Spring 5-15-2022

Dementia Classification through Textual Analysis with Machine Learning Algorithms

Joseph Hurowitz
University of Lynchburg, HUROWITZ_JB@lynchburg.edu

Dementia Classification through Textual Analysis with Machine Learning Algorithms

Joseph Hurowitz

Senior Honors Project

Submitted in partial fulfillment of the graduation requirements of the Westover Honors College

Westover Honors College

May 2022

____
M. Zakaria Kurdi, PhD

____
Randy Ribler, PhD

____
Laura Kicklighter, PhD

Abstract

The goal of this work is to build a classifier that can identify whether a patient is suffering from Alzheimer's Disease of the Dementia Type (AD). A corpus of 2751 texts was used from the DementiaBank database, where each conversation is transcribed and marked using the CHAT format. Each text was analyzed by frequency of disfluencies, use of aphasic language, and lexical features. All parsed data was used to train a Random Forest, Naïve Bayes, and Support Vector Machine algorithm. These classification algorithms will be tested on the combination of all features, as well as each set of features individually.

Keywords: Dementia, Alzheimer's, Machine Learning, Random Forest, Naive Bayes Gaussian, Support Vector Machine

1. Introduction

Dementia is a term used to describe different brain disorders that affect memory, thinking, behavior, and emotion [11]. The most usual form of dementia is known as Alzheimer's disease, which accounts for 60-80% of all cases, and is usually diagnosed in elderly patients over the age of sixty-five [11]. Alzheimer's is a progressive disease, where dementia symptoms gradually worsen over several years [10]. It is characterized by the loss of cognitive ability such as memory, reasoning, language, and behavior [7]. Early-stage Alzheimer's causes mild cognitive impairment, which affects memory. However, as the disease progresses, brain function is severely impacted, which can lead to an inability to converse, changes in behavior, and a loss of motor skills.

According to Alzheimer's Disease International, there are currently over 55 million people living with dementia; however, this number is expected to double every 20 years [11]. To coincide with the substantial number of cases, the estimated cost of dementia worldwide is around $1.3 trillion USD per annum [11]. For comparison, $1.3 trillion USD is larger than the GDP of the 14th largest economy. [11].In the United States, the annual cost of care for an individual is approximately $50,000 [12]. As the number of patients with Alzheimer's is expected to rise, the costs are expected to rise proportionally. To help counteract the expense more cost-efficient testing is necessary. Currently, dementia is diagnosed only after a multi-part test consisting of a neurological examination, brain scans, psychiatric evaluation, and in some cases, genetic and blood tests [10]. Neurological examinations can cover a variety of topics from memory, language skills, as well as balance and reflexes [10]. For example, the Mini-Mental State Examination (MMSE) consists of a series of questions and cognitive tests to test orientation, registration, attention and calculation, recall, language, and constructional ability [5].

However, the accuracy of diagnosis is ultimately dependent on the clinician's ability to diagnose subtypes of dementia [2].

To help prevent possible human error, time, and cost, machine learning algorithms capable of classifying patients have been under study for the past decade. There have been numerous recent publications regarding analysis of speech in dementia patients. Habash, Guinn, et. al. found significant performance rating when using Decision Trees on disfluencies in speech; more specifically, go-ahead utterances, repeats, incomplete words, filler phrases, forward paraphrasing, and syllables per minute [1]. Eyigoz, Mathur, et. al. also used disfluencies in speech to support the hypothesis that machine learning algorithms can be used to detect early onset Alzheimer’s of the dementia type [3]. Chen, Lien, et. al. made a connection between gait parameters and dementia [6]. Gait parameters include pace, rhythm, variability, asymmetry, and posture control [6]. Oriol, Vallejo, et. al. found significant traces between certain genes and blood biomarkers and individuals with Alzheimer’s disease [9].

All of the previous studies have connected an area of simple human data and machine learning with high levels of efficiency. With this in mind, this project is to further build on the ability to use the analysis of features of speech to build a classifier that can distinguish between Alzheimer’s of the dementia type (AD) and healthy (non-AD) patients.

3. Methods

A set was built using a total of 28 features from the areas of speech disfluency, aphasic language, and lexical features. For each conversation, each feature frequency is generated using a simple counting function. The function builds the count by searching for the marker associated with the feature using *regular expression search.* Regular expressions are patterns of symbols that can be searched for in text, hence a regular expression could be the marking associated with a feature. Each feature count is divided by the total word count of the conversation. After all conversations have had their features counted, a data frame is generated using the Pandas library, which can be seen in Figure 1. The data frame consists of rows containing the classification of the main subject from the conversation, the word count of the conversation, and each individual feature count. The data frame is then sent to each of the machine learning algorithms, which include a Random Forest classifier, Naive Bayes Gaussian classifier, and a Support Vector Machine.

Label	Word Count	Prolongation	Pauses	Repetition	Overlapping	...
1	92	0.02173913	0.065217391	0.02173913	0	...
1	239	0.020920502	0	0.012552301	0	...
1	148	0.02027027	0.060810811	0	0.006756757	...
1	102	0.019607843	0.117647059	0.009803922	0.019607843	...
1	103	0.019417476	0	0	0	...
0	68	0.058823529	0.088235294	0.014705882	0	...
0	144	0.055555556	0.020833333	0.027777778	0.020833333	...
0	75	0.053333333	0.306666667	0	0	...
0	75	0.053333333	0.04	0.013333333	0	...
...	...	...	...	...	...	...

Figure 1 - Modified Version of Feature Count Data Frame

4. Data

The data for this project came from DementiaBank, which is a database used to study communication in dementia [15]. A total of seven corpora were selected, which came in the form of five projects with individuals with Alzheimer's, one project with individuals with mild neurocognitive disorder, and one project with a control group (non-Alzheimer's) [17, 18, 19, 20, 21, 22, 23]. Each corpus is a collection of transcribed conversations which use the CHAT format, which was developed by Brian MacWhinney as a part of the CHILDES project [15]. This project is using a total of 2751 transcribed conversations. The CHAT format creates a special encoding that is used to mark areas of speech such as speech disfluency, aphasic language, and lexical features [15]. For example, pauses in speech may be denoted by markers found in Figure 2 [15].

(.)	Denoting a small pause
(..)	Denoting a medium pause
(...)	Denoting a large pause
(N)	Denoting a pause of N seconds

Figure 2 - Markers for pauses in speech.

5. Features

This project looked at 28 features from speech disfluencies, aphasic language, and lexical features. These areas are directly related to the ability to communicate expressions and concepts, which is drastically reduced for people with dementia.

5.1 Speech Disfluencies

Speech disfluencies are stutters, breaks in speech, or any form of interruption to the flow of natural speech. The selected features for speech disfluencies include the following:

- Blocking: Captured by the ≠ symbol in the CHAT format [15]. A stop in the flow of speech, often caused from an awareness of stuttering or a lack of confidence in speaking situations [24].
- Broken Word: Captured by the ^ symbol in the CHAT format [15]. Recognized as a pause that occurs within a word [15].
- Fillers: Captured by the &-(filler) symbol in the CHAT format [15]. Short utterances that are often used to gain time so the speaker can gather or recall the necessary information [15].
- Fragments: Captured by the +... symbol in the CHAT format [15]. Occur when the speaker breaks the word into distinct utterances of syllables and/or uses syllables of a desired word as an utterance. For instance, “sp… speak” or “sh… e (she)” [15].
- Pauses: Captured by the (.), (..), (...), or (N) symbol in the CHAT format [15]. Represents either a short, medium, long, or second duration of a pause between utterances.
- Prolongation: Captured by the : symbol in the CHAT format [15]. The stretching of sounds and syllables, which is usually an act of stuttering [24].
- Reformulation: Captured by the [///] symbol in the CHAT format [15]. Occurs when a speaker redirects an entire concept in an utterance [15]. For example, “All my friends had… uh we all decided to go home for lunch” [15].
- Repetition: Captured by the [/] or [xN] symbol in the CHAT format [15]. This includes repetition of a whole word or a whole phrase. Can be noted as the number of repetitions in a row as [xN] where N is the number of repetitions [15].

- Retracing: Captured by the [//] symbol in the CHAT format [15]. Occurs when the speaker starts to say something, stops, and repeats the basic phrase, changes the syntax but maintains the same idea [15].
- Unclear Retracing: Captured by the [/?] symbol in the CHAT format [15]. Helps distinguish between pauses and retracing [15].

5.2 Aphasic Language

Aphasia is a language disorder caused by damage in a specific area of the brain that controls language expression and comprehension [14]. The loss of expression often turns to the speaker either shortening previously known words, creating words that may not be real or may not have the intended meaning, and or using gestures to fill in place of the missing words. The selected features for aphasic language include the following:

- False Start: Captured by the [/-] symbol in the CHAT format [15]. Occurs when a speaker begins a thought that succeeds a previous-unfinished response or thought [15].
- Gestures: Captured by the &=(gesture) symbol in the CHAT format [15]. These include communication through hands or motions in place of words.
- Interposed Words: Captured by the &* symbol in the CHAT format [15]. Includes added utterance from a speaker during another speaker's utterance [15]. For example, a speaker could be listening and interject "mhm-hm" or "yep" to signal agreement or sympathy for the situation.
- Overlapping: Captured by the +< symbol in the CHAT format [15]. Acts as a subtle interruption and is defined by the second utterance occurring during the first and taking place of the main conversation [15].

- Paraphasias: Captured by the [*] symbol in the CHAT format [15]. Occurs in place of misspoken words. More specifically, when the speaker says the wrong word but is clearly thinking and indicating towards the right word.
- Self-Interruption: Captured by the +//. symbol in the CHAT format [15]. Helps distinguish between trailing off and incompletions. Occurs when the speaker breaks an utterance only to start another [15].
- Shortening: Captured by the (part of word missing) symbol in the CHAT format [15]. Occurs when the speaker speaks in slang and all syllables of the word are not finished, like "goin" instead of "going" [15].
- Trailing Off: Captured by the +//. symbol in the CHAT format [15]. Indicates the speaker getting off track within the utterance; in other words, the speaker is responding to or acting on a thought outside the topic of conversation [15].

5.3 Lexical Features

Lexical ability is related to the ability to use words, verbs, and formulate speech. By measuring lexical density and sophistication, a speaker's level of speech can be measured. This of course, can be telling of whether or not the speaker has dementia, as it would be lowered by the disease. The selected lexical features include the following

- Lexical Density: Ratio of lexical words to total number of words in text [4]. However, there is no general consensus about word classification with regards to being a lexical word [4]. Kurdi discusses this manner as two classes were tested; however, neither were considered significant [4].

- Lexical Sophistication: Ratio of sophisticated words to the total number of lexical words [4]. Kurdi used a list of the 5000 most frequent English words called the Word Frequency Data [4]. Kurdi recognizes a sophisticated word as a word that has a frequency rank of 3000 or higher.
- Number of Different Words: This is the total count of different words in the text. This can be used to measure lexical density or lexical variation, which is a measure of richness of the lexical forms in the text [4].
- Type Token Ratio (TTR): Ratio of number of different words to total number of words [4]. Kurdi implemented two corrections of TTR, which each performed with high influence or relevance [4]. These corrections are included in the project and are known as Guiraud's corrected TTR and Carroll's corrected TTR [4].

6. Machine Learning Algorithms

The machine learning algorithms used in this project were a Random Forest classifier, the Gaussian version of a Naive Bayes classifier, and a Support Vector Machine. In the literature that applies machine learning to dementia and communication, these are popular and commonly used algorithms. Random Forest is a simple algorithm to implement yet is often highly effective. Naive Bayes is known to perform particularly well on small data sets, which are commonly encountered in speech studies. Lastly, Support Vector Machine has often provided the best results in previous studies.

6.1. Random Forest Classifier

The Random Forest classifier is a collection of a random number of randomly generated decision trees. The decision for the random forest comes from the majority classification from all decision trees in the forest. Each decision tree starts at the “root” and moves down the “branches ”until a “leaf” or classification is reached. A diagram of a decision tree used in the project can be seen in Figure 3, where labels have been provided at each node and all arrows and lines represent branches. For example, a decision tree could be used to determine whether or not to wear a hat for a given day. In this case, the algorithm will start at the root with a question asking about if it will be sunny or not. If it is sunny, then it will move down a branch to the left or right depending on the tree. The algorithm will then ask a new question that is dependent on the condition of it being sunny. This process will continue until a leaf is found containing an answer of either no hat or hat. The difference between using decision trees alone versus in a random forest is that all features are used when used alone. However, when generated in a random forest, each decision tree receives a random set of a random number of features, which can help bias when classifying. For this project, a Random Forest classifier with the default parameters from Sklearn was used. This included a base of one hundred decision trees.

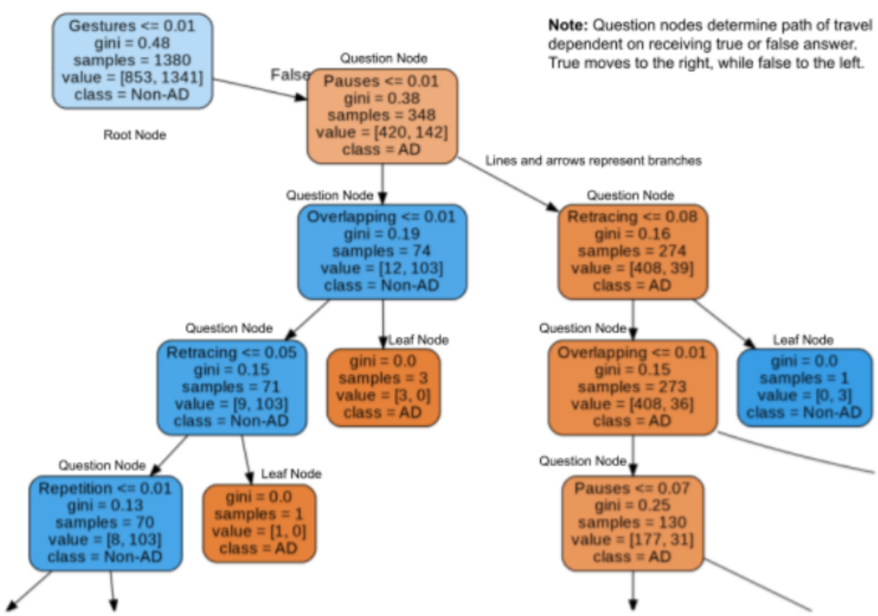


Figure 3 - False side of Decision Tree used in Random Forest

6.2. Naive Bayes Gaussian Classifier

The Naive Bayes classifier is an algorithm that is based on Bayes Theorem, which is used to calculate conditional probabilities and seen in Figure 4. The classifier treats each feature independently, which is great for efficient learning on smaller data sets. A limitation of the Naive Bayes classifier is that it is used on discrete data, which is data that is counted or whole number based. However, since the data used in this project is in the form of ratios, a Gaussian distribution is applied to this classifier in order to allow for the input of continuous data. Continuous data is data that can be measured, like the ratios of feature frequency to total word.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

P(h): independent probability of *h*: *prior probability*
P(D): independent probability of *D*
P(D|h): conditional probability of *D* given h: *likelihood*
P(h|D): conditional probability of *h* given *D*: *posterior probability*

Figure 4 - Bayes Theorem

6.3. Support Vector Machine

The Support Vector Machine (SVM) works to separate classifications of data that is not necessarily linear. The SVM plots data values as classifications along axes that are denoted by features. For example, in this project, the x-Axis could be the ratio of pauses to total word count and the y-Axis be the ratio of fillers to total word count. The data points are then separated by an

N-dimensional (N is the number of features) hyperplane, which allows for a transformation of the graph. Once transformed, the data points will be able to be separated linearly [13]. At this step, the SVM attempts to find the hyperplane that allows for the most margin between the most inner points of each classification, which are referred to as support vectors [13]. An example of the maximum margin and optimal hyperplane can be seen in Figure 5.

7.

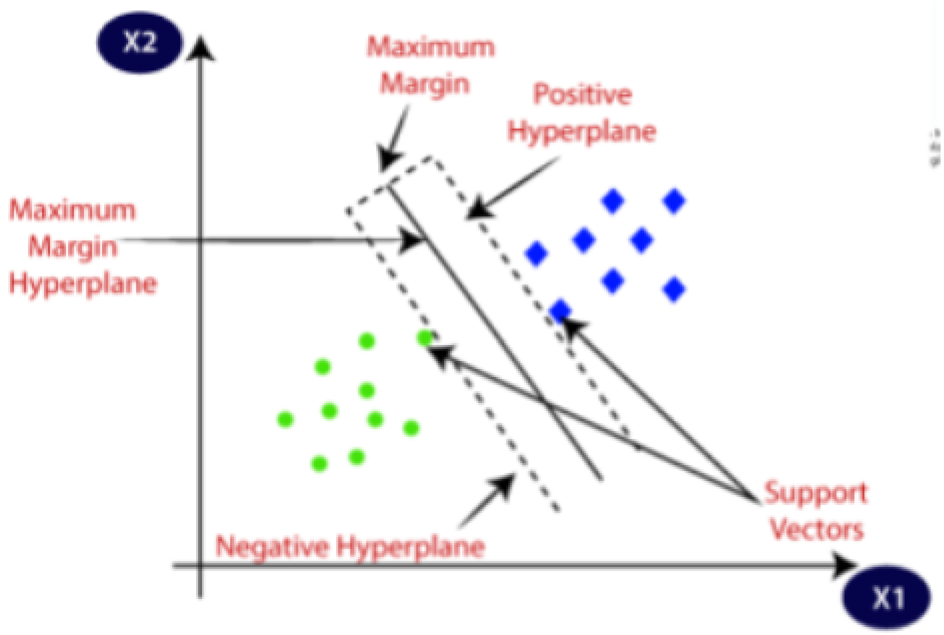


Figure 5 - Support Vector Machine

Results

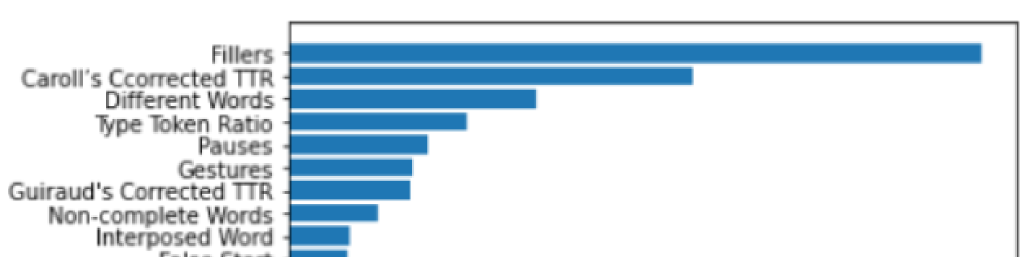
Between the three algorithms, two performed to standards of the current literature. When scoring the accuracy of a machine learning algorithm, there are four areas to consider: actual positives, actual negatives, false positives, and false negatives. Using these areas, algorithms are graded on their precision, recall, and F1 measure. Precision is the ratio of true positives to the sum of true and false positives, which is also the ratio of true positives to total predicted positives [8]. Recall is the ratio of true positives to the sum of true positives and false negatives, which is also the ratio of true positives to total actual positives [8]. Lastly, F1 measure is double the ratio of product of precision and recall to the sum or precision and recall [8]. F1 is the best measure to use when grading algorithms, because it eliminates bias of actual negatives [8].

All machine learning algorithms received an 80/20 train and test split. Overall, the Random Forest classifier performed best in all three measurements with a precision, recall, and f1 measure of 96.2%. Support Vector Machine, despite performing highest in most of the literature, had the lowest ratings in all three measurements as seen in Figure 6.

	Precision	Recall	F1 Measure
Random Forest	96.2	96.2	96.2
Naive Bayes Gaussian	88.7	88.2	87.9
Support Vector Machine	81.2	73.6	69.6

Figure 6 - Performance Scores of Algorithms

Random Forests provide a measure called feature relevance, which is the amount of influence a feature has on the decision classifier. As seen in Figure 7, lexical features have had the highest amount of influence, providing three of the top five features. These are Caroll's Corrected TTR, number of different words, and TTR. This would make sense as a large percentage of the corpus is conversations centered around describing pictures, events, and other activities. More specifically, most of the patients with Alzheimer's would have trouble remembering and communicating ideas and concepts; therefore, buying time with pauses or fillers would be a defense mechanism for the lack of recall.



8. Conclusion

In this project, three classifiers were built and proven to have reliable classification ratings in identifying patients in one of two groups - subjects with Alzheimer's of the dementia type (AD) and healthy (non-AD) subjects. This study employed a few thousand transcribed conversations and pre-built algorithms with default parameters. However, in a continuation of this project, an area of interest would be to study more features and implement a neural network. With regards to features, there are more areas of opportunity to study in speech and features from writing such as syntax, morphology, and phonology might also be included. Depending on the scale, the project could also extend to the physical motor ability of the patients. As previously studied, gait parameters have been studied with three simple tests of a straight distance walk, get up and go test, and jump test [6]. It would be interesting to see the influences of speech, writing, and gait parameters on machine learning as both exclusive sets and combinations of joined sets.

With the estimated rise in cases of dementia, and more specifically Alzheimer's of the dementia type, cost-efficient and high-performance testing methods are going to be necessary in the future. Currently, the testing methods are not only costly and time consuming, but also very intrusive. Patients can feel intimidation and or anxiety when performing examinations, which can potentially lead to skewed results. For instance, brain scans would be bad for patients that have trouble with still motions over extended periods of time and/or patients that suffer from claustrophobia. However, if testing methods could be greatly reduced to speaking, writing, and walking and jumping, it would greatly reduce intrusiveness onto patients.

Machine learning in the medical field is a trending topic of study because there is a wide use for these algorithms in everyday medicine [17, 18, 19, 20, 21, 22, 23]. It is clear that machine learning algorithms can act as reliable sources with regards to classifying patients. With this common and supporting knowledge, the field should look to implement these algorithms with suspected dementia patients. Since early-onset detection can lead to a massive reduction in the effects of dementia, machine learning could be used to keep track of various levels for patients. For example, any person with a family history of dementia or is suspected of being on track for diagnosis can begin recurring testing, where results are stored and processed by the algorithms. Patients can be tested on speech, writing, and gait parameters in very cost efficient and time efficient manners. For instance, some of the portions of the MMSEE, such as the Cookie Theft Description Task. In this task, the patient is tasked with looking at a picture and writing or verbally expressing their observations. The picture contains a family in a kitchen with multiple chaotic events happening, which are the points of emphasis for the patient to describe. In addition to this task, a patient could perform the three simple tests performed by Chen, Lien, Wu, Lee, and Shaw [6]. Over time, the patient's cumulative results or scores will be compared to a sort of baseline. This would not only help the testing methods, but also save money in care because early detection and care would better help main function in the patient over time.

9. References and Work Cited

[1] Anthony Habash, Curry Guinn, Douglas Kline, and Laurie Patterson. 2012. Language Analysis of Speakers with Dementia of the Alzheimer's Type . *Annals of the MS in Computer Science and Information Systems at UNC Wilmington* 6, 1 (2012). DOI:http://dx.doi.org/http://csbapp.uncw.edu/data/mscsis/full.aspx

[2] Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning Predictive Linguistic Features for Alzheimer's Disease and Related Dementias Using Verbal Utterances. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (June 2014), 78–87. DOI:http://dx.doi.org/10.3115/v1/w14-3210

[3] Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic Markers Predict Onset of Alzheimer's Disease. *EClinicalMedicine* 28 (October 2020), 100583. DOI:http://dx.doi.org/10.1016/j.eclinm.2020.100583

[4] M. Zakaria Kurdi. 2020. Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL. *Journal of Data Mining & Digital Humanities* 2020 (September 2020). DOI:http://dx.doi.org/10.46298/jdmdh.6012

[5] Aram So, Danial Hooshyar, Kun Park, and Heui Lim. 2017. Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques. *Applied Sciences* 7, 7 (June 2017), 651. DOI:http://dx.doi.org/10.3390/app7070651

[6] Pei-Hao Chen, Chieh-Wen Lien, Wen-Chun Wu, Lu-Shan Lee, and Jin-Siang Shaw. 2020. Gait-based Machine Learning for Classifying Patients with Different Types of Mild Cognitive

Impairment. *Journal of Medical Systems* 44, 6 (April 2020). DOI:http://dx.doi.org/10.1007/s10916-020-01578-7

[7] Javier De Velasco Oriol, Edgar E. Vallejo, Karol Estrada, José Gerardo Taméz Peña, and The Alzheimer's Disease Neuroimaging Initiative. 2019. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from Genomic Data. *BMC Bioinformatics* 20, 1 (December 2019), 1–17. DOI:http://dx.doi.org/10.1186/s12859-019-3158-x

[8] Farheen Ramzan et al. 2019. A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state firm and residual neural networks. *Journal of Medical Systems* 44, 2 (December 2019). DOI:http://dx.doi.org/10.1007/s10916-019-1475-2

[9] Sylvester O. Orimaye, Jojo S.-M. Wong, Karen J. Golden, Chee P. Wong, and Ireneous N. Soyiri. 2017. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics* 18, 1 (February 2017). DOI:http://dx.doi.org/10.1186/s12859-016-1456-0

[10] NIH National Institute on Aging. 2021. What is dementia? symptoms, types, and diagnosis. (July 2021). Retrieved January 7, 2022, from https://www.nia.nih.gov/health/what-is-dementia

[11] Alzheimer's Disease International (ADI). Adi - Dementia Facts & Figures. Retrieved April 28, 2022, from https://www.alzint.org/about/dementia-facts-figures/

[12] Michael D. Hurd, Paco Martorell, Adeline Delavande, Kathleen J. Mullen, and Kenneth M. Langa. 2013. Monetary costs of dementia in the United States. (April 2013). Retrieved April 28, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3959992/

[13] Rohith Gandhi. 2018. Support Vector Machine - introduction to machine learning algorithms. (July 2018). Retrieved April 28, 2022 from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[14] Johns Hopkins Medicine. 2022. Aphasia. (2022). Retrieved April 28, 2022, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/aphasia

[15] MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

[17] MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*,1286-1307

[18] Kempler, D., Curtiss, S., & Jackson, C. (1987). Syntactic preservation in Alzheimer's disease. *Journal of Speech, Language, and Hearing Research, 30*(3), 343-350.

[19] Lanzi, A., Wallace, S. E., & Bourgeois, M. (2019). Group external memory aid treatment for mild cognitive impairment. *Aphasiology, 33*(3), 320-336.

[20] Becker, J. T., Boller, 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.

[21] Herd, P., Carr, D., & Roan, C. (2014). Cohort profile: Wisconsin Longitudinal Study (WLS). *International Journal of Epidemiology 43*, 34-41.

[22] MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*,1286-1307.

[23] Sebastian, R., Thompson, C. B., Wang, N. Y., Wright, A., Meyer, A., Friedman, R. B., ... & Tippett, D. C. (2018). Patterns of decline in naming and semantic knowledge in primary progressive aphasia. *Aphasiology, 32*(9), 1010-1030.

[24] Candace Ganz. 2019. Home. (2019). Retrieved April 28, 2022, from https://www.cincinnatichildrens.org/health/s/stuttering