INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity

Utkarsh Sarawgi[*], *Wazeer Zulfikar*[*], *Nouran Soliman, Pattie Maes*

Massachusetts Institute of Technology
{utkarshs, wazeer, nouran, pattie} @mit.edu

Abstract

Alzheimer's disease is estimated to affect around 50 million people worldwide and is rising rapidly, with a global economic burden of nearly a trillion dollars. This calls for scalable, cost-effective, and robust methods for detection of Alzheimer's dementia (AD). We present a novel architecture that leverages acoustic, cognitive, and linguistic features to form a multimodal ensemble system. It uses specialized artificial neural networks with temporal characteristics to detect AD and its severity, which is reflected through Mini-Mental State Exam (MMSE) scores. We first evaluate it on the ADReSS challenge dataset, which is a subject-independent and balanced dataset matched for age and gender to mitigate biases, and is available through DementiaBank. Our system achieves state-of-the-art test accuracy, precision, recall, and F1-score of 83.3% each for AD classification, and state-of-the-art test root mean squared error (RMSE) of 4.60 for MMSE score regression. To the best of our knowledge, the system further achieves state-of-the-art AD classification accuracy of 88.0% when evaluated on the full benchmark DementiaBank Pitt database. Our work highlights the applicability and transferability of spontaneous speech to produce a robust inductive transfer learning model, and demonstrates generalizability through a task-agnostic feature-space. The source code is available at `https://github.com/wazeerzulfikar/alzheimers-dementia`

Index Terms: Alzheimer's Dementia Detection, Affective Computing, Human-Computer Interaction, Computational Paralinguistics, Machine Learning, Speech Processing

1. Introduction

Alzheimer's disease is a progressive disorder that causes brain cells to degenerate and is the most common cause of dementia worldwide. It mainly causes cognitive and behavioural deterioration of the patients [1] which is reflected through memory loss, language impairment [2], and a decreased ability to express their needs. This in turn affects their quality of life, prognosis, and social relationships. Consequently, it has been imposing increased health risks [3] and a significant financial burden to patients, caregivers, families, and healthcare institutions [4]. The number of people with dementia worldwide in 2015 was estimated at 47.47 million, and reaching 135.46 million in 2050 [5]. At the time of writing this paper, someone in the U.S. develops Alzheimer's disease every 66 seconds, and by 2050 it is projected to be 33 seconds [6]. According to the World Health Organization, the global economic burden is nearly a trillion dollars which amounts to 1.1% of the global GDP. [7], with 63% of people with dementia living in low- and middle-income countries [8]. In this work, we aim to take a significant step towards more reliable, cost-effective, scalable, and non-invasive technologies to detect the onset of Alzheimer's dementia (AD) and predict the Mini-Mental State Exam [9] scores to estimate the severity of it.

Dementia can be strongly characterized by cognitive degeneration leading to language impairment which primarily occurs due to decline in semantic and pragmatic levels of language processing [10]. It has been widely reported that AD can be more sensitively detected with the help of a linguistic analysis than with other cognitive examinations [11] and also long before the diagnosis is medically confirmed [12]. The temporal characteristics of spontaneous speech, such as speech tempo, number of pauses in speech, and their length are sensitive detectors of the early stage of the disease [13, 14, 15, 16, 17]. Given the relative ease of collecting balanced and representative data of spontaneous speech and their corresponding transcriptions, they can be utilized in early and robust predictions for the onset of AD.

Consequently, our research work:

1. Presents a novel architecture comprising of domain-specific feature engineering and artificial neural networks for Alzheimer's Dementia (AD) detection and its severity through classification and MMSE score regression (Section 3).
2. Evaluates the system in a subject-independent setting with a carefully curated balanced and stratified dataset matched for age and gender, to help minimize common biases in the tasks (Section 3.1).
3. Achieves state-of-the-art test accuracy, precision, recall, and F1-score for AD classification, and state-of-the-art test RMSE for MMSE score predictions on the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) dataset. To the best of our knowledge, the system further achieves state-of-the-art AD classification accuracy when evaluated on the full benchmark DementiaBank Pitt database (Sections 4 and 5).
4. Spans a multimodal feature space to increase generalizability and robustness, and uses ensemble mechanisms to leverage individual feature sets and model performances.
5. Reflects upon the transferability and interdependence of the two tasks of AD classification and MMSE regression.

2. Related work

Many current AD detection studies use medical imaging [18, 19, 20] with deep neural networks and random forests. Several studies claim that AD can be sensitively detected in early stages by doing linguistic analysis which leverages speech and language features to train machine learning models for the detection of AD [13, 14, 15, 16, 17, 21].

In study [22], machine learning methods based on image description were used reaching an accuracy of 75% on a limited

[*]Equal Contribution

http://dx.doi.org/10.21437/Interspeech.2020-3137

number of subjects enrolled in a longitudinal study. Study [23] used logistic regression trained with spectrogram features extracted from audio files reaching accuracy of 83.3% and 84.4% on VBSD and Dem@Care datasets respectively. Data used in each of the above works are limited to around 32 to 36 subjects and highly imbalanced between the classes and across age and gender. In study [14], different traditional classification algorithms like logistic regression, SVM, and more were used to learn speech parameters from dialogues in Carolina Conversations Collection. The best of their solutions reached 86.5% leave-one-out cross-validation (LOOCV) accuracy with 38 subjects. Works based on data extracted from DementiaBank have reported scores of around 0.87, 0.85, 0.82, 0.80, 0.79, 0.64, and 0.62 [24, 25, 13, 26, 27, 28, 29] for AD classification. Study [30] used speech related features to get a mean absolute error (MAE) of 3.83 for MMSE scores with longitudinal data derived from DementiaBank. While a number of works have proposed speech and language based approaches to AD recognition through speech, their studies have used different, often unbalanced and acoustically varied data sets, thereby introducing bias and hindering generalization, reproducibility and comparability of the proposed approaches.

3. Methods and materials

3.1. Dataset

The DementiaBank Pitt database [31] consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [32]. The database consists of multiple samples per subject corresponding to multiple visits. The full database contains 242 speech samples from 99 control healthy subjects and 255 speech samples from 168 AD subjects. The dataset also provides Mini-Mental Status Examination (MMSE) scores, ranging from 0 to 30, of the subjects, which offers a way to quantify cognitive function and screen for cognitive loss by testing the individuals' orientation, attention, calculation, recall, language and motor skills [9]. A 10-fold cross-validation was used on this database for fair comparison with previously reported results.

The ADReSS Challenge Dataset [29] is a balanced subset consisting of 156 speech samples, each from a unique subject, matched for age and gender and evenly spread across the two classes, AD and non-AD. A stratified train-test split of around 70-30 (108 and 48 subjects) for this dataset was provided by the challenge. The test set was held out for all experimentation until final evaluation. Any cross-validation mentioned in the paper refers to cross-validation using the train split. Normalized speech segments are also provided, but we only use full audio samples. The MMSE scores provided are used as labels for the regression task.

We first evaluate on the balanced ADReSS dataset and then extend the evaluation to the full DementiaBank Pitt database.

3.2. Feature engineering

People with dementia show symptoms of cognitive decline, impairment in memory, communication, and thinking [17]. To include such domain knowledge and context, our system extracts cognitive and acoustic features using three different strategies, which are then prepared and fed into their respective neural models. Similarly extracted features have been repeatedly used to propose speech recognition based solutions for automated detection of mild cognitive impairment from spontaneous speech [33, 17]. The following features were extracted upon exploring the data to find the most descriptive set of correlated features for detecting AD and its severity:

• *Disfluency:* A set of 11 distinct and carefully curated features from the transcripts, like word rate, intervention rate, and different kinds of pause rates reflecting upon speech impediments like slurring and stuttering. These are normalized by the respective audio lengths and scaled thereafter.

• *Acoustic:* The ComParE 2013 feature set [34] was extracted from the audio samples using the open-sourced openSMILE v2.1 toolkit, widely used for affect analyses in speech [35]. This provides a total of 6,373 features that include energy, MFCC, and voicing related low-level descriptors (LLDs), and other statistical functionals. This feature set encodes changes in speech of a person and has been used as an important noninvasive marker for AD detection [36, 29]. Our system standardizes this set of features using z-score normalization, and uses principal component analysis (PCA) to project the 6,373 features onto a low-dimensional space of 21 orthogonal features with highest variance. The number of orthogonal features was selected by analyzing the percentage of variance explained by each of the components.

• *Interventions:* Cognitive features reflect upon potential loss of train of thoughts and context. Our system extracts the sequence of speakers from the transcripts, categorizing it as subject or the interviewer. To accommodate for the variable length of these sequences, they are padded or truncated to length of 32 steps, found upon analyses and tuning of sequence lengths.

We evaluated each of these features individually and in a combined fashion to highlight the different configurations and compare their performances.

3.3. Model architecture and training

Figure 1 - (1), (2), and (3) illustrate the architecture of the disfluency, acoustic, and interventions models respectively. The disfluency model is a multi-layer perceptron (MLP) that projects the 11-feature input to a higher dimensional space for better separability of the binary classes. The acoustic model is an MLP with a single hidden layer that adds non-linearity and regularizes the PCA decomposed feature space. The interventions model uses a recurrent architecture to learn the temporal relations from the sequence of interventions. These models were trained with corresponding inputs obtained upon feature engineering (Section 3.2), and one-hot encoded binary class labels.

To leverage the features learnt from classification for regression, transfer learning was done on the trained classification models. The regression module, as shown in Figure 1 - (4) replaced the terminal output layer in the models and the remaining original layers were frozen. The resultant models were then trained with MMSE scores as labels.

A 5-fold cross-validation setting was adopted for evaluation. The models were also evaluated in a leave-one-out cross validation (LOOCV) setting, which in the case of ADReSS dataset is equivalent to leave-one-subject-out cross validation (LOSO) since each datapoint is an independent subject. Each training run used a batch size of 8; and Adam optimizer with a learning rate of 0.01 to minimize categorical cross-entropy loss for classification, and a learning rate of 0.001 to minimize mean squared error loss for regression. The best models were saved by monitoring the validation loss in each fold.

To leverage all sets of features and models together, a parallel ensemble was performed using the outputs of the three models for each of the two tasks independently. We experimented

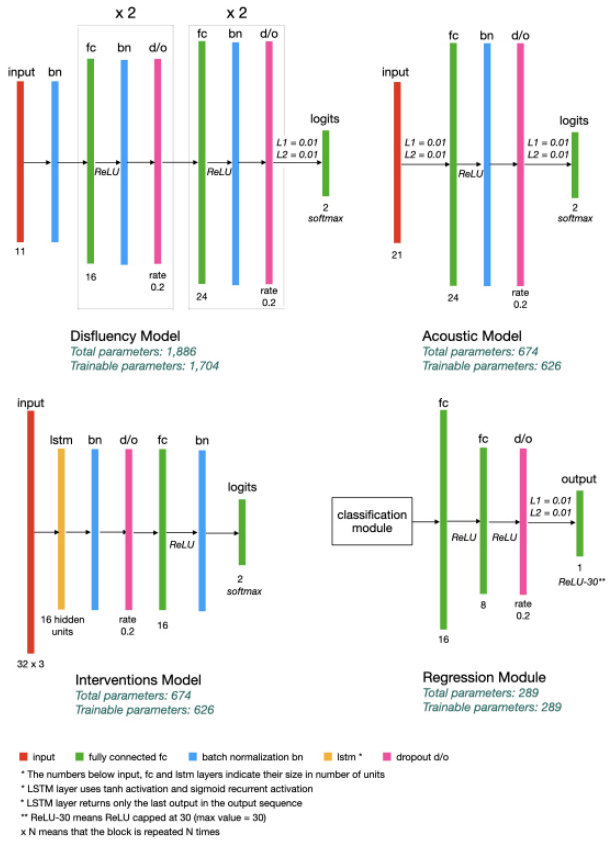

Figure 1: *Architecture of (1) Disfluency, (2) Acoustic, (3) Interventions models, and (4) Regression module.*

with three kinds of ensemble modules for classification:

• *Hard:* A majority vote was taken between the predictions of the three individual models.

• *Soft:* To leverage the confidence of the predictions, a weighted sum of the class probabilities was computed for final decision. The weight used was $1/N$ where N is the total number of models.

• *Learnt:* Instead of weighing the confidence of all the models equally as in soft voting above, we used a logistic regression to learn the weights. A logistic regression voter was trained using class probabilities as inputs.

For regression, the predictions of all the individual models were averaged by the ensemble module.

4. Results

The results of the experiments were recorded using a combination of accuracy, precision, recall and F1-score for classification, and root mean squared error (RMSE) for regression.

4.1. ADReSS Challenge dataset

Table 1 shows the 5-fold cross-validation results for the classification task. The individual features achieved competitive performance, although the acoustic model slightly overfits while the interventions model marginally underfits on the data. The ensemble model counteracted these and achieved an increased 5-fold mean training as well as validation accuracy with comparable variance. The low variance generally observed across all runs signifies high model stability across folds which is essential in small datasets. Similar observations can be seen on the regression task in Table 1, where the ensemble model reduced the train and validation mean RMSE as well as the variance. This is consistent with the intuition behind using transfer learning using the trained classification models through the addition of a regression module.

Table 1: *5-fold cross validation results of the models. Accuracy measures the AD classification performance while RMSE measures the MMSE score regression performance over all 5 folds. Ensemble in this table refers to hard ensemble for classification and the regression ensemble for regression.*

Model	Split	Accuracy	RMSE
Disfluency	Train	0.87 ± 0.08	4.37 ± 0.40
	Val	0.89 ± 0.05	4.87 ± 0.78
Acoustic	Train	0.89 ± 0.03	4.40 ± 0.64
	Val	0.83 ± 0.07	5.63 ± 1.15
Interventions	Train	0.82 ± 0.06	5.05 ± 0.56
	Val	0.89 ± 0.04	4.70 ± 0.96
Ensemble	Train	$\mathbf{0.91 \pm 0.04}$	$\mathbf{3.65 \pm 0.38}$
	Val	$\mathbf{0.92 \pm 0.06}$	$\mathbf{4.26 \pm 0.75}$

Table 2: *5-fold cross-validation accuracies of different ensemble mechanisms for AD classification.*

Ensemble Type	Split	Accuracy
Hard	Train	0.91 ± 0.04
	Val	$\mathbf{0.92 \pm 0.06}$
Soft	Train	0.86 ± 0.04
	Val	0.86 ± 0.04
Learnt	Train	$\mathbf{0.95 \pm 0.03}$
	Val	0.81 ± 0.08

The improvement in performance upon ensembling the three models as compared to the individual models further reflects upon the significance of leveraging acoustic and cognitive features together from multimodal speech and text inputs.

Table 2 shows the 5-fold cross validation results of different parallel ensemble techniques, discussed in Section 3.3, for the classifiation task. The learnt ensemble showed signs of overfitting due to the extra trainable parameters in the model. The soft and hard ensemble helped counter this. However, the hard ensemble proved to be the most competitive by improving training and validation accuracies along with a strong degree of generalization across folds.

Figure 2 shows the receiver operating characteristic (ROC) curve for the individual models on the classification task. The ROC is cumulatively calculated over the validation splits of all 5 folds of cross-validation.

We compare our results with the currently available baseline performance results on this dataset [29]. Amongst our models, the best performing model, the hard ensemble classification model and the ensemble regression model, considerably improved all the metrics on the LOSO as well as the held-out test set on AD classification and regression, as can be seen in Table 3 and Table 4 respectively.

The confusion matrices in Figure 3 provide further insights into the predictions of the hard ensemble classification model that has been compared with the baseline in Table 3.

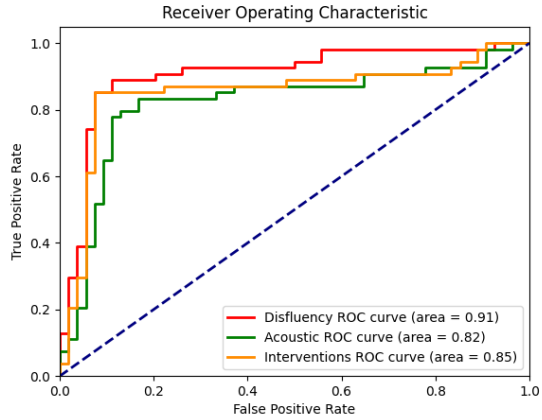

Figure 2: *Receiver Operating Characteristic for Disfluency, Acoustic, and Interventions models, cumulatively calculated over validation splits of all the folds of 5-fold cross-validation.*

Table 3: *Baseline comparison of the AD classification. Our test results below are corresponding to the hard ensemble model.*

	Model	Accuracy	Precision	Recall	F1-Score
LOSO	Luz et al. [29]	0.77	0.77	0.76	0.77
	Ensemble (*ours*)	**0.99**	**0.99**	**1.00**	**0.99**
TEST	Luz et al. [29]	0.75	0.83	0.62	0.71
	Ensemble (*ours*)	**0.83**	**0.83**	**0.83**	**0.83**

Table 4: *Baseline comparison of the MMSE score regression. Our test results are corresponding to the regression ensemble.*

	Model	RMSE
LOSO	Luz et al. [29]	4.38
	Ensemble (*ours*)	**0.82**
TEST	Luz et al. [29]	5.20
	Ensemble (*ours*)	**4.60**

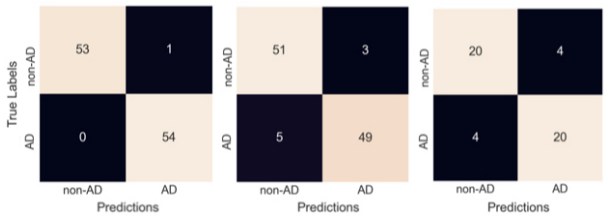

Figure 3: *Confusion matrices for the hard ensemble classification model (1) cumulatively calculated over the validation splits of all the folds of LOOCV and (2) 5-fold cross-validation, and (3) calculated on the held out test set.*

4.2. DementiaBank Pitt database

The same AD classification models were retrained on the DementiaBank Pitt database and a 10-fold cross-validation was performed for fair comparison with previously reported results. To the best of our knowledge, our hard ensemble model achieves state-of-the-art 0.88 $\pm$ 0.04 accuracy, also showing minimal variance across the folds (Table 5).

Table 5: *Comparison of the AD classification on DementiaBank Pitt. All are 10-fold cross-validation results. Our results below are corresponding to the hard ensemble model.*

Model	Accuracy	Precision	Recall	F1-Score
Fraser et al. [13]	0.82	-	-	-
Masrani [25]	0.85	-	-	0.85
Kong et al. [24]	0.87	0.86	**0.91**	**0.88**
Ensemble (*ours*)	**0.88**	**0.92**	0.82	**0.88**

5. Discussion and Future Work

There has been substantial work using spontaneous speech samples and manual transcriptions present in the DementiaBank dataset [31]. Some of the highest reported scores for AD classification are 0.87, 0.85, 0.82, 0.80, 0.79, 0.64, and 0.63 [24, 25, 13, 26, 27, 28, 29]. Many of these previous results were obtained on datasets with variable subject dependencies. In such datasets, a data point corresponds to a session and there can exist multiple sessions per subject. Given the subject independent setting in ADReSS dataset, our LOSO method clearly distinguishes the left-out test subject. Hence, the near perfect LOSO results on classification and regression (Tables 3 and 4) demonstrate that every subject individually can be correctly evaluated with the engineered features. Furthermore, almost all previous results are reported using cross-validation, whereas our work is evaluated on a designated held-out test set as well. This helps overcome 'validation overfitting' which is prone in small dataset settings.

Study [30] used speech related features to obtain a cross-validated mean absolute error (MAE) of 3.83 for MMSE scores with data derived from DementiaBank. Our ensemble regression model recorded a cross-validated MAE of 3.01 on ADReSS dataset.

Through considerable improvements in both the AD classification and MMSE score regression by employing an ensemble of independent models extracting acoustic and cognitive features, our work reveals the potential of multimodal analysis and its applicability to a age and gender balanced subject-independent dataset. Future work would include incorporating automated transcription of speech samples in our system. The continuous range of the MMSE scores can provide more insights into progression of dementia. This can further be leveraged for risk stratification and analyzing potential causal relationships modelling AD with its symptoms and markers, through a longitudinal dataset.

6. Conclusion

We present a novel architecture that uses domain knowledge for inductive transfer learning for AD classification and MMSE score regression. Our work achieves state-of-the-art accuracy, precision, recall, and F1-score of 83.3% each for AD classification, and state-of-the-art RMSE of 4.60 for MMSE predictions on the designated held-out test set of the ADReSS challenge. To the best of our knowledge, the system further achieves state-of-the-art AD classification accuracy of 88.0% when evaluated on the full benchmark DementiaBank Pitt database. Our system spans a multimodal feature space to increase generalization and robustness. We aim to extend our work by adding automated transcription, further textual analysis, and personalized context through longitudinal data.

7. References

[1] J. G. Molinuevo, "Role of biomarkers in the early diagnosis of alzheimer's disease," *Revista espanola de geriatria y gerontologia*, vol. 46, pp. 39–41, 2011.

[2] L. M. V. ESCOBAR and N. P. AFANADOR, "Calidad de vida del cuidador familiar y dependencia del paciente con alzheimer," *Avances en Enfermería*, vol. 28, no. 1, pp. 116–128, 2010.

[3] R. Schulz and S. R. Beach, "Caregiving as a risk factor for mortality: the caregiver health effects study," *Jama*, vol. 282, no. 23, pp. 2215–2219, 1999.

[4] J. M. Atance, A. I. Yusta, and B. G. Grupeli, "Costs study in alzheimer's disease," *Revista clinica espanola*, vol. 204, no. 2, pp. 64–69, 2004.

[5] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: a systematic review and metaanalysis," *Alzheimer's & dementia*, vol. 9, no. 1, pp. 63–75, 2013.

[6] A. Association *et al.*, "2016 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 12, no. 4, pp. 459–509, 2016.

[7] W. H. Organization *et al.*, "The top 10 causes of death. fact sheet no. 310. 2017."

[8] ——, "The epidemiology and impact of dementia. current state and future trends. geneva, switz: World health organization; 2015."

[9] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[10] S. H. Ferris and M. Farlow, "Language impairment in alzheimer's disease and benefits of acetylcholinesterase inhibitors," *Clinical interventions in aging*, vol. 8, p. 1007, 2013.

[11] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[12] M. Mesulam, A. Wicklund, N. Johnson, E. Rogalski, G. C. Léger, A. Rademaker, S. Weintraub, and E. H. Bigio, "Alzheimer and frontotemporal pathology in subsets of primary progressive aphasia," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 63, no. 6, pp. 709–719, 2008.

[13] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[14] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," *arXiv preprint arXiv:1811.09919*, 2018.

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

[16] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[17] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: a review," *Expert Systems with Applications*, p. 113213, 2020.

[18] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[19] A. Ortiz, F. Lozano, J. M. Gorriz, J. Ramirez, F. J. Martinez Murcia, A. D. N. Initiative *et al.*, "Discriminative sparse features for alzheimer's disease diagnosis using multimodal image data," *Current Alzheimer Research*, vol. 15, no. 1, pp. 67–79, 2018.

[20] S. Sarraf and G. Tofighi, "Deep learning-based pipeline to recognize alzheimer's disease using fmri data," in *2016 Future Technologies Conference (FTC)*. IEEE, 2016, pp. 816–820.

[21] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," *arXiv preprint arXiv:1906.05483*, 2019.

[22] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard, "Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology," *Journal of Alzheimer's Disease*, vol. 42, no. s3, pp. S3–S17, 2014.

[23] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying alzheimer's disease," *Simulation Modelling Practice and Theory*, vol. 99, p. 102023, 2020.

[24] W. Kong, H. Jang, G. Carenini, and T. Field, "A neural model for predicting dementia from language," in *Machine Learning for Healthcare Conference*, 2019, pp. 270–286.

[25] V. Masrani, "Detecting dementia from written and spoken language," Ph.D. dissertation, University of British Columbia, 2018.

[26] 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)*, 2016, pp. 2337–2346.

[27] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 260–268, 2018.

[28] 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.

[29] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[30] M. Yancheva, K. C. Fraser, and F. Rudzicz, "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*, 2015, pp. 134–139.

[31] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[32] H. Goodglass, E. Kaplan, and B. Barresi, *BDAE-3: Boston Diagnostic Aphasia Examination–Third Edition*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[33] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[34] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[36] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, M. Faundez-Zanuy, M. Ecay-Torres, C. M. Travieso, A. Ezeiza, A. Estanga *et al.*, "Alzheimer disease diagnosis based on automatic spontaneous speech analysis," 2012.