Acoustic-Linguistic Features for Modeling Neurological Task Score in Alzheimer's

Saurav K. Aryal† Howard Prioleau and Legand Burge

EECS, Howard University,
Washington, DC 20059, USA
†*E-mail: saurav.aryal@howard.edu*
https://howard.edu/

The average life expectancy is increasing globally due to advancements in medical technology, preventive health care, and a growing emphasis on gerontological health. Therefore, developing technologies that detect and track aging-associated disease in cognitive function among older adult populations is imperative. In particular, research related to automatic detection and evaluation of Alzheimer's disease (AD) is critical given the disease's prevalence and the cost of current methods. As AD impacts the acoustics of speech and vocabulary, natural language processing and machine learning provide promising techniques for reliably detecting AD. We compare and contrast the performance of ten linear regression models for predicting Mini-Mental Status Exam scores on the ADReSS challenge dataset. We extracted 13000+ handcrafted and learned features that capture linguistic and acoustic phenomena. Using a subset of 54 top features selected by two methods: (1) recursive elimination and (2) correlation scores, we outperform a state-of-the-art baseline for the same task. Upon scoring and evaluating the statistical significance of each of the selected subset of features for each model, we find that, for the given task, handcrafted linguistic features are more significant than acoustic and learned features.

1. Introduction

People are living longer due to advancements in medical technology, preventive health care, and a growing emphasis on gerontological health. The Administration for Community Living estimates that by 2020, 77 million people in the United States will be 60 years of age or older. Hence, developing technologies that detect and track aging-associated disease in cognitive function among older adult populations is imperative.

For decades scientists have examined the association between psychological well-being and cognition. In prior research, gerontologists have identified a significant relationship between mental acuity, loneliness and depression, and social engagement among older adults. Specifically, late-life dementia has been associated with extended periods of loneliness in older adults.[1] Another cognition study,[2] conducted a longitudinal study of adults aged 60 years or older living in North Manhattan, New York, and who were randomly selected from a dementia registry. Their study assessed the association between depressed mood and the onset of dementia. Physicians collected neuropsychological data to assess the degree of decreased cognitive function and determine the risk of dementia. Study results indicated that of the 1,070 participants, 218 (20%) met the criteria for dementia at baseline assessment. Among the 852 participants

that did not have dementia, depressive symptoms were common among those with cognitive impairment. Two years after the baseline data collection, follow-up data were collected on 478 participants who did not have dementia from the baseline collection. A comparison of baseline and follow-up results concluded that of the 478 participants (93%), the depressed mood was associated with dementia and exhibited symptoms of Alzheimer's disease.[2]

Before the turn of the last century, the only way to ascertain if a person has AD was via posthumous autopsy. Currently, as per the National Institute of Health (NIH), medical professionals ask the patient and their caregivers about overall health, medications, diet, medical history, and changes in behavior and personality. They may also administer a psychiatric evaluation to determine confounding causes and conduct tests on memory, problem-solving, attention, counting, language, blood, urine, and other standard medical tests. Finally, performing computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET) supports an AD diagnosis or rules out other plausible causes.[3] While there are other methods, such as accumulation of amyloid plaques and associated genes, these methods may not be entirely accurate[4].[5] Nonetheless, all methods listed are cost-prohibitive or require at least one dedicated medical professional. Consequently, researchers have been studying and modeling non-invasive methods using speech and linguistic features that do not necessitate human intervention to detect and evaluate AD patients. In addition, caregivers experience feelings of depression and being overwhelmed when caring for an older adult lacking social support mechanisms and are predominantly female and overwhelmingly low-income.[1]

Thus, with an aging world population negatively impacted by the symptoms associated with cognitive decline and an overwhelmed caregiving profession, research into technologies to help alleviate these issues is necessary. As AD affects the acoustics of speech[6] and vocabulary,[7] natural language processing and machine learning provide promising techniques for reliably detecting AD. While significant work has been done on detecting AD, this paper will evaluate and score mental status with ten different linear regression models using a combination of handcrafted or learned acoustic-linguistic features. The statistical significance and relevance of each selected feature are also studied.

The rest of the paper covers a review of related works in Section 2. The models, dataset, feature extraction, feature selection, and training-testing protocol are detailed in Section 3. The performance of our models and features are compared to a state-of-the-art baseline linear model in section 4. The final section outlines the conclusion and future work.

2. Related Works

There has been significant research into the symptoms and manifestations of Alzheimer's Disease (AD) in medical literature and AD detection in interdisciplinary research. The review of relevant literature will be divided into two subsections: the first will cover the well-known acoustic-lingual expression of AD in patients, and the second will cover models and techniques currently used for evaluating and detecting AD. Furthermore, the first subsection helps establish the relevance of acoustic and linguistic features for AD progression, whereas the second subsection supports the reasoning behind our methodology.

2.1. *Acoustic and Linguistic Features in AD*

The relation between loss of memory and AD-associated neurodegeneration is well established. Recent research has studied acoustic and verbal aberrations present in patients with AD. In particular, dysarthria/slurring, stuttering, monotony, higher delay, and associated acoustic features with AD.[7] Additionally, linguistic features such as paucity of words or aggramatism are also present with AD.[6,8] In severe cases, sentences uttered may comprise only nouns; articles, auxiliary verbs, and inflectional affixes are absent or replaced in lesser forms. Unsurprisingly, multiple approaches have utilized acoustic and linguistic features for the automatic detection of AD. We will discuss a few of these approaches in the following subsection.

2.2. *Contemporary Models and Techniques for AD Evaluation*

Speech has been used to distinguish between healthy and AD patients.[9] Some researchers have focused on developing dedicated machine learning model architectures[10–12] while others have focused on language models to classify AD.[13] Some research has been focused on extracting acoustic and textual features that capture information indicative of AD, such as the length of segments and the amount of silence.[13] Other researchers have used linguistic and audio features extracted from English speech.[14,15] Prosodic features have been extracted from English speech[16–18] and so have paralinguistic acoustic features.[19] Other approaches have attempted to focus on collecting speech from people performing multiple normative tasks to improve generalizability.[20] However, most of these approaches utilize unbalanced, non-standardized, and proprietary datasets, which hampers their reproducibility and generalizability. We suggest the reader peruse this survey[21] to get a better understanding of these approaches.

In 2020, The ADReSS Challenge[22] defined shared tasks and standardized datasets with predefined metrics. Different approaches for automated recognition of AD based on spontaneous speech and transcripts can be compared with two tasks: AD Classification (AD vs. not-AD) and the neuropsychological score regression. The challenge provided a baseline using standard machine learning models such as Random Forest and k-Nearest Neighbors on classification metrics (accuracy, precision, recall, F-1) and regression Root Mean Square Error (RMSE) scores. More details pertaining to the dataset are discussed in the Methodology section.

Since the release of the dataset, significant work has been done on the classification task,[23–25] the regression task,[26] or both.[27–30] Of the two tasks, a high degree of accuracy 83% to 92.84% has been obtained on the classification task. However, the regression task, being the more challenging of the two, still has room for improvement and is the focus of this paper. Of the approaches reviewed, the lowest RMSE score of 4.56 was acheived on both training and testing sets and utilizes a linear Ridge Regressor model on a set of the 30 best correlating features.[27] We refer to this work as the baseline and state-of-art for the comparison of our model and feature set through the remainder of the paper.

3. Methodology

The models, dataset, feature extraction, feature selection, and training-testing protocol are detailed in the following subsections. All of the tasks performed were performed on a standard personal laptop machine or a Google Collaboratory notebook.[31] No specific accelerators are

required, however, feature extraction, feature selection, and training-testing could be sped up through the utilization of more computing cores.

3.1. *The ADReSS Dataset and Metrics*

To enable comparison with the baseline, the ADReSS Challenge dataset[22] is utilized. This dataset comprises of audio recordings, transcripts from patients performing the Cookie Theft task from the Boston Diagnostic Aphasia exam.[32] Also provided with the dataset are metadata relating to the subject's age, gender and Mini Mental Status Examination (MMSE) score for both non-AD and AD patients. The regression task for this paper is associated with predicting these MMSE score based on the given audio recording and transcripts. Although the MMSE was originally designed to screen for dementia, it is an instrument currently used extensively to assess cognitive status in clinical settings.[33] According to the Alzheimer's Association (2020), an MMSE score of 20–24 corresponds to mild dementia, 13–20 corresponds to moderate dementia, and a score < 12 is severe dementia.

Furthermore, the dataset comes divided into a Train Set (108 patients - 54 non-AD and 54 AD) and a Test Set (48 patients - 24 non-AD and 24 AD). As per the original challenge's guidelines and our baseline, the RMSE is used to determine and compare the performance of our approach. Since the dataset comes with many-to-one mapping of audio file to transcript files, in contrast to previous work, we opted to consider each unique audio-transcript file pair as a distinct observation. While this approach does limit us to shorter audio files with few utterances per file, the number of observations increases to 1447 for training and 569 for testing.

3.2. *Modeling and Train-Validation-Test Protocol*

Although the we were able to increase the sample size by considering audio-transcript file pairs, the number is still smaller than is demanded by most deep learning methods. While work such as[34] has been done on small sample learning, these methods are still a black box. Interpretability is required to evaluate the association between features and the output of the model. While conventional, non-linear machine learning models such as Random Forest and k-Nearest Neighbors were originally the benchmark provided with the dataset,[22] they have been outperformed by the baseline's linear models[27] likely owing to the small sample size. Thus, we also opt for linear modeling. Similar to,[27] we use regression models with in-built regularization or specific optimizations namely Ridge.[35] Additionally, we also employ Lasso,[36] ElasticNet,[37] LassoLars,[38] Bayesian Ridge,[39] Bayesian Automatic Relevance Determination,[40] Orthogonal Matching Pursuit,[41] Huber,[42] TheilSen,[43] and Stochastic Gradient Descent optimization.[44] The models were trained and evaluated using a combination of the BSD-licensed scikit-learn,[45] numpy,[46] seaborn,[47] scipy,[48] and pandas[49] package, and the PSF-licensed matplotlib.[50] The ISF-licensed regressors[51] was used to evaluate the statiscal signficance of each selected feature . Beyond the default, the hyperparameters for each model can be found through the Appendix.

The training and testing protocol utilizes the provided disjoint sets provided with the dataset. Similar to the baseline, each model is trained using Leave One Subject Out (LOSO) Cross Validation on the training set and the RMSE is evaluated on both the training and test set. Of the models, Ridge, Lasso, ElasticNet, LassoLars, and Orthogonal Matching Pursuit's

L1 or L2 regularization parameters were evaluated during this cross-validation. Additionally, a random 80-20 train-validation split of only the training set is used for feature selection.

3.3. *Feature Extraction, Pre-processing, and Feature Selection*

3.3.1. *Feature Extraction*

To learn from both the audio recording and text transcripts, feature extraction is necessary. The dataset provides audio broken up into normalized audio chunks of the subject's sentences/utterances. Text from each participant's transcripts was combined into one large string separated by a new line for linguistic feature extraction. To aid in our feature extraction a combination of software, and python libraries was used. Each of these third-party software, libraries, and their associated licenses are detailed in the Appendix.

We further classify each feature into Audio Features and Linguistic Features. Each of these features may also either be handcrafted or learned. In total, each audio-transcript pair produced just over 13,000 features. To the best of our knowledge, a significant subset of these features are novel applications for the current task of MMSE score prediction.

* **Audio Features** (11,659 Features):

The learned audio features derived from audio recordings include Articulation,[52,53] Phonation,[52,54] and Prosody[52,55] Features. Articulation features are made up of Bark band energies. Phonation features are composed up of pitch perturbation quotient, logarithmic energy, and derivatives of fundamental frequencies account for 28 features. Prosody features, based on energy and duration, include 103 features. The handcrafted audio features include spectral, Mel Frequency Cepstral Coefficients (MFCCs), and Chroma Vector/Deviation features. While all together these features total to 138, we utilized 80 different combinations of frame sizes and overlaps when the average feature are calculated. This was done to find the optimal frame size and overlap which would provide the most significant association with the given task during feature selection.

* **Linguistic Features** (1,693 Features) Linguistic features include, but are not limited to, Word/Sentence Count, Vocab Set, reading scales, and emotion analysis. These features were all extracted from the textual transcript files and totaled up to 1,693 features.

3.3.2. *Pre-processing*

Since audio data was retrieved from a normalized chunks no further pre-processing was required beyond feature extraction. Each participant's transcript was parsed and combined into one large string separated by a new line characters which was used for linguistic feature extraction. Lacking previous background and for convenient modeling, the features were divided by the maximum value. The scaled features were normalized as required by the modeling library before training. No other pre-processing was performed.

3.3.3. *Feature Selection*

While extracting over 13,000 features provides us with a significant amount of data. Linear models, even with strong regularization, tend to get over-parameterized at this scale and

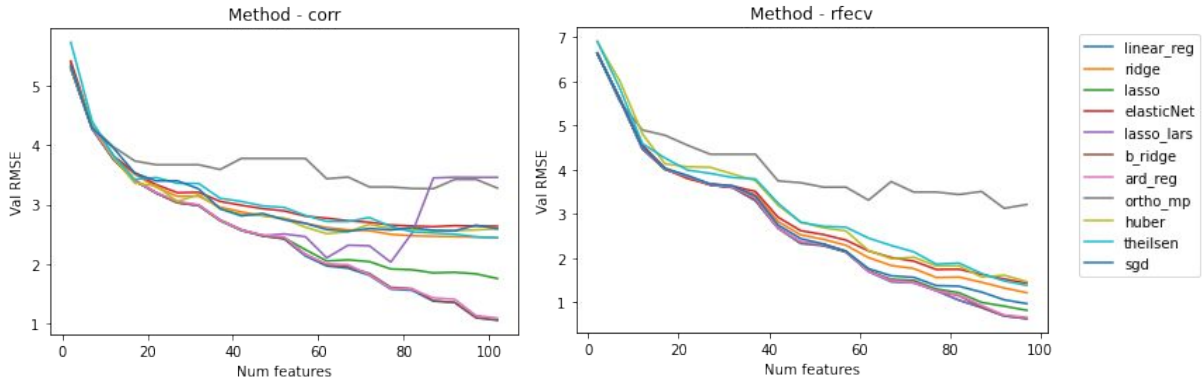

Fig. 1. Validation RMSE vs Num Features using Correlation and Recursive Elimination

require specific adaptation. Thus, we opt to select a subset of 100 due to limitations in available computing power and time. We utilized two methods from[45] for selecting the best features for this problem: (1) Recursive Feature Elimination using a standard Linear Regression estimator and (2) Correlation Scores. For the first method, the best set of features which decreased the RMSE on a standard linear regression model trained on 80% of the training set and minimized RMSE on the 20% validation set was used. We could not get to 100 features since the method only lets us select a minimum number of features required and outputted a set of features > 100. For the second method, we simply selected a the top 100 most correlated features with the output. In order, to further simply the model we trained and validated the models on features from the top 2 features until the all top 100 features selected by the algorithms. Plots of validation RMSE for each of the methods can be seen in Figure 1. As expected, the error does incrementally decrease with the addition of each feature. However, we are better suited taking a cut off around at a few feature after the steep decrease in RMSE. We chose to set this limit at 54 features which is half the number of subjects in the training set. Lacking precedence, we used P-values < 0.05 and coefficient > 0.01 were considered significant. Given page limitations, model summaries, source code, and additional plots are provided via the Appendix. In the following section, we will cover the results of our modeling experiments and perform comparisons with the baseline.

4. Results

All of the models using features selected by both RFECV and Correlation outperformed the baseline model on the training set. Of these models, the standard linear regression model performed the best with an RMSE improvement of 2.37 compared to the baseline of 4.56. The RMSE plot for each model can be seen in Figure 2.

However, for the test set, not all models outperformed the baseline. Interestingly, none of the models which used features selected by recursive elimination outperformed the baseline whereas five models using correlation features outperformed the baseline despite the two methods having an overlap of 17 features selected out of the total 54. Of these models that outperformed the baseline, the stochastic gradient descent optimized model performed the

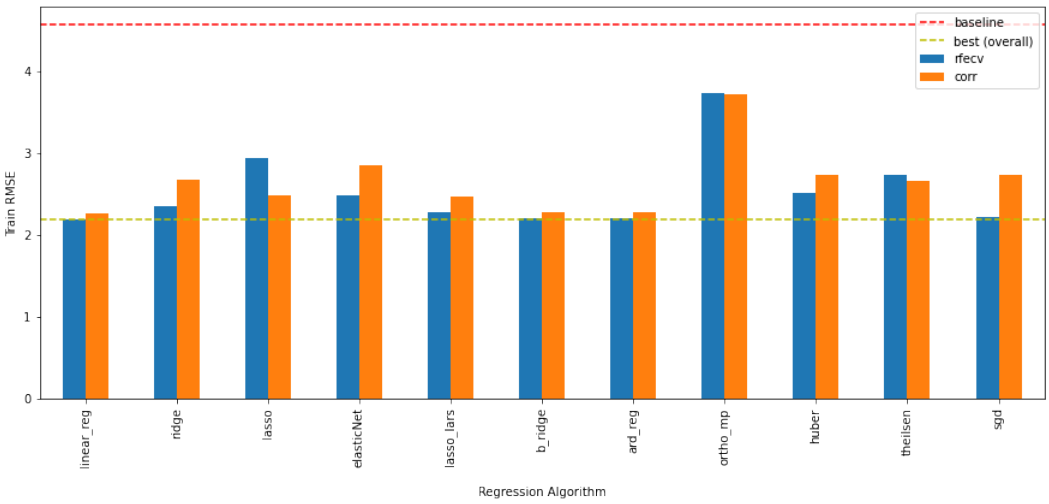

Fig. 2. Train RMSE for each model and each feature selection method

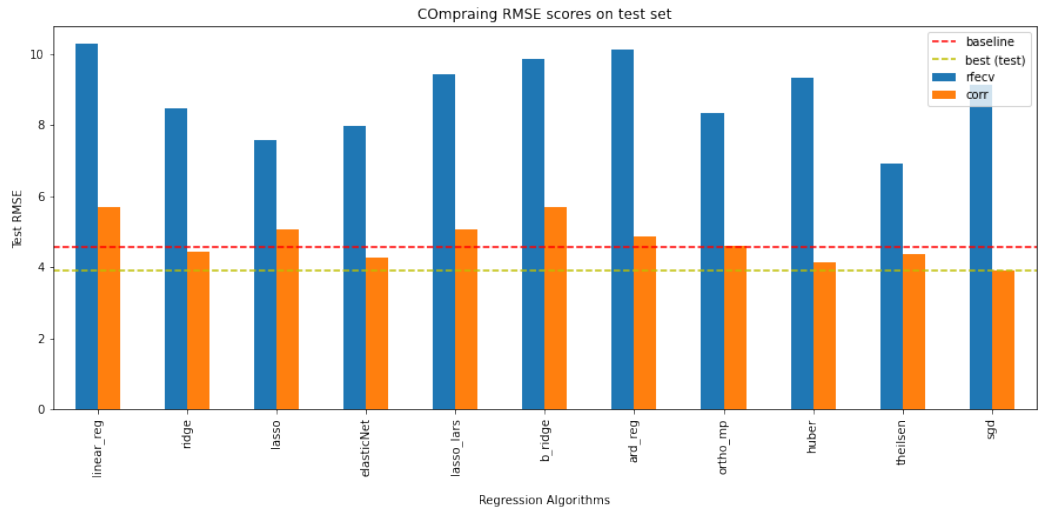

Fig. 3. Test RMSE for each model and each feature selection method

best with an RMSE improvement of 0.66 compared to the baseline RMSE of 4.56. The plot of RMSE can be seen in Figure 3.

Upon a closer look into the the box in Figure 4 and histogram plots in Figure 5 of the residuals of each of the models that outperformed the baseline, we notice that stochastic gradient descent optimization has the most reliable performance. However, the range of prediction is currently too large and unreliable in all of these models for real world application.

Moreover, of the 54 features selected by the methods, it was noticed that all were handcrafted linguistic features related to word usage, readability, and character frequencies. This observation is inline with the observations of both the baseline and speech pathological research[6,8] that linguistic features are better predictors for this task in comparison to acoustic features and is supported. Details results of feature selection can be found via the Appendix

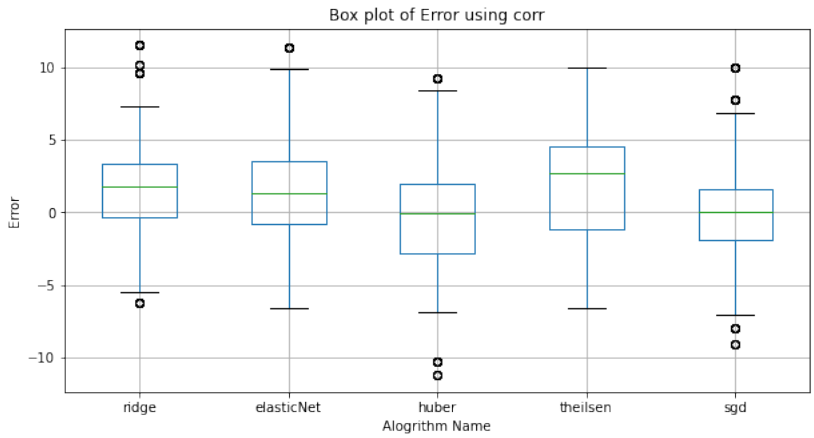

Fig. 4. Boxplot of Residuals on the Test set

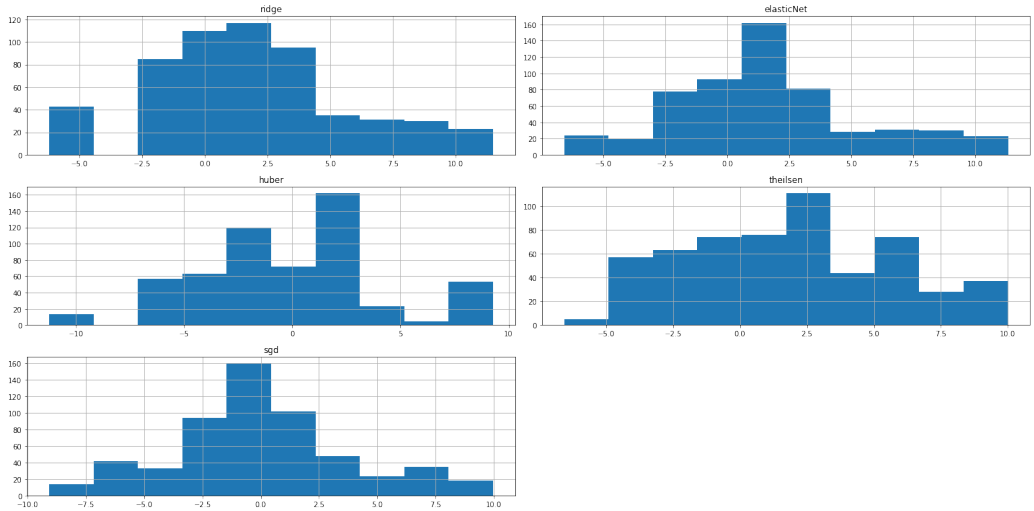

Fig. 5. Histogram of Residuals on the Test set

5. Limitations and Future Work

The major limitation of this work stems from data source. Since the dataset consists of audio recordings of the participants performing a specific task, it is unlikely these findings may be generalizable to recordings that are not obtained from the same task or for non-native English speakers. Furthermore, the standardization based on this task might also explain the proclivity of models to find significance of linguistic features over acoustic features for the prediction of MMSE scores. It is possible that other modes of data capture may be better suited to a general approach for evaluating AD patients.[20]

Although the current dataset is remarkable, the sample size limits researchers from fully realizing and utilizing the most recent advancement in machine learning. While approaches such as early stopping and dropouts could be utilizes, one must question the external validity of such approaches within such a small sample size. Perhaps research into small sample size

algorithms[34] could be applied; however the issues related to interpretability still persists.

Contemporary research has shown the continued need to advance further the study of aging-associated disease effects on cognitive impairment in older adults.[56] Researchers studied older adults who were already enrolled in research projects investigating the onset of Alzheimer's Disease (AD) on cognition under the assumption that the Functional Activities Questionnaire (FAQ) using the Instrumental Activities of Daily Living (IADL) scale to detect and track diminishing capability in managing and remembering daily household tasks and personal responsibilities. Difficulties in managing IADL identified in the FAQ proved helpful in detecting and tracking changes in cognition in healthy older adults at risk for Alzheimer's Disease.[57] Furthermore, social determinants of health such as transportation, education, diet, and other daily factors negatively impact a person's health outlook. Black and Brown persons in the United States are adversely affected by schooling, diet, and disease symptoms associated with hypertension and diabetes that might cause cognitive decline.[58] To further improve the reliability of the models social determinants, facial features, depression, and other correlates can be considered in conjunction with an in-home monitoring and audio-video capture device.

While we do believe that this paper sufficiently advance the state-of-the-art for this task, explores the largest feature space to date, and guides us towards automating the diagnosis of AD and modeling of cognitive status in the elderly, we must note that with automation we should not intend to replace trained medical professionals. We firmly believe that any technology stemming from research should be used as a tool to guide, assist, and ease medical professionals and caregivers to provide the best care possible.

6. Conclusion

While we were able to outperform the baseline with 5 different models, the performance of these models are still not fully suited for real world application. More research needs to be done to find models that work on low resource problems such as neurological evaluation of AD patients using audio and textual features.

7. Acknowledgement

This project was supported (in part) by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number 2U54MD007597, and the Office of Data Science Strategy of the National Institutes of Health under OTA OT2 OD32581-01, and a 2021 Amazon Research Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding organizations.

8. Appendix

All supplemental materials can be found in the link below: https://bit.ly/3Skbaij

References

1. R. S. Wilson, K. R. Krueger, S. E. Arnold, J. A. Schneider, J. F. Kelly, L. L. Barnes, Y. Tang and D. A. Bennett, Loneliness and Risk of Alzheimer Disease, *Archives of General Psychiatry* **64**, 234 (February 2007).

2. D. P. Devanand, M. Sano, M.-X. Tang, S. Taylor, B. J. Gurland, D. Wilder, Y. Stern and R. Mayeux, Depressed mood and the incidence of alzheimer's disease in the elderly living in the community, *Archives of general psychiatry* **53**, 175 (1996).
3. How Is Alzheimer's Disease Diagnosed? | National Institute on Aging.
4. K. R. Thomas, K. J. Bangen, A. J. Weigand, E. C. Edmonds, C. G. Wong, S. Cooper, L. Delano-Wood, M. W. Bondi and f. t. A. D. N. Initiative, Objective subtle cognitive difficulties predict future amyloid accumulation and neurodegeneration, *Neurology* **94**, e397 (January 2020), Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Article.
5. M. Giri, M. Zhang and Y. Lü, Genes associated with Alzheimer's disease: an overview and current status, *Clinical Interventions in Aging* **11**, 665 (May 2016).
6. F. Rudzicz, G. Hirst, P. van Lieshout, G. Penn, F. Shein, A. Namasivayam and T. Wolff, Torgo database of dysarthric articulation (2012).
7. I. Ferrer, A. Aymami, A. Rovira and J. M. Grau Veciana, Growth of abnormal neurites in atypical Alzheimer's disease, *Acta Neuropathologica* **59**, 167 (September 1983).
8. J. O. d. Lira, T. S. C. Minett, P. H. F. Bertolucci and K. Z. Ortiz, Analysis of word number and content in discourse of patients with mild to moderate alzheimer's disease, *Dementia & neuropsychologia* **8**, 260 (2014).
9. 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* **150**, p. 113213 (July 2020).
10. J. Chen, J. Zhu and J. Ye, An attention-based hybrid network for automatic detection of alzheimer's disease from narrative speech., in *INTERSPEECH*, (Not Available, 2019).
11. Y.-W. Chien, S.-Y. Hong, W.-T. Cheah, L.-C. Fu and Y.-L. Chang, An Assessment System for Alzheimer's Disease Based on Speech Using a Novel Feature Sequence Design and Recurrent Neural Network, in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (Not Available, 2018). ISSN: 2577-1655.
12. L. Liu, S. Zhao, H. Chen and A. Wang, A new machine learning method for identifying Alzheimer's disease, *Simulation Modelling Practice and Theory* **99**, p. 102023 (February 2020).
13. Z. Guo, Z. Ling and Y. Li, Detecting Alzheimer's Disease from Continuous Speech Using Language Models, *Journal of Alzheimer's Disease* **70**, 1163 (January 2019), Publisher: IOS Press.
14. K. C. Fraser, J. A. Meltzer and F. Rudzicz, Linguistic Features Identify Alzheimer's Disease in Narrative Speech, *Journal of Alzheimer's Disease* **49**, 407 (January 2016), Publisher: IOS Press.
15. G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán and I. Hoffmann, Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features, *Computer Speech & Language* **53**, 181 (January 2019).
16. R. Nagumo, Y. Zhang, Y. Ogawa, M. Hosokawa, K. Abe, T. Ukeda, S. Sumi, S. Kurita, S. Nakakubo, S. Lee, T. Doi and H. Shimada, Automatic Detection of Cognitive Impairments through Acoustic Analysis of Speech, *Current Alzheimer Research* **17**, 60 (January 2020).
17. Y. Qiao, X.-Y. Xie, G.-Z. Lin, Y. Zou, S.-D. Chen, R.-J. Ren and G. Wang, Computer-Assisted Speech Analysis in Mild Cognitive Impairment and Alzheimer's Disease: A Pilot Study from Shanghai, China, *Journal of Alzheimer's Disease* **75**, 211 (January 2020), Publisher: IOS Press.
18. R. Ossewaarde, R. Jonkers, F. Jalvingh and R. Bastiaanse, Classification of Spontaneous Speech of Individuals with Dementia Based on Automatic Prosody Analysis Using Support Vector Machines (SVM), in *The Thirty-Second International Flairs Conference*, (Not Available, 2019).
19. 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* **14**, 272 (February 2020), Conference Name: IEEE Journal of Selected Topics in Signal Processing.
20. A. Balagopalan, J. Novikova, F. Rudzicz and M. Ghassemi, *The Effect of Heterogeneous Data for*

Alzheimer's Disease Detection from Speech, Tech. Rep. arXiv:1811.12254, arXiv (November 2018), arXiv:1811.12254 [cs, eess, stat] type: article.
21. S. de la Fuente Garcia, C. W. Ritchie and S. Luz, Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review, *Journal of Alzheimer's Disease* **78**, 1547 (2020).
22. 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).
23. E. Edwards, C. Dognin, B. Bollepalli, M. K. Singh and V. Analytics, Multiscale system for alzheimer's dementia recognition through spontaneous speech., in *INTERSPEECH*, (Not Available, 2020).
24. J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye and K. Church, Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease., in *INTERSPEECH*, (Not Available, 2020).
25. A. Pompili, T. Rolland and A. Abad, The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge (May 2020).
26. S. Farzana and N. Parde, Exploring mmse score prediction using verbal and non-verbal cues., in *INTERSPEECH*, (Not Available, 2020).
27. A. Balagopalan, B. Eyre, F. Rudzicz and J. Novikova, To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection, *arXiv preprint arXiv:2008.01551* (2020).
28. M. S. S. Syed, Z. S. Syed, M. Lech and E. Pirogova, Automated screening for alzheimer's dementia through spontaneous speech., in *INTERSPEECH*, (Not Available, 2020).
29. T. Searle, Z. Ibrahim and R. Dobson, Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech, *arXiv preprint arXiv:2006.07358* (2020).
30. G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah and A. Karpov, Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition, *arXiv preprint arXiv:2009.03432* (2020).
31. E. Bisong, Google Colaboratory, in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, ed. E. Bisong (Apress, Berkeley, CA, 2019) pp. 59–64.
32. B. MacWhinney, The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database (2000).
33. R. Y. Wood, K. K. Giuliano, C. U. Bignell and W. W. Pritham, Assessing cognitive ability in research: use of mmse with minority populations and elderly adults with low education levels., *Journal of Gerontological Nursing* **32**, 45 (2006).
34. R. Keshari, S. Ghosh, S. Chhabra, M. Vatsa and R. Singh, Unravelling small sample size problems in the deep learning world, in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, (Not Available, 2020).
35. D. W. Marquardt and R. D. Snee, Ridge regression in practice, *The American Statistician* **29**, 3 (1975).
36. J. Ranstam and J. Cook, Lasso regression, *Journal of British Surgery* **105**, 1348 (2018).
37. H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301 (2005).
38. R. J. Tibshirani and J. Taylor, The solution path of the generalized lasso, *The annals of statistics* **39**, 1335 (2011).
39. C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning* (Springer, 2006).
40. D. Wipf and S. Nagarajan, A new view of automatic relevance determination, *Advances in neural information processing systems* **20** (2007).
41. T. Blumensath and M. E. Davies, On the difference between orthogonal matching pursuit and

orthogonal least squares (2007).
42. A. Owen, A robust hybrid of lasso and ridge regression (technical report) (2006).
43. X. Wang, X. Dang, H. Peng and H. Zhang, The theil-sen estimators in multiple linear regression models, *Manuscript available at: http://home. olemiss. edu/˜ xdang/papers/MTSE. pdf* (2009).
44. L. Bottou, Stochastic gradient descent tricks, in *Neural networks: Tricks of the trade*, (Springer, 2012) pp. 421–436.
45. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011).
46. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, Array programming with NumPy, *Nature* **585**, 357 (September 2020).
47. M. L. Waskom, seaborn: statistical data visualization, *Journal of Open Source Software* **6**, p. 3021 (2021).
48. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods* **17**, 261 (2020).
49. T. pandas development team, pandas-dev/pandas: Pandas (February 2020).
50. J. D. Hunter, Matplotlib: A 2d graphics environment, *Computing in Science & Engineering* **9**, 90 (2007).
51. N. Haas, regressors: Easy utilities for fitting various regressors, extracting stats, and making relevant plots.
52. J. C. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet and E. Nöth, Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease, *Journal of communication disorders* **76**, 21 (2018).
53. J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei *et al.*, Neurospeech: An open-source software for parkinson's speech analysis, *Digital Signal Processing* **77**, 207 (2018).
54. T. Arias-Vergara, J. C. Vásquez-Correa and J. R. Orozco-Arroyave, Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech, *Cognitive Computation* **9**, 731 (2017).
55. N. Dehak, P. Dumouchel and P. Kenny, Modeling prosodic features with joint factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing* **15**, 2095 (2007).
56. Z. Li, Z. Zhang, Y. Ren, Y. Wang, J. Fang, H. Yue, S. Ma and F. Guan, Aging and age-related diseases: from mechanisms to therapeutic strategies, *Biogerontology* **22**, 165 (April 2021).
57. G. A. Marshall, A. S. Zoller, N. Lorius, R. E. Amariglio, J. J. Locascio, K. A. Johnson, R. A. Sperling, D. M. Rentz and for the Alzheimer's Disease Neuroimaging Initiative, Functional Activities Questionnaire Items that Best Discriminate and Predict Progression from Clinically Normal to Mild Cognitive Impairment, *Current Alzheimer Research* **12**, 493 (June 2015).
58. G. Landsberg, Therapeutic options for cognitive decline in senior pets, *Journal of the American Animal Hospital Association* **42**, 407 (2006).