2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Using Verb Fluency, Natural Language Processing, and Machine Learning to Detect Alzheimer's Disease

Aradhana Soni[1], Benjamin Amrhein[1], Matthew Baucum[1], Eun Jin Paek[2], Anahita Khojandi[1]

Abstract— **Alzheimer's disease (AD) causes significant impairments in memory and other cognitive domains. As there is no cure to the disease yet, early detection and delay of disease progression are critical for management of AD. Verbal fluency is one of the most common and sensitive neuropsychological methods used for detection and evaluation of the cognitive declines in AD, in which a subject is required to name as many items as possible in 30 or 60 seconds that belong to a certain category. In this study, we develop an approach to detect AD using a verb fluency (VF) task, a specific subset of verbal fluency analyzing the subjects' listing of verbs in a given time period. We use machine learning techniques including random forest (RF), neural network (NN), recurrent NN (RNN), and natural language processing (NLP) to detect the risk of AD. The results show that the developed models can stratify subjects into the corresponding AD and control groups with up to 76% accuracy using RF, but at a cost of having to preprocess the data. This accuracy is slightly lower, but not significantly, at 67% using RNN and NLP, which involves almost no manual preprocessing of the data. This study opens up a powerful approach of using simple VF tasks for early detection of AD.**

Key Words—**Alzheimer's disease, early detection, random forest, neural networks, natural language processing**

I. INTRODUCTION

Alzheimer's disease (AD) is the leading cause of dementia, accounting for 60-80 percent of cases [1]. Dementia generally refers to a patient's decline in memory and cognitive skills such as their ability to reason, think, or speak clearly. AD is a degenerative brain disease that originates from damage to brain cells. While no cure for AD currently exists, earlier detection of the disease means earlier intervention and more effective care. Despite the growing number of cases of AD, approximately only a quarter of the patients are typically diagnosed [2]. Worse yet, the mortality rate of AD in the United States has significantly increased between 2000 and 2018 from 17.6 to 37.3 deaths per 100,000 population [3].

A large research body is dedicated to studying the utilization of language tasks in improving early detection of AD [4]. This research has shown promise as AD results in cognitive impairment and typically has negative implications on how patients produce or use language. In general, past studies have covered recording a patient's speech over a period of time and analyzing the number and types of words they produce to detect AD [5]. Such an approach to AD detection is promising because there is generally no need for expensive equipment or invasive procedures, and the data collection and analysis can be done even remotely. However, existing works on detecting AD from recorded speech data generally use time-intensive tasks, such as open-ended interviews with clinicians [6].

We leverage a verb fluency (VF) task data analysis to detect AD that simply relies on the way patients list verbs in bursts of 30 seconds [7]. Although verbal fluency, e.g., semantic fluency and phonemic fluency, has been commonly used to detect AD, analyzing the listing of verbs is much less explored. This task of listing verbs has the potential to simplify the evaluation process and can be more readily transferable and generalizable to a large array of languages.

Therefore, in this study, we aim to leverage machine learning (ML) and natural language processing (NLP) along with VF for early AD detection. ML is a branch of artificial intelligence that allows for eliciting patterns from the data. It can draw associations between a set of input variables (e.g., the choice of verbs, the pattern with which they are produced, etc.) and output (response) variables (e.g., at risk for AD or not). NLP is a field at the intersection of artificial intelligence and linguistics that concerns with the interactions between human (natural) language and computers. Both ML and NLP, either separately or jointly, have been used in health care applications to much success, e.g., to detect or predict various outcomes or risks for patients using electronic medical records [8].

In this study, we develop an approach to detect AD using the data from a 30-second VF task. First, we develop ML models that detect AD using psycholinguistic features of the input verbs, extracted by experts from the VF task data. We specifically develop random forest (RF) and neural networks (NN) models. Next, we leverage NLP and ML jointly to develop an end-to-end ML pipeline. That is, we use NLP on the concatenated text string of verbs from subjects to elicit information. We then use this elicited information along with the (raw) sequence of verbs produced in a recurrent neural network (RNN) model to detect AD.

II. LITERATURE REVIEW

A number of studies have already investigated the efficacy of subjects' recall and choice of words or types of speech to detect AD. For instance, Jarrold et al. [4] collected recordings of patients' and controls' speech during simple interviews that asked open-ended questions. They applied three ML techniques, logistic regression, NNs, and decision trees to find that AD patients use more pronouns than nouns, with 88% accuracy.

[1] Department of Industrial & Systems Engineering, University Of Tennessee, Knoxville, TN, U.S.A (asoni5@vols.utk.edu, bamrhein@vols.utk.edu, mbaucum1@vols.utk.edu, khojandi@utk.edu)

[2] Eun Jin Paek is with The University Of Tennessee Health Science Center Knoxville, TN, U.S.A (epaek@uthsc.edu)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

The study by Shibata et al. [5] focused on the number and frequency of word categories used by patients and controls during recorded conversations in which subjects were asked 11 different questions. The authors found a statistically significant difference between AD patients and healthy controls regarding the number of impersonal pronouns and verbs used between the two groups. These studies lend an imperative basis to our continued work in this field as they not only underscore the utility of analyzing patients' selection of words to detect AD, but they also illuminate a path forward to explore work in VF with NLP.

NNs are especially effective in text classification models and are employed by Apple's Siri, Google Assistant, and Amazon Alexa [9]. RNNs are a special type of NNs and are commonly used for automated time series classification, especially when used along with NLP. For instance, they have been used for tasks such as predicting whether movie reviews are positive or negative [10]. To the best of our knowledge, there has been little work on the application of RNNs with NLPs in VF data analysis and AD diagnoses. One study utilized RNN to analyze AD progression with and without missing data [11], while another focused on patients with mild cognitive impairment that were given questionnaires [12], but neither concentrated on data centered around patients' listings of verbs.

III. Methodology

A. Data Collection and Preprocessing

The subject cohort includes a total of 20 AD patients (mean age = 77.85 years) and 25 age-matched controls (mean age = 72.68 years). Each subject is asked to say as many verbs as possible in a 30-second block. The responses are recorded verbatim. The study protocol was reviewed and approved by The University of Tennessee Health Science Center.

The data are analyzed by subject matter experts to extract psycholinguistic properties. The analysis is performed to elicit properties pertaining to VF responses of individuals with amnestic AD and cognitively healthy older adults. Specifically, The English Lexicon project, a multi-university effort to provide a standardized behavioral and descriptive data set for 40,481 words and 40,481 non-words [13], is used for the psycholinguistic analysis. To extract psycholinguistic properties, the root forms of the verbs are used. The properties extracted include [7]: Total number of correctly produced words, length of the word, the number of phonological neighbors that a word has, the number of orthographic neighbors that a word has, how pleasant a word is, the extent to which the word denotes something that is weak or strong, number of phonemes in the pronunciation, word frequency, and the age of acquisition of the word.

B. Models

Two types of ML models were developed in this study. The first type relied on features extracted from the psycholinguistic properties. Specifically, we calculated the average, standard deviation, and range of each of the psycholinguistic properties reported for any given subject. This resulted in 60 initial features. We then used these features to develop two ML models, namely RF and NN. RF is an ensemble classifier that uses a large number of decision trees, each fitted on a randomly selected subset of the data, for classification [14]. RF is generally highly robust against overfitting. For the RF model, based on preliminary results using out-of-bag (OOB) error, 100 trees were included in the model.

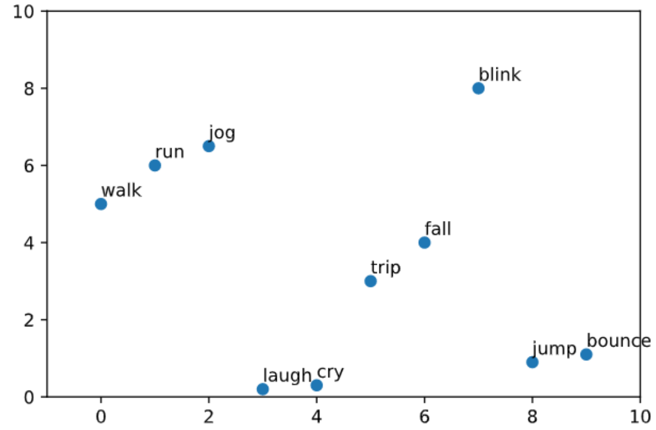

Fig. 1. Example 2D word embedding space, where similar words are closer together

In addition, we used NNs, another non-linear learning model for classification. NNs transfer the information from an input layer into a hidden layer and finally outputs the results [15]. For the NN model, one hidden layer with 16 hidden nodes was used. The activation function was set to rectified linear unit (ReLU). Also, the learning rate was set to 0.001. Adam optimizer was used for model training. Lastly, features were normalized before feeding them into the model.

The second type of ML models did not rely on features extracted from the psycholinguistic properties. Specifically, we developed an RNN directly using the recorded verbatim. This involved using the concatenated string of verbal responses for any given subject, plus the corresponding *word embeddings* obtained from NLP. In particular, we used word embeddings to convert the words into vectors, allowing the RNN to form a relationship between different verbs produced by subjects. Fig. 1 provides an example of how this relationship is established in a two-dimensional word embeddings. As seen in the figure, the word 'walk', for example, is closer to the word 'jog' than the word 'laugh.' Therefore, words that are closer in meaning (or are related in some way) have more similar vector representations. The RNN includes one hidden layer with 50 hidden nodes. The activation function was set to 'sigmoid'. Adam optimizer was used for model training. The learning rate was again set to 0.001. All models are developed in Python. For ML models, we use Keras [16] with the TensorFlow backend [17]. In addition, we use pre-trained, 300-dimension word embeddings from the spaCy package, which are trained on a corpus of web page data [18].

TABLE I

TOP 15 MOST IMPORTANT FEATURES FOR RF AND NN MODELS IN THE ORDER OF IMPORTANCE

Feature	**Description** [13]
Nphon Sd	Standard deviation of number of phonemes in the pronunciation.
Emotional Valence sd	Standard deviation of how pleasant a word is.
Ortho Range	Range of the number of orthographic neighbors a word has.
Score Average	Average of the indicator of when a person said a new verb.
AoArate Range	Range of age of acquisition obtained through adults' ratings.
AoArate sd	Standard deviation of age of acquisition obtained through adults' ratings.
Phono_N Average	Average of the number of phonological neighbors that a word has.
Length sd	Standard deviation of the length of the word.
Emotional Dominance Range	Range of the extent to which the word denotes something that is weak or strong.
Ortho Average	Average of the number of orthographic neighbors a word has.
Phono_N Range	Range of the number of phonological neighbors that a word has.
Ortho sd	Standard deviation of the number of orthographic neighbors a word has.
Emotional Valence Range	Range of how pleasant a word is.
Length Average	Average of the length of the word.
Phono_H Average	Average of the number of phonological neighbors that a word has including homophones.

C. Feature Selection for RF and NN

RF allows for ranking feature importance per the total decrease in the Gini measure of node impurities [19]. We use this feature ranking to perform feature pruning. Specifically, the 60 initial features are first ranked based on their importance using RF. The 15 most important features are then selected to be used in both RF and NN models.

D. Input Data Tuning for the RNN and NLP Model

Recall that concatenated string of verbal responses for any given subject was used in the RNN and NLP model. To improve model performance, various text string combinations were explored. This included using concatenated strings with and without stumbling (such as "um" and "uh"), and with and without repeated verbs, if they occurred.

E. Model Evaluation and Metrics

For all models, we employ five-fold cross validation. In each fold, training is done using balanced sets, i.e., equal numbers of AD patients and healthy controls. This helped to avoid favoring the more representative group. We consequently provide mean and standard deviations across the five folds.

Evaluation metrics include accuracy, F1 score, and area under the receiver operating characteristic curve (AUC). Accuracy is the ratio of correct predictions over total predictions. F1 is the harmonic mean of precision and recall, where precision is the proportion of positive predictions that are correct and recall is the proportion of positive predictions that are correctly classified. AUC is the value that reflects the overall ranking performance of a classifier [20].

IV. RESULTS

Fig. 2 displays the importances of the top 15 features that are included in the RF and NN models. Table I presents the descriptions of these features. As seen in the figure and table, the features are generally drawn from psycholinguistic properties relating to age of acquisition, number of phonemes, how pleasant the word is, phonological neighbors that a word has, among others. [13]. In addition, per the preliminary results, the text strings *without* stumbling and *with* repeated verbs resulted in best performance. We used this approach for the rest of the study.

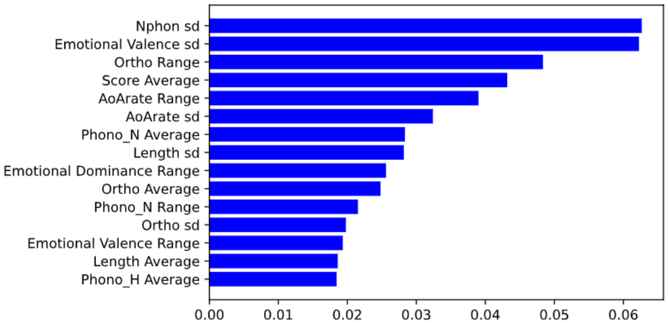

Fig. 2. Feature Importances. Table I presents the descriptions of these features.

Table II presents the averages and standard deviations of the evaluation metrics for RF, NN, and RNN models. As seen in the table, RF slightly outperforms NN and RNN models. This model is able to detect AD participants with an accuracy of 76%. Note that the RF model relies on features extracted from psycholinguistic properties that require considerable preprocessing of the data by subject matter experts. However, even with minimal preprocessing of data, RNN model is able to correctly detect AD with an accuracy of 67%.

TABLE II

AVERAGES (AND STANDARD DEVIATIONS) OF METRICS OF THE THREE ML MODELS

	RF	NN	RNN
Accuracy	76.00% (11.00%)	68.89% (8.00%)	66.67% (9.94%)
F1 score	71.44% (9.94%)	66.10% (7.89%)	71.88% (7.76%)
AUC	75.00% (10.37%)	69.00% (7.84%)	60.00% (13.04%)

Further, we performed paired t-test to compare the results of the three models. Table III lists the p-values of these tests. As seen in the table, the differences between the RF model and the other models is not statistically significant. This concludes that the results from the RF model are not significantly better than the other two models.

TABLE III
t-TEST COMPARISON OF THREE MODELS

	RF vs NN p-value	RF vs RNN p-value
Accuracy	0.0705	0.3739
F1 score	0.2083	0.9461
AUC	0.0801	0.1244

V. CONCLUSIONS

Our results demonstrate that we can correctly detect AD with above-average chance accuracy using NLP, even when using an RNN that requires almost no preprocessing of subjects' VF data. Our accuracy scores fall within the reported accuracy ranges of several clinical AD detection methods, such as EEG and brain scans, that are considerably more costly and time-consuming than VF tasks [21]. Our results thus show promise for detecting AD using data-driven methods without resorting to cost prohibitive, invasive or time-consuming clinical procedures.

As indicated by our results, RF performs as the slightly better method for detecting AD when compared with NN and RNN. However, the differences are not significant. It is worth noting that the RF requires considerable data preprocessing. While the RF requires analysis and computation of psycholinguistic properties, the RNN simply requires the concatenation of the subjects' verbs. The latter methodology provides a much more efficient, time and cost saving means to detect AD with 67% accuracy, and can easily be conducted remotely.

Given these benefits and the insights derived here regarding its potential effectiveness, further exploration into using an RNN with an NLP after collecting subjects' verb listings stands out as a worthy venture. More specifically, future work may include further refining and tuning the RNN and NLP while studying the patient-specific covariates including age and comorbidities more comprehensively. It is also worth investigating whether analyzing different categories of verbal fluency tasks (e.g., semantic, phonemic, and verb fluency) simultaneously adds value to the detection of AD, given the distinctive psycholinguistic processes demanded upon each task and each task's sensitivity to different aspects of cognitive declines in AD. Lastly, we acknowledge that one of the limitations of this study is the small sample size. Hence, further studies using larger data sets are needed to reproduce the current findings and build upon them.

REFERENCES

[1] A. Association, "Dementia vs. alzheimer's disease: What is the difference?" 2021. [Online]. Available: https://www.alz.org/alzheimers-dementia/difference-between-dementia-and-alzheimer-s

[2] A. N. Today, "Alzheimer's disease statistics." [Online]. Available: https://alzheimersnewstoday.com/alzheimers-disease-statistics

[3] J. Elflein, "Death rate due to alzheimer's disease in the u.s. 2000-2019," *Alzheimer's Association*, 2021.

[4] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 27–37. [Online]. Available: https://www.aclweb.org/anthology/W14-3204

[5] D. Shibata, S. Wakamiya, A. Kinoshita, and E. Aramaki, "Detecting Japanese patients with Alzheimer's disease based on word category frequencies," in *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 78–85. [Online]. Available: https://www.aclweb.org/anthology/W16-4211

[6] K. Palmer, L. Bäckman, B. Winblad, and L. Fratiglioni, "Detection of alzheimer's disease and dementia in the preclinical phase: population based cohort study," *Bmj*, vol. 326, no. 7383, p. 245, 2003.

[7] E. J. Paek and L. L. Murray, "Quantitative and qualitative analysis of verb fluency performance in individuals with probable alzheimer's disease and healthy older adults," *American journal of speech-language pathology*, vol. 30, no. 1S, pp. 481–490, 2021.

[8] Q.-Y. Zhong, E. W. Karlson, B. Gelaye, S. Finan, P. Avillach, J. W. Smoller, T. Cai, and M. A. Williams, "Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing," *BMC medical informatics and decision making*, vol. 18, no. 1, pp. 1–11, 2018.

[9] V. Këpuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018, pp. 99–103.

[10] T. Gadekallu, A. Soni, D. Sarkar, and L. Kuruva, "Application of sentiment analysis in movie reviews," in *Sentiment Analysis and Knowledge Discovery in Contemporary Business*. IGI Global, 2019, pp. 77–90.

[11] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, B. T. Yeo, A. D. N. Initiative *et al.*, "Predicting alzheimer's disease progression using deep recurrent neural networks," *NeuroImage*, vol. 222, p. 117203, 2020.

[12] H. Li and Y. Fan, "Early prediction of alzheimer's disease dementia based on baseline hippocampal mri and 1-year follow-up cognitive measures using deep recurrent neural networks," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 368–371.

[13] D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The english lexicon project," *Behavior research methods*, vol. 39, no. 3, pp. 445–459, 2007.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] B. Müller, *Neural Networks An Introduction*, ser. Physics of Neural Networks. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995.

[16] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[17] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[18] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.

[20] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.

[21] J. E. Gaugler, R. L. Kane, J. A. Johnston, and K. Sarsour, "Sensitivity and specificity of diagnostic accuracy in alzheimer's disease: a synthesis of existing evidence," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 28, no. 4, pp. 337–347, 2013.