benjamin Eyre*[1], *Frank Rudzicz*[2,3], *Jekaterina Novikova*[1]

[1]Winterlight Labs Inc, Toronto, Canada
[2]Department of Computer Science / Vector Institute for Artificial Intelligence, Toronto, Canada
[3]Li Ka Shing Knowledge Institute, St Michaels Hospital, Toronto, Canada

aparna@winterlightlabs.com, benjamin@winterlightlabs.com, frank@cs.toronto.edu, jekaterina@winterlightlabs.com

Abstract

Research related to automatically detecting Alzheimer's disease (AD) is important, given the high prevalence of AD and the high cost of traditional methods. Since AD significantly affects the content and acoustics of spontaneous speech, natural language processing and machine learning provide promising techniques for reliably detecting AD. We compare and contrast the performance of two such approaches for AD detection on the recent ADReSS challenge dataset [1]: 1) using domain knowledge-based hand-crafted features that capture linguistic and acoustic phenomena, and 2) fine-tuning Bidirectional Encoder Representations from Transformer (BERT)-based sequence classification models. We also compare multiple feature-based regression models for a neuropsychological score task in the challenge. We observe that fine-tuned BERT models, given the relative importance of linguistics in cognitive impairment detection, outperform feature-based approaches on the AD detection task.

Index Terms: Alzheimers disease, ADReSS, dementia detection, MMSE regression, BERT, feature engineering, transfer learning.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease that causes problems with memory, thinking, and behaviour. AD affects over 40 million people worldwide with high costs of acute and long-term care [2]. Current forms of diagnosis are both time consuming and expensive [3], which might explain why almost half of those living with AD do not receive a timely diagnosis [4].

Studies have shown that valuable clinical information indicative of cognition can be obtained from spontaneous speech elicited using pictures [5]. Several studies have used speech analysis, natural language processing (NLP), and ML to distinguish between healthy and cognitively impaired speech of participants in picture description datasets [6, 7]. These serve as quick, objective, and non-invasive assessments of an individual's cognitive status. However, although ML methods for automatic AD-detection using such speech datasets achieve high classification performance (between 82%-93% accuracy) [6, 8, 9], the field still lacks publicly-available, balanced, and standardised benchmark datasets. The ongoing ADReSS challenge [1] provides an age/sex-matched balanced speech dataset, which consists of speech from AD and non-AD participants describing a picture. The challenge consists of two key tasks: 1) Speech classification task: classifying speech as AD or non-AD. 2) Neuropsychological score regression task: predicting Mini-Mental State Examination (MMSE) [10] scores from speech.

In this work, we develop ML models to detect AD from speech using picture description data of the demographically-matched ADReSS Challenge speech dataset [1], and compare the following training regimes and input representations to detect AD:

1. **Using domain knowledge**: with this approach, we extract linguistic features from transcripts of speech, and acoustic features from corresponding audio files for binary AD vs non-AD classification and MMSE score regression. The features extracted are informed by previous clinical and ML research in the space of cognitive impairment detection [6].

2. **Using transfer learning**: with this approach, we fine-tune pre-trained BERT [11] text classification models at transcript-level. BERT achieved state-of-the-art results on a wide variety of NLP tasks when fine-tuned [11]. Our motivation is to benchmark a similar training procedure on transcripts from a pathological speech dataset, and evaluate the effectiveness of high-level language representations from BERT in detecting AD.

In this paper, we evaluate performance of these two methods on both the ADReSS train dataset, and on the unseen test set. We find that fine-tuned BERT-based text sequence classification models achieve the highest AD detection accuracy with an accuracy of 83.3% on the test set. With the feature-based models, the highest accuracy of 81.3% is achieved by the SVM with RBF kernel model. The lowest root mean squared error obtained for the MMSE prediction task is 4.56, with a feature-based L2 regularized linear regression model.

The main contributions of our paper are as follows:

- We employ a domain knowledge-based approach and compare a number of AD detection and MMSE regression models with an extensive list of pre-defined linguistic and acoustic features as input representations from speech (Section 5 and 6).

- We employ a transfer learning-based approach and benchmark fine-tuned BERT models for the AD vs non-AD classification task (Section 5 and 6).

- We contrast the performance of the two approaches on the classification task, and discuss the reasons for existing differences (Section 7).

Table 2: *Feature differentiation analysis results for the most important features, based on ADReSS train set. μ_{AD} and μ_{non-AD} show the means of the 13 significantly different features at p<9e-5 (after Bonferroni correction) for the AD and non-AD group respectively. We also show Spearman correlation between MMSE score and features, and regression weights of the features associated with the five greatest and five lowest regression weights from our regression experiments. * next to correlation indicates significance at p<9e-5.*

Feature	**Feature type**	μ_{AD}	μ_{non-AD}	**Correlation**	**Weight**
Average cosine distance between utterances	Semantic	0.91	0.94	-	-
Fraction of pairs of utterances below a similarity threshold (0.5)	Semantic	0.03	0.01	-	-
Average cosine distance between 300-dimensional word2vec [28] utterances and picture content units	Semantic (content units)	0.46	0.38	-0.54*	-1.01
Distinct content units mentioned: total content units	Semantic (content units)	0.27	0.45	0.63*	1.78
Distinct action content units mentioned: total content units	Semantic (content units)	0.15	0.30	0.49*	1.04
Distinct object content units mentioned: total content units	Semantic (content units)	0.28	0.47	0.59*	1.72
Average cosine distance between 50-dimensional GloVe utterances and picture content units	Semantic content units)	-	-	-0.42*	-0.03
Average word length (in letters)	Lexico-syntactic	3.57	3.78	0.45*	1.07
Proportion of pronouns	Lexico-syntactic	0.09	0.06	-	-
Ratio (pronouns):(pronouns+nouns)	Lexico-syntactic	0.35	0.23	-	-
Proportion of personal pronouns	Lexico-syntactic	0.09	0.06	-	-
Proportion of RB adverbs	Lexico-syntactic	0.06	0.04	-0.41*	-0.41
Proportion of ADVP_− >_RB amongst all rules	Lexico-syntactic	0.02	0.01	-0.37	-0.74
Proportion of non-dictionary words	Lexico-syntactic	0.11	0.08	-	-
Proportion of gerund verbs	Lexico-syntactic	-	-	0.37	1.08
Proportion of words in adverb category	Lexico-syntactic	-	-	-0.4*	-0.49

Table 3: *10-fold CV results averaged across 3 runs with different random seeds on the ADReSS train set. Accuracy for BERT is higher, but not significantly so from SVM ($H = 0.4838, p > 0.05$ Kruskal-Wallis H test). Bold indicates the best result.*

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
SVM	10	0.796	0.81	0.78	0.82	0.79
NN	10	0.762	0.77	0.75	0.77	0.76
RF	50	0.738	0.73	0.76	0.72	0.74
NB	80	0.750	0.76	0.74	0.76	0.75
BERT	-	**0.818**	**0.84**	**0.79**	**0.85**	**0.81**

Table 4: *LOSO-CV results averaged across 3 runs with different random seeds on the ADReSS train set. Accuracy for SVM is significantly higher than NN ($H = 4.50, p = 0.034$ Kruskal-Wallis H test). Bold indicates the best result.*

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	0.574	0.57	0.52	-	0.54
SVM	509	0.741	0.75	0.72	0.76	0.74
SVM	10	**0.870**	**0.90**	**0.83**	**0.91**	**0.87**
NN	10	0.836	0.86	0.81	0.86	0.83
RF	50	0.778	0.79	0.77	0.79	0.78
NB	80	0.787	0.80	0.76	0.82	0.78

Table 5: *Results on unseen, held-out ADReSS test set .We present test results in same format as the baseline paper [1]. Bold indicates the best result.*

Model	#Features	Class	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	non-AD	0.625	0.67	0.50	-	0.57
		AD		0.60	0.75	-	0.67
SVM	10	non-AD	0.813	0.83	0.79	-	0.81
		AD		0.80	0.83	-	0.82
NN	10	non-AD	0.771	0.78	0.75	-	0.77
		AD		0.76	0.79	-	0.78
RF	50	non-AD	0.750	0.71	**0.83**	-	0.77
		AD		0.80	0.67	-	0.73
NB	80	non-AD	0.729	0.69	**0.83**	-	0.75
		AD		0.79	0.63	-	0.70
BERT	-	non-AD	**0.833**	**0.86**	0.79	-	**0.83**
		AD		0.81	**0.88**	-	**0.84**

Table 6: *LOSO-CV MMSE regression results on the ADReSS train and test sets. Bold indicates the best result.*

Model	#Features	α	RMSE	MAE	RMSE
			Train set		Test set
Baseline [1]	-	-	7.28		6.14
LR	15	-	5.37	4.18	4.94
LR	20	-	4.94	3.72	-
Ridge	509	12	6.06	4.36	-
Ridge	35	12	4.87	3.79	**4.56**
Ridge	25	10	**4.56**	**3.50**	-

most of the important lexico-syntactic and semantic features. It is thus able to use information present in the lexicon, syntax, and semantics of the transcribed speech after fine-tuning [31]. We also see a trend of better performance when increasing the number of folds (see SVM in Table 4 and Table 3) in cross-validation. We postulate that this is due to the small size of the dataset, and hence differences in training set size in each fold (N_{train} = 107 with LOSO, N_{train} = 98 with 10-fold CV).

7.3. Regression Weights

To assess the relative importance of individual input features for MMSE prediction, we report features with the five highest and five lowest regression weights in Table 2. Each presented value is the average weight assigned to that feature across each of the LOSO CV folds. We also present the correlation with MMSE score coefficients for those 10 features, as well as their significance, in Table 2. We observe that for each of these highly weighted features, a positive or negative correlation coefficient is accompanied by a positive or negative regression weight, respectively. This demonstrates that these 10 features are so distinguishing that, even in the presence of other regressors, their relationship with MMSE score remains the same. We also note that all 10 of these are linguistic features, further demonstrating that linguistic information is particularly distinguishing when it comes to predicting the severity of a patient's AD.

8. Conclusions

In this paper, we compare two widely used approaches – explicit features engineering based on domain knowledge, and transfer learning using fine-tuned BERT classification model. Our results show that pre-trained models that are fine-tuned for the AD classification task are capable of performing well on AD detection, and outperforming hand-crafted feature engineering. A direction for future work is developing ML models that combine representations from BERT and hand-crafted features [32]. Such feature-fusion approaches could potentially boost performance on the cognitive impairment detection task.

9. Acknowledgements

This paper benefited greatly from feedback and review from multiple people. Most notably, Dr. J. Robin (Winterlight), L. Kaufman (Winterlight) and M. Yancheva (Winterlight). F. Rudzicz is supported by a CIFAR Chair in AI.

Table 7: *Summary of all lexico-syntactic features extracted. The number of features in each subtype is shown in the second column (titled "#features").*

Feature type	#Features	Brief Description
Syntactic Complexity	36	L2 Syntactic Complexity Analyzer [33] features; max/min utterance length, depth of syntactic parse tree
Production Rules	104	Number of times a production type occurs divided by total number of productions
Phrasal type ratios	13	Proportion, average length and rate of phrase types
Lexical norm-based	12	Average norms across all words, across nouns only and across verbs only for imageability, age of acquisition, familiarity and frequency (commonness)
Lexical richness	6	Type-token ratios (including moving window); brunet; Honors statistic
Word category	5	Proportion of demonstratives (e.g., "this"), function words, light verbs and inflected verbs, and propositions (POS tag verb, adjective, adverb, conjunction, or preposition)
Noun ratio	3	Ratios nouns:(nouns+verbs); nouns:verbs; pronouns:(nouns+pronouns)
Length measures	1	Average word length
Universal POS proportions	18	Proportions of Spacy univeral POS tags [34]
POS tag proportions	53	Proportions of Penn Treebank [35] POS tags
Local coherence	15	Avg/max/min similarity between word2vec [28] representations of utterances (with different dimensions)
Utterance distances	5	Fraction of pairs of utterances below a similarity threshold (0.5,0.3,0); avg/min distance
Speech-graph features	13	Representing words as nodes in a graph and computing density, number of loops etc.
Utterance cohesion	1	Number of switches in verb tense across utterances divided by total number of utterances
Rate	2	Ratios – number of words: duration of audio; number of syllables: duration of speech,
Invalid words	1	Proportion of words not in the English dictionary
Sentiment norm-based	9	Average sentiment valence, arousal and dominance across all words, noun and verbs

Table 8: *Summary of all acoustic features extracted. The number of features in each subtype is shown in the second column (titled "#features").*

Feature type	#Features	Brief Description
Pauses and fillers	9	Total and mean duration of pauses;long and short pause counts; pause to word ratio; fillers(um,uh); duration of pauses to word durations
Fundamental frequency	4	Avg/min/max/median fundamental frequency of audio
Duration-related	2	Duration of audio and spoken segment of audio
Zero-crossing rate	4	Avg/variance/skewness/kurtosis of zero-crossing rate
Mel-frequency Cepstral Coefficients (MFCC)	168	Avg/variance/skewness/kurtosis of 42 MFCC coefficients

Table 9: *Summary of all semantic features extracted. The number of features in each subtype is shown in the second column (titled "#features").*

Feature type	#Features	Brief Description
Word frequency	10	Proportion of lemmatized words, relating to the Cookie Theft picture content units to total number of content units
Global coherence	15	Avg/min/max cosine distance between word2vec [28] utterances and picture content units, with varying dimensions of word2vec