Automatic Classification of Dementia Using Text and Speech Data

Hee Jeong Han, Suhas B. N., Ling Qiu, and Saeed Abdullah

Abstract Dementia is a serious public health concern. It does not have any cure. Early detection of dementia is, thus, critical for effective symptom management as well as delaying cognitive and functional decline. This paper focuses on detecting onset of dementia using text and speech features provided by two publicly available datasets from the AAAI 2022 hackallenge. Our approach resulted in developing ACOUSTICS (AutomatiC classificatiOn of sUbjectS with demenTIa and healthy Controls using text transcriptions and Speech data)—an ensemble model with two deep learning-based architectures for text and speech analysis. ACOUSTICS achieved 89.8% accuracy when classifying individuals with dementia and health controls. Our approach outperforms current state-of-the-art methods in dementia detection.

1 Introduction

Dementia is a chronic neurodegenerative disorder that gradually causes cognitive and functional decline [34]. Dementia results in gradual decline in cognitive and functional abilities including memory loss, cognitive impairment, and worsening communication and language skills. More than 55 million individuals worldwide have dementia, with nearly 10 million new cases diagnosed each year [34]. It is currently the seventh leading cause of mortality among all illnesses, as well as one of the major causes of impairment and reliance among the elderly [34]. There is

H. J. Han (✉) · S. B. N. · L. Qiu · S. Abdullah
College of Information Sciences and Technology, Pennsylvania State University, E397 Westgate Building, University Park, PA 16802, USA
e-mail: heejeonghan@psu.edu

S. B. N.
e-mail: sxb701@psu.edu

L. Qiu
e-mail: lzq5034@psu.edu

S. Abdullah
e-mail: saeed@psu.edu

A. Shaban-Nejad et al. (eds.), *Multimodal AI in Healthcare*, Studies in Computational Intelligence 1060, https://doi.org/10.1007/978-3-031-14771-5_29

currently no cure for dementia. Early detection of dementia is, thus, critical for effective symptom management as well as delaying cognitive and functional decline [5].

Current methods for dementia detection uses different assessment strategies methods including cognitive assessments (e.g., Mini-Mental State Examination—MMSE [15]), self-report questionnaires, and neuroimaging (e.g., Positron Emission Tomography—PET [30]). However, these diagnosis methods can be infrequent and time-consuming [27], which can hinder early detection of dementia. Furthermore, the lack of accessible methods can be particularly problematic for individuals living at remote locations. As such, there is an urgent need for developing methods that can be deployed at scale without adding burden to individuals.

Research shows that changes in speech or language usage can indicate early sign of cognitive decline [1]. As such, recent studies have leveraged language and speech characteristics to develop novel machine learning-based approaches to identify dementia onset [2, 8, 12, 14, 19, 33]. In this work, we aim to advance the state-of-the-art in developing computational approaches to detect dementia onset using text and speech features. Specifically, we have developed an ensemble called ACOUSTICS using two publicly available datasets—the Pittsburgh (Pitt) corpus [3], and the Wisconsin Longitudinal Study (WLS) corpus [22]. ACOUSTICS consists of two deep learning-based architectures for text and speech data processing.

To process text data, we generated a BERT model based on transcription and build a deep learning model based on a miniature version of the Xception network [9]. To process speech data, we converted audio data to spectrograms. Using spectrograms features allows us to exploit spatio-temporal structures and relations present in the speech data. The resultant ensemble model demonstrates 89.8% accuracy in classifying participants living with dementia and healthy controls. This model performance represents an improvement over prior work using a similar dataset from the previous two years with the baseline scores of 84.8% [18], and 87.2% [10]. Our approach, thus, outperforms current state-of-the-art methods in dementia detection.

2 Related Work

In recent years, there has been some significant progress in developing machine learning models for assessing dementia onset using audio and text features [2, 8, 12, 14, 19, 33]. For example, Karlekar et al. [23] proposed to use a Convolutional Neural Network (CNN), a long short-term memory network (LSTM), and CNN-LSTM to detect dementia using the conversational transcripts with word embeddings and parts-of-speech (POS) tags. However, the lack of standardized datasets and performance benchmarks has been a challenge for this domain.

Luz et al. [25] developed the ADReSS challenge dataset to support standardized model development and evaluation focusing on dementia assessment. The ADReSS challenge focused on differentiating between individuals with dementia and healthy controls using a subset of the Pitt corpus. A number of recent papers have used this

Table 1 A comparison of previous works based on the DementiaBank dataset in the ADReSS challenge looked at developing models and evaluation metrics for dementia assessment

Work	Accuracy (%)	Methodology
Haulcy and Glass [20]	85.4	i-vectors and x-vectors with SVM and RF
Yuan et al. [35]	**89.6**	Encoding pauses in transcripts using BERT and ERNIE
Shah et al. [29]	85.4	Acoustic and Language features with Regression
Meghanani et al. [26]	83.3	n-grams from transcripts with CNN's

dataset to develop and evaluate machine learning models for dementia assessment (see Table 1).

For example, Haulcy et al. [20] investigated the use of i-vectors and x-vectors, acoustic features initially designed for speaker identification, and linguistic features to address dementia detection and MMSE prediction. The i-vectors and x-vectors were pre-trained on datasets unrelated to dementia as well as data from the domain. Several classification and regression models were tested, with SVM and Random Forests yielding 85.4% accuracy in dementia detection and a gradient boosting regressor having 4.56 RMSE. The authors hypothesized that the poor performance of i-vectors and x-vectors was caused by a mismatch in in- and out-of-domain training data.

Yuan et al. [35] used BERT and ERNIE to fine-tune the training of language models by encoding filled and unfilled pauses in transcripts. In the dataset, the authors observed that individuals with dementia used the phrase "um" much less frequently than "uh", and their language samples contained more pauses. The detection of dementia has improved to 89.6% accuracy (with ERNIE).

Shah et al. [29] used speech samples from the DementiaBank database for binary classification and MMSE regression. Despite developing models that combined acoustic and language-based features, their best performing model for binary classification used only language-based features with a regularized logistic regression and achieved 85.4% accuracy on a hold-out test set. With an RMSE of 5.62, their best performing model for the regression task was a more limited set of language features.

Meghanani et al. [26] compared two approaches to the challenge tasks based on the use of manual, non-automatic transcripts. Both methods relied on n-grams of varying lengths (n = 2, 3, 4, and 5) extracted from transcripts. The first method used CNNs with a single convolutional layer, with the kernel size adjusted to fit the n-gram size. The fastText model was used with bigrams and trigrams in the second method. The fastText models outperformed the CNN models, achieving 83.3% classification accuracy and an RMSE of 4.87 for predicting MMSE scores.

3 Materials and Methods

3.1 Dataset: DementiaBank Pitt and WLS Corpora

In this work, we used two datasets: The Pitt and WLS corpora. The dataset consists of participants' demographic information, diagnostic data (e.g., MMSE), audio recordings, and transcriptions of the audio recordings. The audio records in the two datasets are from participants conducting the "Cookie Theft" task in the Boston Diagnostic Aphasia Exam [17]. Prior work has used the "Cookie Theft" task to identify recurrent cognitive-linguistic impairments including dementia [4, 11, 16]. Both datasets also provide various diagnostic data (e.g., diagnostic code, MMSE score, and fluency score).

3.2 Data Pre-processing

To label the Pitt corpus, We leveraged MMSE scores [15] to distinguish between healthy controls and individuals with dementia. There are two reasons for this approach. Firstly, MMSE is a clinically established approach to evaluate the cognitive functionality for various populations [31]. Secondly, in this dataset, MMSE has much fewer missing values compared to other variables. Individuals with MMSE scores $\leq$24 (including 24) are labeled as having dementia [24]. The rest of the individuals were labeled as healthy controls. The Pitt corpus contains 292 participants with 552 audio recordings. Participants with multiple visits can have multiple audio recordings. Since cognitive functionality might change overtime, we decided to label their audio recordings with their corresponding MMSE scores. We removed 93 out of 552 audio files with missing MMSE scores. There were 242 audio files for healthy controls and the remaining 217 audio files were classified as individuals with dementia. Lastly, we selected 323 audio files (152 dementia and 171 non-dementia) from the Pitt corpus dataset as the training set. We used the remaining 136 audio files (65 dementia and 71 non-dementia) as the test set.

For the WLS corpus, we used verbal fluency to distinguish between healthy controls and patients with dementia. Similar approach has been used by prior work (e.g., [18]). Participants from the WLS corpus completed two category verbal fluency cognitive tests. They were asked to name as many words as they could that belong to a specified category (animal and food in this case) within 1 minute. Research shows that the verbal fluency test can effectively detect dementia in the clinical setting [21]. Following prior work [18], we used fluency cutoff scores accounting for age. This is, we used decreasing cutoff scores for higher age with thresholds of 16, 14, and 12 for participants aged less than 60, between 60 and 79, and more than 79, respectively. Using this approach, we identified 23 participants in the WLS corpus as individuals with dementia. The rest of the participants (N = 93) were labeled as healthy controls. We used the same train-test split approach for the WLS corpus, which resulted in 79

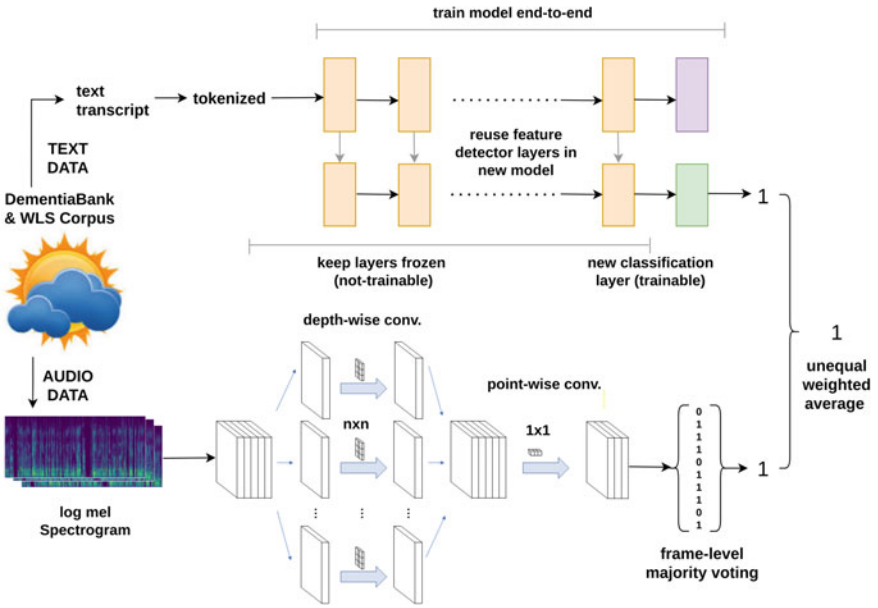

Fig. 1 Our approach resulted in developing ACOUSTICS (AutomatiC classificatiOn of sUbjectS with demenTIa and healthy Controls using text transcriptions and Speech data)—an ensemble model with two deep learning-based architectures for text and speech analysis. ACOUSTICS is trained on two separate BERT and Xception Networks. This is followed by majority-voting the frame-wise predictions and ensembling the outputs to provide one prediction for each subject

participants (16 dementia and 63 non-dementia) in the training set. The remaining 37 participants (7 dementia and 30 non-dementia) were used for the test set.

3.3 Feature Extraction

We used the Harmonized Pre-processing Toolkit[1] for Dementia Bank to extract text features from transcription data. For the audio preprocessing, we first converted the audio file format from mp3 to wav. We also downsampled the audio data from 44.1 to 16 kHz. This is due to the fact that human speech ranges from 0 to 8 kHz. Furthermore, it also reduces file size considerably without losing valuable information. We used the provided timestamp data (i.e., participant start-stop times in .cha files) to trim corresponding audio files. We only retained participant speech information following this step. We then extracted the log-mel spectrogram features by using use overlapping windows with 1-sec duration (Fig. 1).

[1] https://github.com/LinguisticAnomalies/harmonized-toolkit

4 Ensemble Model

4.1 Deep Learning Model

For data processing, we developed ACOUSTICS—an end-to-end ensemble model that consists of two deep learning-based architectures. First, we generated a deep learning model to classify participants based on transcribed audio recordings using pre-trained Transformer-based architectures [32], focusing on the Bidirectional Encoder Representations from Transformers (BERT) model [13]. The BERT model with Transformer-based architectures allows contextual learning relations between words [32]. Therefore, the model can examine the larger context while resolving ambiguities in contextualized meaning. We implemented a classification layer to get binary class labels corresponding to "positive" (has dementia) and "negative" (no dementia). The model's accuracy is 77.5%, the equal error rate is 0.25, and the AUC score is 0.75.

Second, we generated a deep learning model to classify participants based on speech using log-mel spectrogram features. Using spectrograms features allows us to exploit spatio-temporal structures and relations present in the speech data. We adopted a miniature version of the Xception network [9] for developing our deep learning model. The resultant model has an accuracy of 94.2%, the equal error rate is 0.32, and the AUC score is 0.918.

The ensemble model the combines two deep-learning-based models to detect whether an individual has dementia. As mentioned above, both the two deep learning models have excellent accuracy. Combining their outcomes in an ensemble can lead to a robust classification.

4.2 Model Evaluation

For the audio features, we considered all the spectrogram images of a given subject which resulted in a list of predictions for each image. We obtained a single score of dementia (1) or healthy control (0) by performing majority voting on all the speech frames. This is the first intermediate output of our ensemble model. We obtain a single score of dementia (1) or healthy control (0) by feeding in tokenized data for the text model. This is the second intermediate output of our ensemble model.

Table 2 Comparison of our dataset selection with baseline data evaluated by training and testing on the same network architecture. For the same network architecture, there is an improvement in our data highlighting the usefulness of our selection criteria based on the MMSE scores

Work	Accuracy (%)	Methodology
Baseline data + our ensemble network	87.2	Acoustic (Modified Xception Network) and Language Model (BERT) with unequal weighted average
Ours − Audio only	**94.2**	
Ours − Text only	77.5	
Ours − Ensemble of Audio + Text	89.8	

Based on the two intermediate outputs from the two models, we perform an unequal weighted average of their accuracies at the classification layer to provide a single output prediction. In these datasets, our ensemble model achieves an average five-fold accuracy of 89.8% (S.D = 3.3%) (Table 2).

5 Discussion

There is much work to be done in the field of identifying dementia. With an aging population, the number of people affected by this disease will continue to grow [6, 28]. It is crucial that we continue researching new ways to identify and diagnose dementia as early as possible. This will allow us to provide treatment and support for those affected and help us find a cure for this devastating disease. Our development of ACOUSTICS is a step toward this goal of early detection of dementia at scale.

Many promising new technologies focusing on different types of data may help us in our quest to identify dementia early. For example, brain imaging technology has improved dramatically in recent years and can now be used to detect changes in brain function that may indicate the presence of dementia. Another promising avenue of research is investigating biomarkers—both invasive (ex. blood) and non-invasive (such as sweat/urine or speech used in this work) that may serve as indicators of Alzheimer's disease or other forms of dementia. Future work should aim to develop models that combine these biomarkers with speech and language data to identify onset of dementia.

6 Conclusion

In this project, we developed two deep learning models using the Pitt corpus and the WLS corpus to classify participants living with dementia and healthy controls.

Our approach led to developing ACOUSTICS—an ensemble model with two deep-learning architectures for text and speech data. The resultant performance of ACOUSTICS was highly promising with an accuracy of 89.8%. Future work should evaluate our approach across more diverse datasets to further assess its robustness in identifying dementia onset.

References

1. Antonsson, M., Lundholm Fors, K., Eckerström, M., & Kokkinakis, D. (2021). Using a discourse task to explore semantic ability in persons with cognitive impairment. *Frontiers in Aging Neuroscience*, *495*.
2. Balagopalan, A., Eyre, B., Rudzicz, F., & Novikova, J. (2020). To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. arXiv:2008.01551.
3. Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51*(6), 585–594.
4. Bird, H., Ralph, M. A. L., Patterson, K., & Hodges, J. R. (2000). The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language, 73*(1), 17–49.
5. Briggs, R., Kennelly, S. P., & O'Neill, D. (2016). Drug treatments in Alzheimer's disease. *Clinical Medicine, 16*(3), 247.
6. Brodaty, H., Breteler, M. M., DeKosky, S. T., Dorenlot, P., Fratiglioni, L., Hock, C., ... & De Strooper, B. (2011). The world of dementia beyond 2020. *Journal of the American Geriatrics Society, 59*(5), 923–927.
7. Calzà, L., Gagliardi, G., Favretti, R. R., & Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language, 65*, 101113.
8. Chen, L., Dodge, H. H., & Asgari, M. (2020). Topic-based measures of conversation for detecting mild cognitive impairment. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations (Virtual)* (pp. 63–67).
9. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1251–1258).
10. Cohen, T., & Pakhomov, S. (2020). A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. arXiv:2005.03593.
11. Cummings, L. (2019). Describing the cookie theft picture: Sources of breakdown in Alzheimer's dementia. *Pragmatics and Society, 10*(2), 153–176.
12. De la Fuente Garcia, S., Ritchie, C., & Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease, 78*, 1547–1574. https://doi.org/10.3233/JAD-200888
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.
14. Eyre, B., Balagopalan, A., & Novikova, J. (2020). Fantastic features and where to find them: detecting cognitive impairment with a subsequence classification guided approach. In *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020) (Virtual)* (pp. 193–199). https://doi.org/10.18653/v1/2020.wnut-1.25.
15. Folstein, M. F., Robins, L. N., & Helzer, J. E. (1983). The mini-mental state examination. *Archives of General Psychiatry, 40*(7), 812–812.
16. Giles, E., Patterson, K., & Hodges, J. R. (1996). Performance on the boston cookie theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology, 10*(4), 395–408.

17. Goodglass, H., Kaplan, E., & Weintraub, S. (2001). *BDAE: The Boston diagnostic aphasia examination*. Philadelphia, PA: Lippincott Williams & Wilkins.
18. Guo, Y., Li, C., Roan, C., Pakhomov, S., & Cohen, T. (2021). Crossing the "Cookie Theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Frontiers in Computer Science, 3*, 26. Chicago.
19. Haider, F., De La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 272–281.
20. Haulcy, R. M., & Glass, J. (2021). Classifying Alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology, 11*, 624137.
21. Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia, 42*(9), 1212–1222.
22. Herd, P., Carr, D., & Roan, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology, 43*(1), 34–41.
23. Karlekar, S., Niu, T., & Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. arXiv:1804.06440.
24. Kvitting, A. S., Fallman, K., Wressle, E., & Marcusson, J. (2019). Age-normative MMSE data for older persons aged 85 to 93 in a Longitudinal Swedish Cohort. *Journal of the American Geriatrics Society, 67*(3), 534–538.
25. Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The adress challenge. arXiv:2004.06833.
26. Meghanani, A., Anoop, C. S., & Ramakrishnan, A. G. (2021). Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fasttext and cnn models. *Frontiers in Computer Science*, *7*.
27. Prabhakaran, G., & Bakshi, R. (2018). Analysis of structure and cost in a longitudinal study of Alzheimer's disease. *Journal of Health Care Finance*, *44*(3).
28. Prince, M., Guerchet, M., & Prina, M. (2013). *The global impact of dementia 2013-2050*.
29. Shah, Z., Sawalha, J., Tasnim, M., Qi, S. A., Stroulia, E., & Greiner, R. (2021). Learning language and acoustic models for identifying Alzheimer's dementia from speech. *Frontiers in Computer Science, 4*.
30. Silverman, D. H., Small, G. W., Chang, C. Y., Lu, C. S., de Aburto, M. A. K., Chen, W., & Phelps, M. E. (2001). Positron emission tomography in evaluation of dementia: Regional brain metabolism and long-term outcome. *Jama, 286*(17), 2120–2127.
31. Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: A comprehensive review. *Journal of the American Geriatrics Society, 40*(9), 922–935.
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).
33. Weiner, J., Herff, C., & Schultz, T. (2016). Speech-based detection of Alzheimer's disease in conversational german. In *Interspeech* (pp. 1938–1942).
34. World Health Organization. (2021). Dementia Fact Sheet.
35. Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., & Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In *INTERSPEECH* (pp. 2162–2166).