An Automatic System for Dementia Detection using Acoustic and Linguistic Features

Miriam Gonzalez-Atienza, Jose A. Gonzalez-Lopez, and Antonio M. Peinado

Dept. of Signal Theory, Telematics and Communications
University of Granada

myriamgonzalez@correo.ugr.es, joseangl@ugr.es, amp@ugr.es

Abstract

Early diagnosis of dementia is crucial for mitigating the consequences of this disease in patients. Previous studies have demonstrated that it is possible to detect the symptoms of dementia, in some cases even years before the onset of the disease, by detecting neurodegeneration-associated characteristics in a person's speech. This paper presents an automatic method for detecting dementia caused by Alzheimer's disease (AD) through a wide range of acoustic and linguistic features extracted from the person's speech. Two well-known databases containing speech for patients with AD and healthy controls are used to this end: DementiaBank and ADReSS. The experimental results show that our system is able to achieve state-of-the-art performance on both databases. Furthermore, our results also show that the linguistic features extracted from the speech transcription are significantly better for detecting dementia.

Index Terms: Acoustic voice analysis, speech-based disease diagnosis, dementia, Alzheimer's disease, word embeddings.

1. Introduction

Dementia is a type of neurodegenerative disease, whose most common cause is Alzheimer's Disease (AD). Although memory impairment is the main early symptom for AD, it has been found that language and speech abilities also decline, even in the very early stages of the disease [1]. This is known to affect object naming, noun production and rates of verb usage. In general, loss of vocabulary, simplified syntax/semantics, and overuse of semantically empty words are commonly found in the language of people with dementia [2, 3]. Speech is, therefore, a promising candidate as a source of information for new approaches to diagnosing dementia. As no curative therapy is known, early secondary prevention measures are of great importance. Examinations typically include a large number of neuropsychological tests, which are very time consuming and expensive. In order to enable longitudinal cognitive status monitoring on a large scale, a fast and cheap method for diagnosing the disease needs to be found.

Automatic speech processing has been shown to be a promising way in the diagnosis of dementia. Approaches have used acoustic, prosodic and linguistic features [4, 5, 6] in a classification task that aims to distinguish people affected by dementia from cognitively healthy subjects using just their speech. Patient's speech and language are obtained from written texts and speech recordings. For example, *picture description* is a constrained task that relies less on episodic memory, but requires more semantic knowledge and retrieval ability. The most commonly used picture prompt is a line drawing called "Cookie Theft" [7]. During the test, the patients are asked to describe what they see in the picture, while the answer is recorded. As reported [8], people suffering from dementia tend to hesitate more often and make longer pauses. Thus, features based on the occurrence and duration of pauses, extracted from the output of a Voice Activity Detector (VAD), are very promising for automatic detection of this disease.

In this paper, we present the details of our system for dementia detection using features extracted from speech recordings and its corresponding transcriptions. We use both acoustic and linguistic features and compare the results to those obtained in previous investigations. Feature selection of the most relevant features is performed in order to achieve better classification results. We evaluated our proposed system on two well-known databases containing speech material recorded by AD patients and healthy controls (HC): DementiaBank [9] and the recently released Alzheimer's dementia recognition through spontaneous speech (ADReSS) database [10].

This paper has the following structure. In Section 2, we present the details of our system for automatic dementia detection. Section 3 is dedicated to explain the process followed to select the most important features in classification. The experimental results are presented in Section 5, whereas the main conclusions of this work are listed in Section 7.

2. Dementia detection system

Figure 1 shows a block diagram for the proposed automatic system for speech-based dementia detection. Firstly, the voice of the participant is recorded and transcribed (manually, in our case) while the subject performs a cognitive task (i.e., describing a drawing such as the Cookie Theft from the Boston test). Then, a noise reduction technique is applied to the audio signal. The enhanced signal is passed to the feature extraction block where a set of acoustic and linguistic features are extracted. Finally, a selection of this features are sent to a machine learning classifier trained to discriminate between AD and HC subjects.

The details of our system are given in the following sections. Table 1 shows a summary of the acoustic and linguistic features extracted by our system for automatic dementia detection.

2.1. Acoustic features

Acoustic features measure how participants speak. They are extracted from the de-noised speech signals after applying the spectral noise gating method described in [11] to the audio signals to improve their quality. In our system, we consider the

This work was funded by the Spanish State Research Agency (SRA) under the grant PID2019-108040RB-C22/SRA/10.13039/501100011033. Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporation Fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-32926).

Table 1: *Description of the 319 speech features extracted by our system for automatic dementia detection.*

Feature set	Description	No.
Pause based	% of pauses in speech	1
	% of pauses in utterance	1
	Speech and pauses duration	2
Speech rhythm	Beats per minute	1
Spectral features	MFCCs mean	1
	MFCCs variance	1
	MFCCs skewness	1
	MFCCs kurtosis	1
	Mean of ΔMFCCs	1
	Variance of ΔMFCCs	1
	Mean of first 24 spectral centroids	1
Prosodic features	Avg. of F_0 values	1
	Std. of F_0 values	1
	Acoustic features:	**14**
Word count	% of unique words	1
POS tags	% of verbs	1
	% of determinants	1
	% of nouns	1
Word embeddings	Avg. of word embeddings	300
	Std. of word embeddings	1
	Linguistic features:	**305**

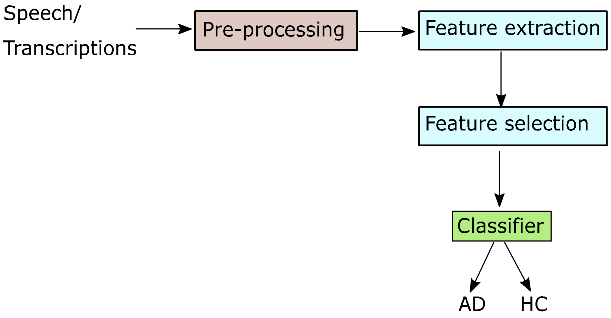

Figure 1: *Block diagram of the dementia detection system*

following four types of acoustic features:

Pause-based features: People afflicted with dementia tend to hesitate more often and make longer pauses than HC subjects [6, 12]. Thus, we computed four pause-based features for each utterance, including total pause and speech duration (in seconds), rate between total duration of pauses and duration of the utterance, and rate between the total duration of pauses and the duration of speech for each utterance. These features were automatically computed from the outputs of an energy-based VAD technique applied to the de-noised signals. The energy threshold was manually defined as 60% of the signal energy.

Speech rhythm: Longer pauses and more hesitations in speech from people with dementia imply that more time is needed to convey words. Thus, we measured the speech rhythm as number of beats per minute using the Predominant Local Pulse (PLP) algorithm, first introduced by Grosche and Müller [13].

Spectral features: Mel-frequency cepstral coefficients (MFCCs) [14, 11, 15, 16] were extracted from windows of 25 ms with 15 ms overlap to capture the spectral content of the speech signal. We then computed the average across time for the first 24 MFCCs. Then, we computed the mean, variance, skewness and kurtosis of the resulting average vector. We also extracted their first order derivatives with their mean and variance. Spectral centroid parameters[17], which aim to locate the spectrum centre of mass and have been found to be valuable in measuring the cognitive load [18, 11], were also extracted from the audio waveforms and computed their mean.

Prosodic features: We computed the speech fundamental frequency (F_0) using the autocorrelation method [19]. From the F_0 values, we computed the mean and standard deviation through the signal.

2.2. Linguistic features

Linguistic features are used to measure changes in vocabulary and sentence structure that are caused by dementia. Our linguistic features operate at the word level of transcriptions. In this work, the manual transcriptions provided with the databases are used to extract the linguistic features. In general, AD patients tend to make shorter phrases than HC and also have a less-rich vocabulary. The linguistic features described below are computed as proportions considering the total number of words spoken in each utterance, e.g., percentage of adjectives w.r.t. the total number of spoken words.

Word count: Word count is a common used feature to classify people with dementia [16]. Verbal repetition is a hallmark of dementia and AD at all stages, but is most commonly targeted for monitoring and treatment effects in its mild stage [20]. For that reason, we have extracted the proportion between unique words and the total number of words in each transcription.

Part-of-Speech (POS) Tags: Words with similar grammatical properties can be grouped together by POS tags. Each tag represents the grammatical role a word can take in a sentence and thus POS tags can be used to indicate grammatical properties of participant's speech. We used TreeTagger [21] to automatically extract POS tags and calculate the frequency of occurrence of each tag. We have considered three POS categories: verbs, determinants and nouns. Once the frequency of occurrence is calculate for each participant, we measure the proportion between the number of POS categories and the number of words in each sentence.

Word embeddings: Word embedding is a technique widely used to convert written words into feature vectors. Recently, successful approaches have used deep learning techniques to produce vectors representing words. In this work, we have used the *word2vec* technique [22], which is based on the co-ocurrences of words taking into account the context, to extract 300-dimensional embeddings for every word spoken by the participant in each sentence. Then, we average all the embeddings obtained in each sentence to finally obtain a 300-dimensional vector representing the linguistic content of that sentence. Besides, we calculate the mean and standard deviation for each word embedding vector of the subject. Thus, we obtain a 300-dimensional vector for each subject, with two additional values: the mean and standard deviation of this vector.

3. Feature selection

We applied a feature selection procedure to select the most meaningful acoustic and linguistic features and discard the features that contribute less to the classification accuracy. We trained an extremely randomized trees (ERT) classifier [23], which is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. Feature selection was applied to the training set and then, the optimum

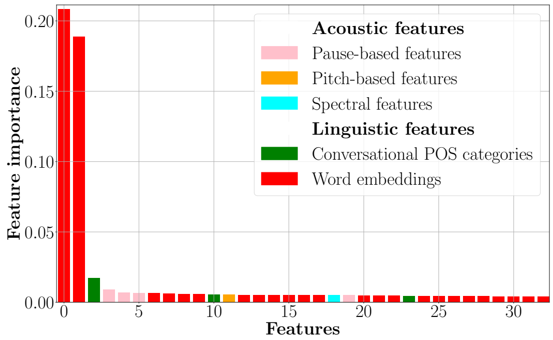

Figure 2: *Ranking of the top 15% features for DementiaBank database.*

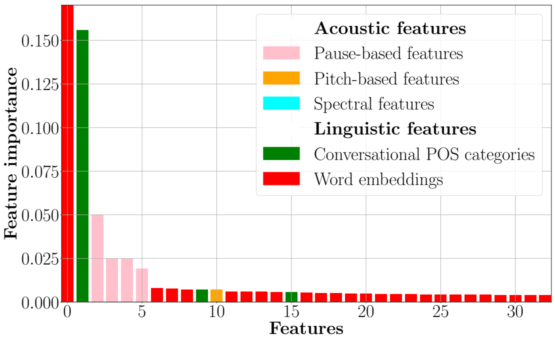

Figure 3: *Ranking of the top 15% features for the ADReSS database.*

set of features selected by this procedure, was used during evaluation time to detect dementia based on a person's speech.

Figures 2 and 3 show the ranking of the top 15% features selected by the the ERT technique for the DementiaBank and ADReSS databases, respectively, from the whole set of 319 acoustic and linguistic features extracted by our system. As can be seen, the majority of selected features are linguistic. This, as also reported by other authors [16, 20], makes sense because dementia is known to have a profound impact on a person's speech, even from its first stages [1]. In particular, POS features based on the proportions of nouns, verbs, and determinants are very relevant for detecting dementia on both databases, with relative importance of 0.01, 0,005 and 0.004 in DementiaBank, and 0.150, 0.015, and 0.014 in ADReSS. Word embedding based features are also among the best features for both databases, thus highlighting again the importance of linguistic-based features for the task of automatic detection of neurodegenerative diseases. Although acoustic features are deemed less relevant, the ERT technique also selected a significant number of them, particularly for the ADReSS database, where pause-based features related to pause and speech ratio are among the top 5-best features.

4. Experimental setup

4.1. Databases

To evaluate the proposed dementia detection system, we used the audio recordings and transcriptions from the DementiaBank [11, 24] and ADReSS [10] databases. DementiaBank is a free-access, large existing database for Alzheimer's and related dementia diseases collected during longitudinal study conducted by the University of Pittsburgh School of Medicine. Verbal descriptions of the Boston Cookie Theft picture were recorded from people with different types of dementia with an age span from 49 to 90 years as well as from elderly healthy control subjects within an age range from 46 to 81 years. During the interviews, patients were given the picture and were told to discuss everything they could see happening in the picture. There are a total of 473 recordings from 97 healthy controls and 233 speech samples from 167 AD patients diagnosed as possible or probable AD. We splitted this database randomly into training and evaluation subsets. In the training subset, we have selected 150 subjects, as well as for the evaluation subset, in order to have the same number of subject in both groups.

Similarly, the ADReSS database is a subset of DementiaBank with acoustically enhanced audio recordings and matched in terms of age and gender (i.e., major factors in recognising dementia) to avoid bias towards them. The dataset contains speech recordings from 78 non-AD subjects and 78 AD subjects while describing the Cookie Theft drawing. This database already defines a training subset, containing data for 54 AD and 54 HC subjects, and an evaluation subset with 24 subjects for each group.

A summary of the characteristics of both databases in terms of number of subjects in each class in shown in Table 2.

Table 2: *Basic characteristics of the patients in each group in the ADReSS challenge dataset and DementiaBank*

Dataset		**No. subjects**	
		AD	**Non-AD**
ADReSS	Train	54	54
	Test	24	24
DementiaBank	-	322	229

4.2. Evaluation

We evaluated four state-of-the-art classifiers for the task of classifying dementia from the set of extracted linguistic and acoustic features: Linear Discriminant Analysis (LDA), Support Vector Machines (SVMs), Random Forests and Adaptive Boosting, previously used for dementia detection in [12, 25, 26]. We also evaluated the effect of the feature selection procedure described in Section 3 on classification accuracy. In particular, the following configurations were evaluated: (i) using only either acoustic features (Acoustic) or (ii) linguistic features (Linguistic) with no feature selection; or (iii) a combination of both types of features where only a fraction of them are used for classification, as provided by the feature selection algorithm (Feature selection).

The following classification metrics are reported for each configuration: accuracy, recall (sensitivity) and specificity, also considered in [1, 11, 27]. Accuracy is the percentage of correct predictions. Sensitivity describes what proportion of patients with AD are correctly identified as having AD, while specificity describes what proportion of HC persons are correctly identified as belonging to that class.

5. Experimental Results

Table 3 shows the classification results achieved by our system on the DementiaBank dataset. With LDA classifier and using only linguistic features, we achieve an accuracy of 96%, 97% sensitivity and 92.3% specificity, providing the higher classifi-

Table 3: *Classification results for DementiaBank dataset. Accuracy/Sensitivity/Specificity (%). In bold are shown the higher classification results by rows*

Features	LDA	SVM	RF	AdaBoost	Avg.
Acoustic	65/65/78	67/65/60	**71/71/70**	65/65/73	67/67/70
Linguistic (with word embeddings)	**96/97/92**	62/62/92	74/60/93	78/78/75	78/74/88
Feature selection (15%)	**95/95/93**	67/65/57	78/56/48	81/77/74	80/73/68
Feature selection (25%)	**93/93/94**	75/74/81	81/67/64	78/78/74	82/78/78
Feature selection (50%)	**87/86/94**	83/83/88	78/66/68	68/67/75	79/76/81
Feature selection (75%)	**94/94/95**	72/71/71	81/67/64	85/78/77	83/78/77
Feature selection (100%)	82/82/71	71/65/86	**90/87/85**	79/66/62	81/75/76
Al-Hammed et al [11, 14]	-	86/-/-	96/-/-	86/-/-	-

Table 4: *Classification results for ADReSS dataset. Accuracy/Sensitivity/Specificity (%). In bold are shown the higher classification results by rows*

Features	LDA	SVM	RF	AdaBoost	Avg.
Acoustic	65/65/67	**65/65/78**	59/57/78	59/58/78	62/61/75
Linguistic (with word embeddings)	72/71/80	**78/76/67**	61/59/80	65/65/78	69/68/76
Feature selection (15%)	76/55/54	**79/65/60**	57/56/80	57/55/80	67/58/69
Feature selection (25%)	53/53/50	**66//60/56**	57/55/80	56/54/75	58/56/65
Feature selection (50%)	59/53/47	55/53/50	**69/68/80**	57/58/80	60/58/64
Feature selection (75%)	**65/58/54**	61/58/56	58/58/80	56/57/80	60/58/68
Feature selection (100%)	60/55/50	62/60/58	**66/65/80**	57/53/80	62/58/67
Martinc *et al.* [28]	77/-/-	51/-/-	55/-/-	-	
Luz *et al.* [10]	-	75/-/-	-	-	-

cation metrics among all the classifiers and features combinations. Furthermore, the best classification results with Feature Selection are obtained when using only 15% of the features. Although we achieve high classification metrics with Feature Selection, the best classification results are obtained for the configuration using the LDA classifier and linguistic features only.

Table 4 shows the classification results obtained on the ADReSS dataset. With SVM classifier and 15% of the features, we achieve a maximum accuracy of 79%, 65% sensitivity and 60% specificity. As well as with DementiaBank dataset, the best classification results are obtained when a small subset of the features is selected. This is also the case for related works, such as Luz *et al.*[10], where the accuracy for their best performing system drops 4% (relative) when feature selection is not performed on their original set of 370 features. Again, linguistic features showed more capability to differentiate between HC and AD than the acoustic features.

6. Discussion

From the comparison between acoustic and linguistic features, we conclude that linguistic features provide significantly better classification results. Hence, they present more capability to distinguish between HC and AD than acoustic features. Recently and similar to our work, Fraser *et al.* [16] studied the potential of using linguistic features to identify Alzheimer's disease. In total, a set of 370 acoustic, lexical and semantic features were extracted and they obtained a highest accuracy of 92% in distinguishing between HC subjects and AD patients using the top 25 ranked features.

From the comparison between the results obtained on the ADReSS and DementiaBank datasets, it can be observed that better classification results are obtained on DementiaBank. One possible explanation is the number of training subjects in each dataset, which is considerably larger in DementiaBank (300 subject in DementiBank vs. 108 in ADReSS). Another possible explanation could be that ADReSS is gender and age balanced, whereas DementiaBank is not. Thus, it could be that, in the case of DementiaBank, the classifiers are able to infer the age and gender of each subject from the e.g., acoustic features to achieve a better performance. Besides, the recordings from DementiaBank have a higher level of background noise than the recordings from ADReSS, which causes that acoustic features contribute less to the classification results in comparison with the acoustic features on ADReSS dataset. Finally, as can be seen from Table 3 and 4, we have achieved higher classification results than the ones presented in the literature.

7. Conclusions

In this paper, we have proposed a binary classifier approach based on speech for dementia detection using acoustic and linguistic features derived from audio recordings and text transcriptions. Feature Selection has showed that the best features for this task are pause-based features, word embeddings and Conversational POS categories (nouns, verbs and determinants). We have evaluated our system on two well-known databases achieving state-of-the-art results for both. We have obtained the best classification results with LDA classifier and linguistic features for DementiaBank dataset, achieving an accuracy of 96%, a sensitivity of 97% and 92% of specificity. On the other hand, with ADReSS dataset we have achieved the maximum classification results with SVM classifier and 15% of the features, with an accuracy of 79%, a sensitivity of 65% and 60% of specificity. We therefore conclude from this results that linguistic features have the capability to distinguish from people with AD and healthy control.

In the future, we plan to investigate the use of more features as well as other machine learning algorithms. Furthermore, it would be interestint to evaluate the robustness of the system for detecting dementia in other databases or, even, in different languages.

8. References

[1] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[2] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in alzheimer patients," *Brain and Language*, vol. 17, no. 1, pp. 73–91, 1982.

[3] R. Jaffard, "Communication and cognition in normal aging and dementia," vol. 28, pp. 229–230, 1990.

[4] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," vol. 14, pp. 71–91, 2000.

[5] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech," vol. 2015, 2015, j AUDIO SPEECH MUSIC PROC. 2015, 9 (2015).

[6] J. Weiner and T. Schultz, *Selecting Features for Automatic Screening for Dementia Based on Speech: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings*, 01 2018, pp. 747–756.

[7] L. Cummings, "Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia," *Pragmatics and Society*, vol. 10, no. 2, pp. 153–176, 2019.

[8] J. Rodríguez, H. Martínez, and B. Valles, "Las pausas en el discurso de individuos con demencia tipo alzheimer. estudio de casos," *Revista de investigación en logopedia*, vol. 5, no. 1, pp. 40–59, 2015.

[9] [Online]. Available: https://talkbank.org/DementiaBank/. [Accessed:10-Dec- 2015]Dementia Bank."

[10] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2571

[11] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for alzheimer's disease," sLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies. 13 September 2016, San Francisco, USA.

[12] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german," in *Proc. Interspeech 2016*, 2016, pp. 1938–1942. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-100

[13] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1688 – 1701, 09 2011.

[14] S. Al-Hameed, M. Benaissa, and H. Christensen, "Detecting and predicting alzheimer's disease severity in longitudinal acoustic data," in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, ser. ICBRA 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 57–61. [Online]. Available: https://doi.org/10.1145/3175587.3175589

[15] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[16] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech." *Journal of Alzheimer's disease : JAD*, vol. 49, pp. 407–422, 2016.

[17] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, 2011.

[18] S. Al-Hameed, M. Benaissa, H. Christensen, B. Mirheidari, D. Blackburn, and M. Reuber, "A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints," vol. 14, p. e0217388, pLoS One. 2019;14(5):e0217388. Published 2019 May 24.

[19] L. Tan and M. Karnjanadecha, "Pitch detection algorithm: autocorrelation method and amdf," in *Proceedings of the 3rd international symposium on communications and information technology*, vol. 2, 2003, pp. 551–556.

[20] E. Reeve, P. Molin, A. Hui, and K. Rockwood, "Exploration of verbal repetition in people with dementia using an online symptom-tracking tool," vol. 29, pp. 959–966, 2017.

[21] H. Schmid, "Improvements in part-of-speech tagging with an application to german," pp. 13–25, 1999.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, pp. 3111–3119, 2013.

[23] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, p. 3–42, Apr. 2006. [Online]. Available: https://doi.org/10.1007/s10994-006-6226-1

[24] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations," 09 2018, pp. 1893–1897.

[25] M. Yancheva and F. Rudzicz, "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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2337–2346. [Online]. Available: https://www.aclweb.org/anthology/P16-1221

[26] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, "Diagnosing people with dementia using automatic conversation analysis," 09 2016, pp. 1220–1224.

[27] L. Breiman, "Random forests," vol. 45, pp. 261–277, 2001, machine Learning volume 45, pages5–32(2001).

[28] M. Martinc and S. Pollak, "Tackling the adress challenge: a multimodal approach to the automated recognition of alzheimer's dementia," *Proc. Interspeech 2020*, pp. 2157–2161, 2020.