SMU Data Science Review

Volume 5 | Number 1 Article 5

2021

Rule out Screening for Undiagnosed Dementia and Alzheimer's Disease Using an EHR Based Machine Learning Solution

Branum Stephan
Southern Methodist University, bstephan@smu.edu

David A. Julovich
Southern Methodist University and University of North Texas Health Science Center at Fort Worth, david.julovich@unthsc.edu

Dustin Bracy
Southern Methodist University, dbracy@smu.edu

Jeff Nguyen
Southern Methodist University, jeffn@smu.edu

Recommended Citation

Stephan, Branum; Julovich, David A.; Bracy, Dustin; and Nguyen, Jeff (2021) "Rule out Screening for Undiagnosed Dementia and Alzheimer's Disease Using an EHR Based Machine Learning Solution," *SMU Data Science Review*: Vol. 5: No. 1, Article 5.
Available at: https://scholar.smu.edu/datasciencereview/vol5/iss1/5

Rule out Screening for Undiagnosed Dementia and Alzheimer's Disease Using an EHR Based Machine Learning Solution

Branum Stephan, B.S.[1], David A. Julovich, B.S.[1,2], Dustin Bracy, B.S.[1], Jeff Nguyen, B.S.[1],
1 Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
2 Institute for Translational Research - Department of Pharmacology & Neuroscience,
University of North Texas Health Science Center
Fort Worth, TX USA
{bstephan, djulovich, dbracy, jeffn}@smu.edu

Abstract. Current detection methods for Dementia and Alzheimer's disease include cerebral spinal fluid (CSF) markers and/or the use of positron emission tomography (PET) imaging, both being high-cost, highly invasive testing methods. The need for low-cost, minimally invasive methods to prescreen individuals for cognitive impairment has been a challenge for many years. Today's costs associated with an annual screen for all adults 65 and above using current methods (CSF, PET) reach well beyond trillions of dollars per year. Motivated by the limited accessibly and high costs, an alternative tool presented within this paper demonstrates an effective rule out screening for Dementia and Alzheimer's disease. Leveraging Electronic Health Records (EHR) data, low-cost computing, modern statistical modeling, and useable Machine Learning algorithms were able to derive a screening tool that effectively detects 98 percent of individuals without Dementia and Alzheimer's disease. Approximately 43,000 EHR patient records' totaling 5,000 patients from the University of North Texas Health Science Center were evaluated using this new rule out screening method, which consists of traditional Machine Learning models: Random Forest, AdaBoost, SVM combined with the application of Natural Language Processing of physician's notes. The findings from this study help define a new paradigm in medical practice where an effective rule out screening method for Dementia and Alzheimer's disease can be used as an initial screening tool. This study effectively cuts the cost of current detection methods by 75 percent, gives access to all adults age 65 and above while still leaving expensive methods as secondary lines of detection.

1 Introduction

For over 20 years, there has been the need for low-cost, minimally invasive Dementia and Alzheimer's disease screening methods for patients to better determine who does not have the disease. By identifying those not at-risk from those at-risk within differing healthcare environments, it is now possible to generate savings and better utilize current high-cost testing methods (CSF/PET). A rule out approach greatly reduces the number of patients escalated for higher cost neurodiagnostic testing. Rule out screening of Dementia and Alzheimer's disease gives patients peace of mind because the results indicate the disease is not present. While at-risk patients are not diagnosed, they would be referred on for additional neurodiagnostic testing; consistent with other screening methods currently employed in other areas such as oncology.

Machine Learning models trained on EHR data using unstructured doctor's notes can provide a useful method for ruling out those without Dementia and Alzheimer's disease within a traditional clinic and hospital setting. Such an approach, to date, has not been evaluated particularly with advanced statistical analytic methods (i.e. Machine Learning). Despite the potential for the proposed applied Machine Learning solutions, there are industry-wide hurdles regarding access to data that needs to be overcome when using Electronic Health Records.

The Department of Health & Human Services (HHS) is committed to increasing interoperability including the ability to exchange, interpret, and use medical data cohesively, across all healthcare settings, servicers, and providers. As of 2015, 96 percent of hospitals and healthcare providers along with 78 percent of office-based physicians use digital records [21]. The current disconnects in customer records is a major issue. Neither patients nor doctors have consistent access to a patient's entire record history from varying institutions. Therefore, disconnections exist between the different institutions, which means there not consistent data standards and or means. of data collection. Unique EHR systems will require custom installation and setup of the rule out tool within each environment.

A low-cost, highly accessible screening tool is presented here. EHR records can be successfully utilized and quickly assembled into a functional resource that is used to screen for Dementia and Alzheimer's disease. The addition of Natural Language Processing shows great promise as a new component to the screening tool, creating untapped features that are highly important and produce better model predictions. The findings discussed within add value to current medical and research literature by demonstrating a screening tool that effectively detects individuals without Dementia and Alzheimer's disease for patients that visit their doctor.

2 Alzheimer's disease

As an overview, Alzheimer's disease is a neurodegenerative disease that is particularly common in the senior population of 50 years and above. The disease was discovered in 1906 by Dr. Alois Alzheimer during an inspection of neural tissue from a deceased elderly woman who had reported unusual mental illnesses. Some of her most notable characteristics were memory loss, language difficulty, and erratic behavior. Physical characteristics of her brain included abnormal amyloid plaques and neurofibrillary tangles which today are the hallmark characteristics synonymous with Alzheimer's disease [22].

2.1 Overview and Impact

As Alzheimer's disease progresses in the brain, a buildup of amyloid plaques occurs along with a noticeable destruction of neuronal connections reflecting neurodegeneration. Typically, the hippocampal brain region (associated with memory) is the first area to have noticeable damage followed by other areas of the brain. Once Alzheimer's disease has progressed to the state of being clinically detectable, there may be damage throughout the brain along with a reduction in overall tissue volume [24]. The neurodegeneration occurs over time and eventually leads to a decline in cognitive and functional abilities [16]. Alzheimer's disease is known to begin years before symptoms are detectable. The lack of detection allows degenerative biological states to exist within individuals where disease pathologies can go undetected and unchecked for several decades before symptom presentation, such as memory loss, are seen [2],[3],[6],[11],[27].

The 2020 Alzheimer's Association Report identifies the most common causes of Dementia to be Alzheimer's disease, Cerebrovascular Disease, Lewy Body Disease, Frontotemporal lobar degeneration, Parkinson's Disease, and Hippocampal sclerosis. Mixed pathologies occur when an individual shows more than one cause of Dementia [16]. This study highlights the fact and figures of Alzheimer's disease due to the high prevalence of disease; however, Alzheimer's disease was not the only form of Dementia within the data. Dementia, in general was used for analysis. Alzheimer's disease is the most notable and prevalent form of Dementia and contributes to around 60 to 70 percent of all Dementia cases. The term Dementia describes the neurocognitive and functional decline of patients and is the overarching terminology used to characterize symptoms of memory loss along with changes in thinking, language skills, and problem solving. This study utilizes Alzheimer's disease as an additional outcome due to the high prevalence of the disease; however, Alzheimer's disease was not specifically identified within the data. Dementia, in general was therefore used for analysis as the primary outcome variable.

Most concerning is the seriousness of the damage caused by Alzheimer's disease and the staggering occurrence rates. As of 2017, experts have estimated 5.5 million Americans have Alzheimer's disease, and the disease is not just limited to the senior community. The incident rates of Alzheimer's disease continue to increase as the global population grows and continues to age [22]. To put these large numbers into perspective, 1 in 10 people age 65 and above experiences Alzheimer's disease or some form of Dementia [9]. The prevalence (existing cases) of Alzheimer's disease also increases with age; ranging around 3% of people aged 65−74 to 32% of people age 85 and above [9]. Patient lifespan post diagnosis may range from 3−10 years, Alzheimer's disease is estimated as the sixth leading cause of death in the United States and possibly the third leading cause for the older community. The only other causes of death that ranked higher for the elderly population was heart disease and cancer [22].

It is estimated that more than 16 million unpaid caregivers provide care for individuals with Alzheimer's disease and Dementia. These unpaid caregivers provided an estimated 18.6 billion hours of unpaid care, valued at $244 billion dollars in contributions in 2016 [16]. Caring for a person with Alzheimer's disease has special challenges and levels of care will likely increase with the progression of the disease. As the disease worsens, more serious health problems usually arise. Incomes and finances regularly become depleted causing high levels of emotional stress and depression for many of these caregivers [16]. Direct care (paid) workers, usually nursing assistants, consistently have a difficult job due to a lack of training and education required for Alzheimer's disease and Dementia care.

Making matters worse, as of 2016, there is a shortage of certified geriatricians with only 7,293 in the United States. That is one geriatrician for every 1,924 Americans age 65 or older in need of care [16]. The American Geriatrics Society estimates that by 2030 an additional 23,750 geriatricians are needed to meet the aging population [16]. In 2017, Medicare started reimbursing healthcare professionals for comprehensive Dementia health care plans for care visits to help address short comings in healthcare planning. Medicare stated prehensive care planning is a core element of effective Dementia care management and can result in the delivery of services that potentially enhance quality of life for people with Dementia and their caregivers [16].

The economic impact and cost have enormous implications for individuals and their families affected by Alzheimer's disease. Lifetime costs for care of a single individual were estimated at $357,297 in 2019 [14]. Alzheimer's disease and Dementia have become the costliest condition, with over $290 billion made in payments 2020 for health and long−term care [10]. Considering the population of individuals currently with Alzheimer's disease today (5.8 million) and the number of individuals forecasted in future (88 million affected by 2050), the estimated cost is projected to reach $4.6 trillion. The estimated number of future cases, high costs of care and ageing population push the importance of an early diagnostic test to a real priory for public healthcare.

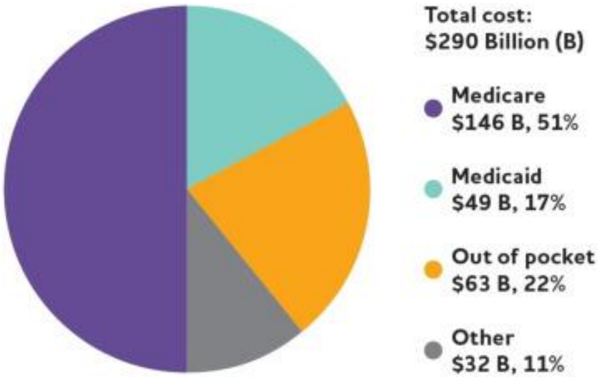

Figure 1: Aggregated costs of healthcare payments made by Americans age 65 and older with Alzheimer's disease or dementia in 2020. Reported in the Alzheimer's Association Report and data displayed from the Lewin Model [22].

Despite the dire consequences of developing Dementia or Alzheimer's disease, there is currently no cure, and there is still much to learn about the disease itself. There are currently only 5 FDA approved medications, none of which has shown supportable evidence for the permanent reduction/remission of Alzheimer's disease symptoms. Rather, the only limited achievements of these medications are to manage and slow the symptoms of neurocognitive degradation characterized by Alzheimer's disease and Dementia. To say medication development has been slow would be an understatement, as there are no medications that prevent or delay the progression of the disease. The most recently FDA approved Alzheimer's disease medication is simply a cocktail of two other Alzheimer's disease drugs from 23 and 16 years ago. There are efforts to develop new medications. However, clinical trials have seen a staggering 99.6% failure rate.

Many research documents cite major challenges in that there is still not much known about the cause of Alzheimer's disease and that finding Alzheimer's disease patients at early development stages is crucial. It has been estimated that potentially half of patients with Alzheimer's disease are undiagnosed or experience anosognosia (inability to recognize they have memory changes) [16],[11]. Not having access to effective screening solutions perpetuates missed diagnoses. A rule out screening method that can effectively identify most of the population without the presence of Alzheimer's disease or Dementia allows the focus to shift towards those at-risk patients. Finding suitable clinical trial participants has been difficult and has high costs associated with it. Without a suitable pool of subjects, it becomes extremely difficult to carry out research on Dementia and Alzheimer's disease. The rule out screening method is a novel way to overcome some clinical trial hurdles by lowering the cost of inclusion and exclusion criteria for trials ensuring the focus and expense is used on individuals who can help drive development of potential lifesaving measures, treatments, and medications [1].

2.2 Detection Approaches and Related Works

The gold standard screening methods used to identify Dementia and Alzheimer's disease are Positron Emission Tomography (PET) and CSF marker analysis, each have their own respective tradeoffs. To date, there have not been any predictive models approved by the FDA, which creates a gap in screening. Gold standard methods are expensive and may not be covered all by insurance providers. In recent years, research advancements of noninvasive screening methods have been developed utilizing predictive analytics.

The major advantage in utilizing Machine Learning methods is that screening can be done on larger populations and at-risk individuals can be identified in a low-cost, noninvasive manner. Therefore, using predictive analytics to identify undiagnosed Alzheimer's disease and Dementia can provide real value to the realm of research and treatment by way of earlier diagnosis. Although the field of predictive analytics has been around for quite some time, recent advancements in computing capabilities and research in the field have led to much more widespread adoption in the industry. Predictive Analytics and coined terms such as "Machine Learning" and "artificial intelligence," consists of preparing, cleaning, and processing suitable training data in order to train a series of predictive mathematical models. From there, the models will typically "learn" by calculating a loss function to measure the total accuracy (or other chosen performance metric) of the model and then back propagation seeks to adjust model weights in order minimize aforenoted loss function.

Several groups have demonstrated success using Machine Learning and predictive analytics. McCoy et al (2018, 2019) aimed to help prevent Dementia by identifying pre-symptomatic and/or individuals with minimal symptoms that were at high-risk for Dementia. McCoy and colleagues identified that traditional current detection of Alzheimer's disease and Dementia are expensive, and that reliability and scalability are not consistent. McCoy subsequently proposed that an Electronic Health Record (EHR) method of screening could be a potential solution for early detection [18]. In their work, Natural Language Processing (NLP) was used on free text from hospital discharge notes where a "Cognition Score" was generated and compared with other medical records. These discharge notes capture key indicators of Alzheimer's disease that may not be captured by ICD−10 codes. ICD-10 codes are a standardized set of codes that classify and document disease diagnoses. For analysis, each study site data was merged to form a holistic view [17]. Association between model score and Alzheimer's disease related death rates were then studied to better understand the progression of Alzheimer's disease. The McCoy paper further states additional validation is needed to show the effectiveness of the NLP approach and the need for future iterations of the model to incorporate other biomarkers and medical records. From McCoy's work, a lexicon containing Alzheimer's disease and Dementia key words and stop words has been made available. Incorporating the limitations mentioned by McCoy, use of additional fields and NLP analysis may build a more robust and predictive model.

The Scripps Research Translational Institute utilized over 1 million patient medical records representing 11 years of data. The incidence of Alzheimer's disease was low in the study population. To allow the models to better classify the low volume of Alzheimer's records, the data was balanced using bootstrap sampling. Random Forest, Support Vector Machine, and Logistic Regression were utilized on 5 groups of data with time points zero, one, two, three, and four subsequent years incidence of Alzheimer's disease. Within each group the models classified Alzheimer's records with the following operational definitions: definite Alzheimer's disease and probable Alzheimer's disease based on patient medical records. The top 20 features were selected using Logistic Regression, then used in other models for comparison. Several key features that were identified were elevated urine protein, Zotepine, and several ICD-10 codes. These features were deemed statistically significant and produced high odds ratios for Alzheimer's disease suggesting that these features were more strongly associated with incidence of disease.

The Scripps group showed Random Forest models had the highest accuracies for most of the subsequent years of incidence for definite and probable Alzheimer's disease. Predictions are acceptable for 0−and 1−year incidence with an AUC of 0.898 and 0.775 respectively [18]. However, as time increases to 4 years the model performance starts to decrease. This trend was consistent across all models for prediction of definite or probable Alzheimer's disease. This research does not mention the use of any free form text field or the use of Natural Language Processing techniques to model this data. Relevant information captured from a doctor's observations could provide key indications for Alzheimer's disease incidence.

Another interesting point presented by the Park et al paper described feature selection using a Logistic Regression approach for comparison with other models. Several key risk factors were identified and have shown to contribute to Alzheimer’s through their high odds ratios; this can be used as a potential starting point for our group’s models [23].

The papers from the Center for Quantitative Health at Massachusetts General Hospital, Harvard, and Scripps Research Translational Institute both suggest that document classification can be used to improve the efficiency and increase robustness of predictive models. The paper by Peng et al shows how document classification accuracy can be improved over traditional "bag−of−words" using Natural Language Processing techniques to extract word senses and from the context of the entire document. This approach provides a potential 14% increase in accuracy over bag−of−words techniques [24]. Peng suggests a shift in modeling from word frequencies to sense frequencies. Senses rely on the meaning of words and the relationships between groups of meanings [24].

Shifting from a lexical to a semantic approach allows for a better understanding of how words are related in the context of their meanings. There are situations where the same word is used but has different meanings. A quarter can mean an American coin, twenty−five percent, or a football formation. Peng’s research in semantic analysis suggests gains in accuracy over lexical techniques and offers a high−level approach to handle documents to help generate better accuracy. These techniques can be used on doctor notes contained in EHR to help understand the relationship between comments captured in the notes.

In lexical analysis this information is not utilized, and the frequency of a word is measured. However, in semantic analysis the meaning of the word is utilized and is dependent on the context of the words surrounding it as seen in Figure 2.

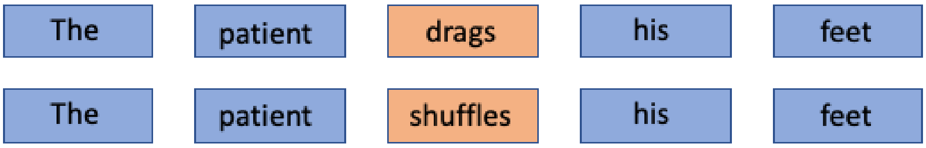

Figure 2: Example of semantic similarity of words in context. (target orange, context blue)

To represent a document, a hierarchy of word senses is first constructed. Words are then mapped to synsets, groupings of words with similar meaning, which are then labeled with a word that captured the meaning of their relationship. Following this, the relationships are captured, and documents are then compared to predefined categories with each synset document representation. Documents with the closest similarities to the category were then grouped with that particular category [24]. Peng et al mention that the methods studied were limited by the lexicon utilized. Language is constantly growing and evolving. If a lexicon is not updated, lexical or semantic approaches may not be sufficient to cluster future documents as new words and their meanings may not be captured. This argument is held for any NLP project as most projects rely on current lexicons.

2.3 Combining Approaches

For research that utilized NLP, authors mentioned incorporating other biomarkers or features that were contained in other medical records to build more robust models. Conversely, in papers that utilized continuous or categorical features, researchers often mentioned incorporating free form text into their models to increase model performance. In either case, free form text in EHRs contained in doctor notes could contain insights related to symptoms associated with Alzheimer's disease or Dementia, and biomarkers and other health records may contain attributes characteristic of Alzheimer's disease or Dementia. This research aims to improve upon the limitations mentioned in the previous research by incorporating both free form text from doctor notes in EHR records or other biomarkers and continuous data a patient may have.

3 Methods

This research makes use of several Machine Learning algorithms which were chosen due to their widespread acceptance in medical industry and transparency compared to more complex approaches, such as neural networks. Considering the need for ease of interpretation by the doctors and medical staff, several models meeting this criterion were evaluated on EHR medical records. Given the prevalence of Alzheimer's and Dementia are low, negative predictive value was deemed the best metric to judge model performance. Negative predictive value and positive predictive value statistics are regularly used to describe the performance of diagnostic tests. Negative predictive value also considers the prevalence of the disease and therefore considers what percent of the time the test correctly detects no disease. It does so by comparing the total number of true negative to all negative predictions (true and false). The Negative Predictive value was used to measure model performance for those individuals who do not have Alzheimer's or Dementia

3.1 Logistic Regression

Logistic Regression is a classification algorithm commonly used in current detection approaches which may be used to solve single or multiple variable classification problems. Logistic Regression improves upon multiple linear regression (MLR) for categorical predictors by forcing all predictions to have values between 0 and 1, mitigating ordinal effects of multiple categories which may occur in a quantitative algorithm such as MLR [12]. This study will utilize multiple classifications with Logistic Regression and one-hot encoding, assigning a new feature representing a condition in the binary states of the condition is present or the condition is not present. The Logistic Regression models will assign coefficient weights to each feature and its effect on the outcome probability.

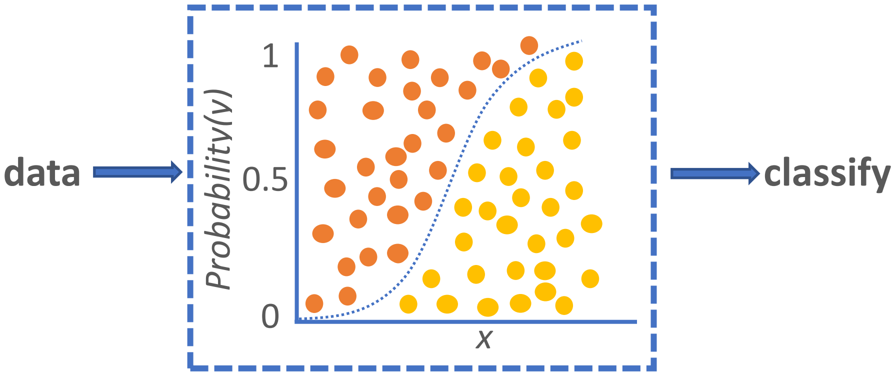

Figure 3: Example of classification using Logistic Regression.

This paper will utilize models making use of the log–odds ratio as it relates to the response, which may be interpreted as: adding one unit to X changes the log odds of the response by β1 or multiplying the odds by $e^{\beta1}$ [25]. This gives us the ability to interpret models using log odds. Ramsey and Schafer provide a practical example of interpreting log odds:

> *For example, it is possible with retrospective samples of lung cancer patients and of patients with no lung cancer to make a statement such as: The odds of lung cancer are estimated to increase by a factor of 1.1 for each year that a person has smoked. This result has tremendous bearing on medical case–control studies and the field of epidemiology. (p. 609) [25].*

Logistic Regression carries assumptions similar to MLR, in that high leverage outliers can severely impact prediction results, and the assumption is that residuals have equal mean and variance across the X predictors. These assumptions are checked and can be validated in the notebooks containing the code for this research paper.

3.2 Random Forest

Random Forest models will be used to benchmark model quality and will act as the baseline for comparison to other models and applicable research in other papers due to its ease of implementation and highly explainable nature. The model itself is comprised of an ensemble of decision trees which can either yield a continuous or categorical response [29]. These decision trees are often referred to as Classification and Regression Trees (CART) algorithms [7].

The anatomy of a decision tree consists of a series of decision nodes that split into branches/edges. Navigation down the decision tree continues through each node and down the branches until there are no more splits at which case a response, continuous or categorical, is the final step [28]. The manner in which a decision tree does this is based on the training features and conditions required for splitting each decision node. As a result of a technique coined recursive binary splitting, all provided features are considered and split points are created by finding the lowest cost for all possible splits [7].

For classification-based decision trees, a common cost function is the Gini score, which calculates a measure of the homogeneity of the response classes in each group of a potential split [7]. In the formula provided below, pk is the proportion of homogeneous response in a group. An ideal split will produce a pk value of 1 or 0, but a 0.5 for the worst fit on a binary classification problem [7].

$$\textit{Classification: } G = sum(pk * (1 - pk))$$

Due to the greedy, recursive nature of the algorithm to excessively minimize the cost function, a large number of splits may be present that could make model interpretation difficult and result in an overfitted model [7]. Decision trees can have exceptional performance on training data and are often quite easy to explain, but often under-perform in real-life scenarios due to their tendency to overfit the training data. In order to achieve a more useable model, various adjustments and techniques can be employed to increase model predictability and robustness.

One such possibility utilized in this research is an ensemble (or collection) of uncorrelated decision trees, often referred to as a Random Forest (aforementioned). For classification tasks, the Random Forest will have each decision tree create a prediction of a response and the response with the most votes is the ensemble output [29]. In essence, the ensemble model allows for the errors of an individual decision tree to be reduced or masked by the other trees in the ensemble [29]. By using the "public opinion" of the collection of the models as a whole, in contrast to that of a specific model, the ensemble method is much more robust against model overfitting and often leads to a higher model performance in real world applications [29].

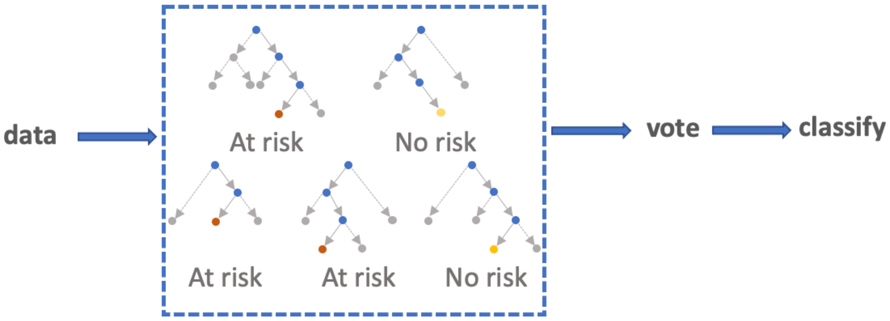

Figure 4: Example of an ensemble of decision trees created by a Random Forest classifier.

AdaBoost was also utilized and is similar to Random Forest in that it combines multiple weak learners to generate a strong learner. Weak learners don't generalize well to new data because they are overfit or have poor accuracy, but when many weak learners are combined, they generate a good, generalized representation of the data.

3.3 Support Vector Machine

Support Vector Machines (SVM) are a type of Machine Learning algorithm that can be used for both regression and classification. Although it can work in a regression problem, it is much more commonly employed in classification problems that utilize the creation of hyperplanes to differentiate between classes. Observations that are present on either side of the hyperplane are associated with a class.

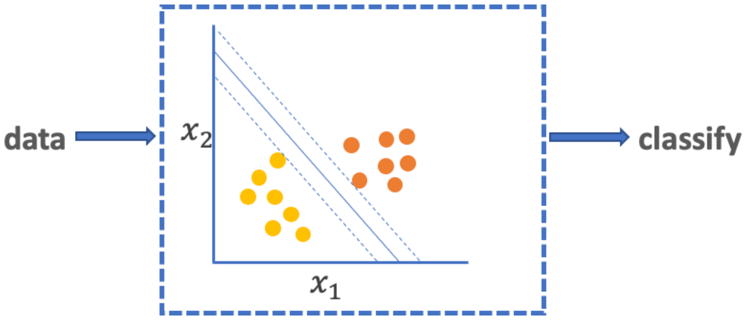

Figure 5: Example of a classification problem using Support Vector Machine.

The creation of the hyperplane is typically of n-1 dimensions where n is the number of dimensions present in the data [4]. For example, if a collection of observations with 3 features (excluding response) are trained with the algorithm, the hyperplane that splits the classes will be 2 dimensional [4]. Due to the infinite nature of a hyperplane, it is often more practical to describe the hyperplane as a sequence of vectors, often referred to as support vectors [4]. To improve model performance, the algorithm utilizes a hinge loss function to maximize the distance between the support vectors that achieves the greatest separation of classes [4].

In SVM, outputs are normalized to a range of -1 and 1 upon which margins of the support vectors run [4]. The cost function of the SVM yields a value of 0 if the prediction and actual values have the same sign, whereas opposite signs yield a measurable loss value [4]. The cost function will seek to maximize the margin of separation between each class and subsequently minimize loss [4].

In simpler terms, SVM is designed to maximize the distance between each class separating hyperplane, and the nearest points to it, also known as the support vectors. This distance is called the margin, and SVM attempts to draw the optimal line through the observed data that provides the most margin for all support vectors. This line is not always 100% linearly separable, so we can incorporate slack variables, which allow for observations to be on the wrong side of the hyperplane [8]. These slack variables are used in conjunction with residuals in the cost function to penalize large residuals by a user defined, constant C amount [15]. This constant C is one of the primary parameters of SVM that by changing, may result in better classification results as it adjusts the functional margin between support vectors.

3.4 Natural Language Processing

Natural Language Processing (NLP) is a process that allows for analysis of text data at scale and is the use of machine-based methods to process natural language – an example of this would be the use of NLP on doctor's notes from EHR. NLP consists of two main sub-categories of language processing: Natural Language Understanding (NLU) and Natural Language Generation (NLG) each comprised of subcategories that serve similar but different purposes, utilize different techniques, and generate different outputs.

NLU is a process used to create a useful representation of ingested natural language that is understandable to a machine. This requires the use of artificial languages, such as python, R, Java, etc. to provide instructions to a machine on how to cluster, classify, tag, and query text. NLG is the transformation of structured data into human comprehensible words, phrases, or sentences. Common application of this form of language processing includes text summarization, labeling of clusters, and question answering.

Text data was prepared on the lexical level to allow for unsupervised clustering of EHR features into groups that were most similar to each other. Term frequency (TF) and term frequency inverse document frequency (TF-IDF) values were used with k-means clustering, and TF and trigrams with Latent Dirichlet Allocation (LDA) for topic modeling. The clusters were captured as engineered features for utilization by Random Forest, Logistic Regression, or SVM model that contains other risk factors and features associated with a patient record like age, BMI, medical tests taken, and ICD-10 codes as seen in Table 1.

Table 1: Example of Electronic Health Record schema with NLP-engineered feature.

Patient Number	Gender	BMI	ICD-10	Physician Comments	NLP-Engineered Feature – At-risk for Alzheimer's
12345	F	22	F63.0	Patient described risky behavior when gambling	0
67891	M	26	F20	Patient described memory issues and difficulty speaking	1
01112	F	25	F03	Patient mentioned forgetfulness	1

3.5 Model Training and Hyperparameter Training

Data for modeling was partitioned into a train and test sets, where 70% of the data was used for training and 30% was used for testing. The national prevalence rate reported for Alzheimer's disease is only 10 percent. Since the number of records containing Alzheimer's disease or Dementia were infrequent it was important to up-sample these the training set to improve model performance. This was done using Synthetic Minority Oversample Technique (SMOTE), where synthetic records containing minority class attributes are generated. Up-sampling is performed when there is a class imbalance in the response. This is done because it can be difficult for a model to effectively classify when there is a class imbalance. If a model is run on data that is imbalanced performance metrics typically would underperform. The data set generated by SMOTE was then used in the grid search cross validation process during the training phase.

To maximize model performance an exhaustive grid search and cross validation was performed. Parameters for each model were selected and a range of values within each parameter were assigned to a parameter grid. The up-sampled training data was then loaded into a 3-fold cross-validation object. During the grid search, all permutations of the parameters were generated then trained on one portion of a fold and tested on the remaining portion. This process is repeated until all fold permutations are trained and tested. Once this process is complete, the grid search is able to identify which permutation of parameters for a particular model performs best in terms of a user specified metrics.

4 Data Description

The Electronic Health Records utilized for this research were obtained from the University of North Texas Health Science Center (UNT HSC) in Fort Worth. The data is a convenience sample from 3 different clinical settings: Geriatrics, Family Medicine and Seminary Medicine consisting of EHR for patients 65 years and older. The dataset spans from January 2016 to December 2019 and contains 4,995 unique patients totaling 46,690 patient visits. There were 3,950 normal patients, 516 Alzheimer’s disease patients and 889 Dementia (Alzheimer’s disease and others) patients. The total number of patients that are diagnosed with Alzheimer’s disease or Dementia as a proportion of the total is at 17.79%. The total proportion of patients with Alzheimer’s disease is calculated at 10.33%, which closely matches the 10% expected in the general population.

Table 2: All tables were relationally linked by Encounter_ID and Person_ID.

Table	Definition
Encounters	Data related to a patient visit where a patient could have multiple visits.
Vitals	Data related to a patient's height, weight, BMI, pulse, etc.
Assessments	Related ICD-10 and doctors' free-from assessment detailing the visit.
Labs	Data related to a patient's ordered laboratory tests and corresponding results.
Medications	Data related to patient's prescribed drug(s), dose, current or not current.
CPT_codes	Codes related to patient's tests, surgeries, evaluations for billing and the patient's free-from reason for visit.
ICD-10	Codes related to patient's diagnosis recorded as codes.

4.1 Data Preparation

To adequately utilize all the data available, the tables containing the EHR data outlined in Table 2 were joined using the unique encounter id that linked all the tables together in a relational manner. There was a one-to-many relationship between the unique patient id and the encounter id, meaning that it is possible for multiple iterations of a single patient's information to be present in the dataset. The central table onto which we planned to join all of the other tables was the "Encounters" table, which houses all patients and visit metadata. Some fields from this table were gender, age, reason for visit, and more. After dropping columns that lack diagnostic properties (insurance provider, location, etc.), from the encounters table we then moved onto merging the other tables using the unique encounter id as the main key.

The vitals table is comprised of commonly taken information from when the patients entered the facility such as temperature, weight, height, body mass index, and blood pressure. It was possible for patients to have multiple vital measurements taken within a single visit. When multiple measurements occurred, we took the median of the continuous vital measurements to "flatten" the vitals to a one-one relationship between the unique encounter and the vital information. After this process was completed, the vitals table was joined onto the encounters table.

The assessments table contains the doctor's "free form" notes about the patient. This table is a unique perspective that the study offers because the opportunity to leverage the text data can allow for relationships that humans understand, but previous research has not used because much of its content is not necessarily measurable and tends to be more subjective in nature. In order for Machine Learning models to ingest unstructured text data, the "bag of words" must be transformed into a matrix of continuous features. To achieve this, all words were stemmed and converted to lowercase. Following, "stop words", words that do not add to the context of a physician note such as: "a", "and", "the", "he"," she" and other articles and pronouns were removed. Next, the remaining nouns and verbs after stop word removal were then put into a Term Frequency – Inverse Document Frequency (TF-IDF) which created a dense matrix of continuous representations for each unique word remaining in the corpus. Due to the excessive size of this table, the dense matrix was used to create engineered features by utilization of k-means clustering. The clusters created logical, separate bins to group various text fields to a higher-level summary in a much more dynamic way. Once the clusters were created, each text derived feature was one-hot encoded to generate a sparse, yet reduced representation of the original matrix. The resultant structure was joined back onto the original encounters table for use in modeling.

In addition to k-means clustering, Latent Derelict Allocation topic modeling for physician notes in the assessments table was also performed. Data was cleaned using the same aforementioned process as the k-means clustering. Tokens were then formed into trigrams, or, three-word chunks, to identify the most common three-word topics. The trigrams were then loaded into a Term Frequency (TF) matrix. Once the topics were generated, each record in the assessments table was labeled with its corresponding topic, then encoded, and joined to the encounters table using the unique encounter ID.

The labs table contains patient lab information, according to the tests ordered by the patient's physician. The lab information constitutes a variety of features that required different forms of feature engineering to tidy data in a format that would work well with a Machine Learning model. When concentration ranges were given, the median of that range was used; if a maximum or minimum range was given, the high or low cut-score was used. Besides continuous lab results, some labs contain information of a more categorical nature. Among these are results of "normal" or "abnormal". Utilizing these specific labs required one-hot encoding of the potential values available for that lab. The continuous types of lab result were found to have much less feature importance than the abnormal lab work ind. Many of the labs with a categorical response were observed to have non-significant feature importance and were not utilized in the final dataset. After reduction all labs were joined onto the encounters table.

The medication table is comprised of the entire medication history for each unique patient. In order to ensure data leaks are not present, medications needed to be filtered using the dates provided in both the medication table and the encounters table. This ensured that only current medications up until the encounter were used in the final table. Once the "current" medications for that individual patient and encounter were subset, the output was joined onto the encounters table using encounter id.

The CPT codes table is comprised of various procedures performed during the patients' visit for billing purposes. It was suggested that use of the procedures in the dataset may help determine if other ailments could be used for screening purposes. Several diseases (diabetes, hypertension, depression) are known comorbidities and occur more frequently for individuals with neurocognitive disease. The CPT description was joined back onto the encounters table using the encounter id.

The ICD-10 table is comprised of documented diagnoses for a patient. In all, there were 4,753 unique ICD-10 codes with the data set. This caused excessive addition of features when one-hot encoding. To reduce feature space, we used lower granularity representation of the ICD-10 code by mapping to a Clinical Classifications Software Refined (CCSR) code which reduced the dimensionality of ICD codes from 4,753 to just 415 unique codes [26]. To prevent target leakage, records containing the NVS011 CCSR code, a group of 25 ICD-10 codes specific to neurodegenerative diseases, were removed. Once the feature reduction was complete, the ICD-10 table was joined back onto the encounters table.

4.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) has identified trends showing strong correlation in several features. Age is the number one factor, and greatly increases the likelihood of dementia. The median age for those diagnosed with neurocognitive disorders is almost 10 years older than those without the diagnosis.

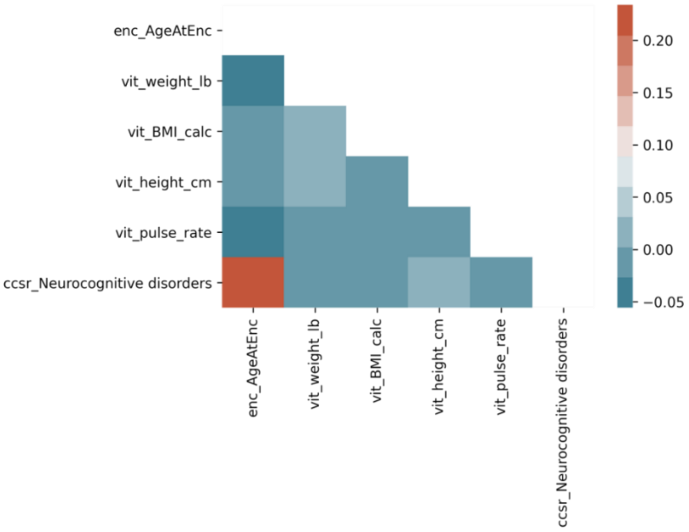

Figure 6: Correlation matrix of Age, Weight, BMI, height, pulse, and neurocognitive disorder.

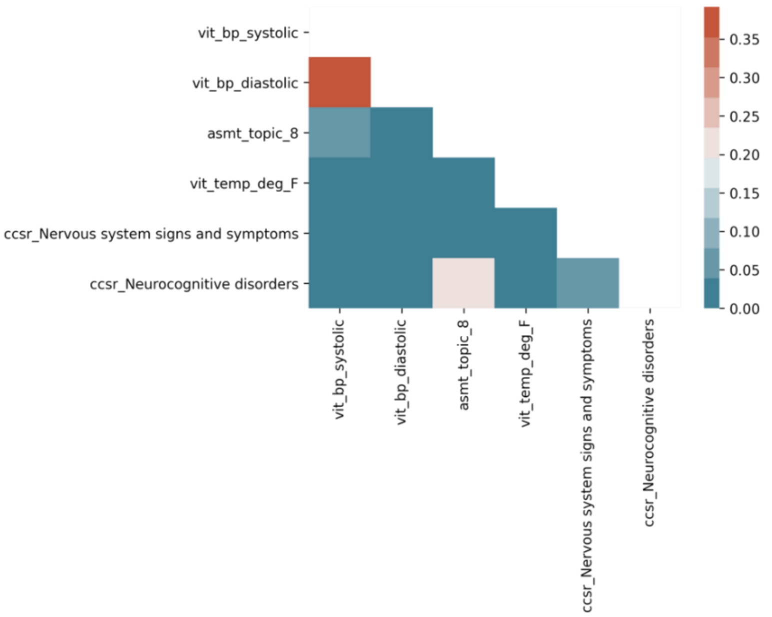

Figure 7: Correlation matrix of systolic and diastolic blood pressure, assessment topic 8, temperature (F), diagnosis of nervous system signs and symptoms, and neurocognitive disorders.

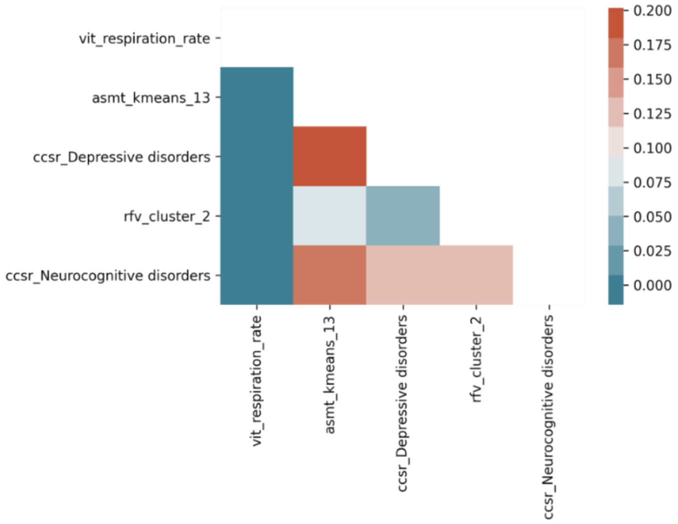

Figure 8: Correlation matrix of respiration rate, assessment group 13, diagnosis of depression, reason for visit group 2, and neurocognitive disorder.

Figure 9: Visualization of age distribution for Alzheimer’s disease and Dementia patients.

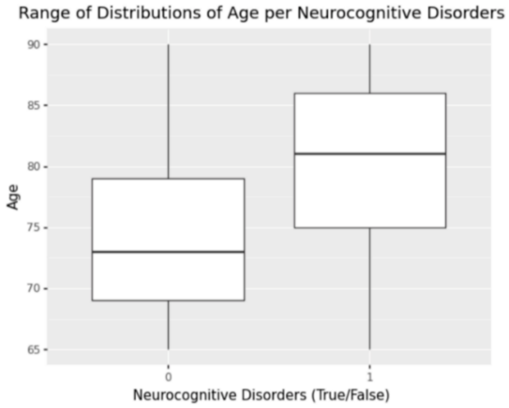

Figure 9: Visualization of median age differences between patients with/without Alzheimer’s disease and Dementia patients.

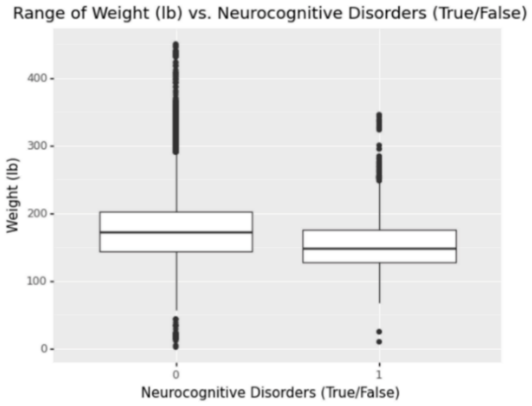

Figure 10: Visualization of median weight distribution for Alzheimer’s disease and Dementia patients.

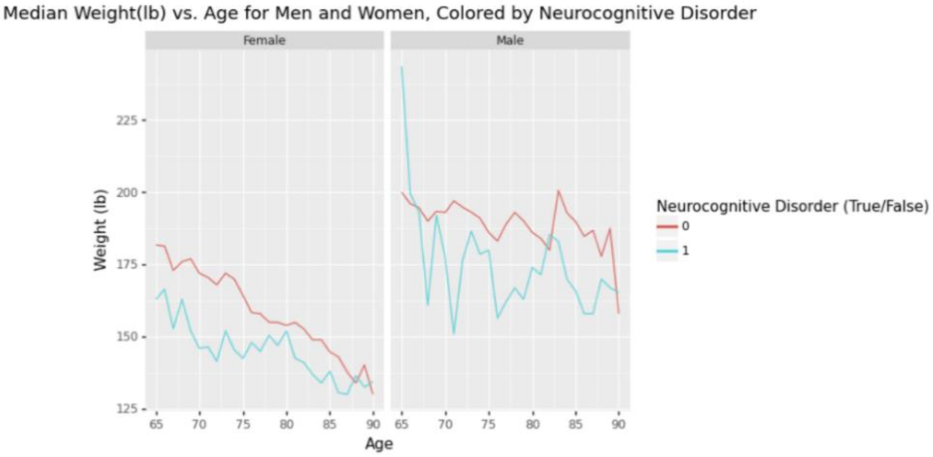

Figure 11: Visualization of weight vs age by gender for each patient type
The median age for those diagnosed with neurocognitive disorders is almost 10 years older than those without the diagnosis.

Research also shows that Alzheimer's disease and Dementia patients are typically underweight compared to their healthy peers. Figures 10-12 display compelling evidence that weight could be a factor in predicting neurocognitive disorders.

Alzheimer's disease and Dementia patients have a higher percentage of depression than healthy peers in the sample population. Figure 14 exemplifies our finding support research by Johnson et al [13].

Features identified from EDA turned out to be influential within each Machine Learning algorithm. Random Forest and Logistic Regression reported weights strongly associated with each of these features as shown in the results section.

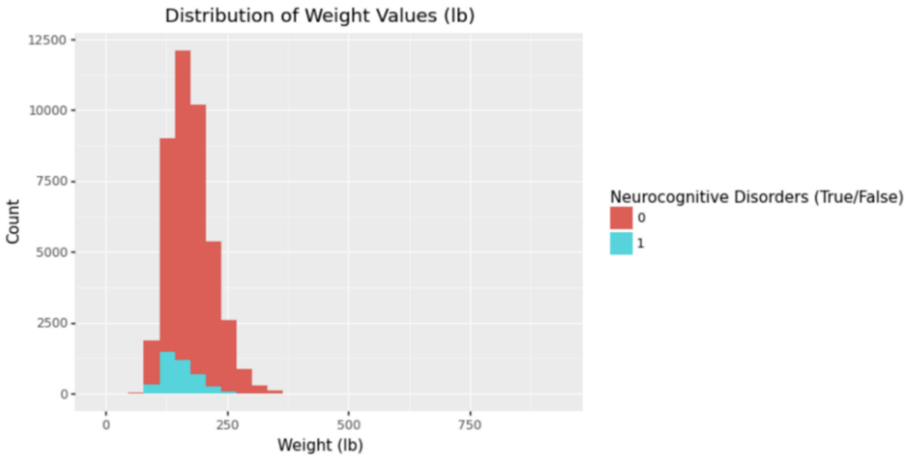

Figure 12: Visualization of weight distribution for Alzheimer's disease and Dementia patients.

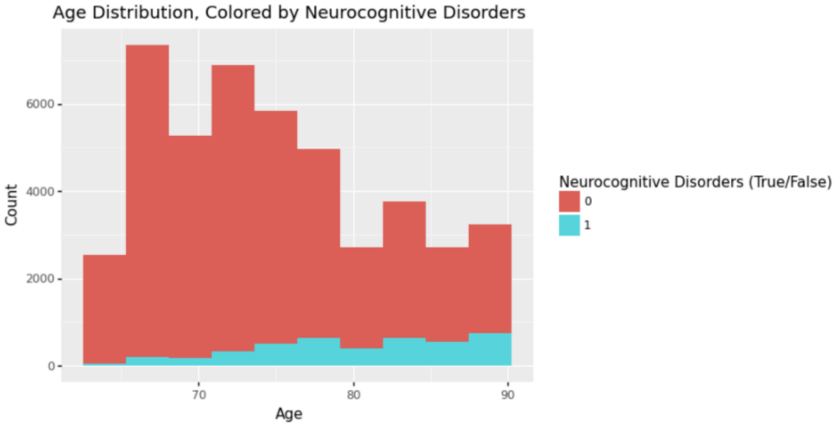

Figure 13: Visualization of age distribution for Alzheimer's disease and Dementia patients.

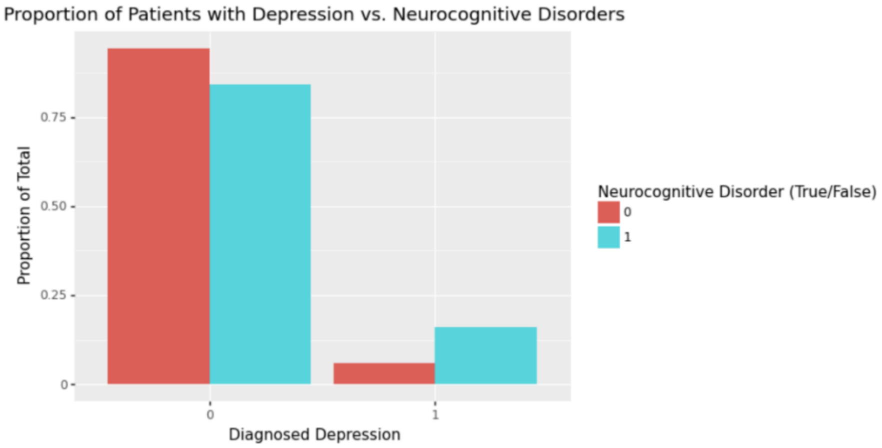

Figure 14: Visualization of the effect of depression on each patient type. The features identified in this EDA turned out to be influential features within each Machine Learning algorithm. Random Forest and Logistic Regression reported weights strongly associated with each of these features as shown in the results section.

5 Results

To build the initial model encompassing all years, the entire dataset was used in a 3-fold cross validation using stratified splits to maintain target ratios. Initial model building started with a stepwise selection approach using all 1,296 predictors available in the data set. All features were investigated and scrubbed of any information that could leak the response variable Alzheimer's disease or Dementia such as NVS011 CSSR codes and text fields that had direct indication or mention of the disease. Text features were reviewed for response variable before and after NLP processing. After obtaining coefficients for all features, feature reduction was applied, narrowing down coefficients to a smaller subset of 150 of the features with highest correlation to reduce overfitting and model complexity.

Multiple sets of data were created from patient encounters by grouping patients into their first year, second year, third year, fourth year and beyond visits, and all encounter occurrences. The intent of sub-setting by annual visit number allows for clinicians to stratify patients into the most logical bin for rule out of Alzheimer's disease or Dementia. Additionally, inclusion of an all-data superset allows for model comparison against yearly data. Each year's subset was then split into a 70/30 set in which the training set was up-sampled to improve model performance by ensuring the model is not biased towards the most common occurrence (individuals without Alzheimer's disease or Dementia). Each model for each training set was fitted using cross validation, and features were manually reduced using highly correlated medication and CPT code data resulting in 661 total features. The best model fit resulting from the cross validation was then measured using the hold out set.

After performing an exhaustive grid search on Logistic Regression, Random Forest, and SVM, the best models for their respective subset are listed in Table 3.

Table 3: Best model type by year and best model for all data.

SUBSET	YEAR 1	YEAR 2	YEAR 3	YEAR 4+	ALL DATA
MODEL	Random Forest	Random Forest	Random Forest	SVM	AdaBoost
NPV	0.9798	0.9741	0.9891	0.9635	0.9886
ACCURACY	0.8424	0.8352	0.8710	0.8754	0.9081
SPECIFICITY	0.8511	0.8480	0.8759	0.8997	0.9167
SENSITIVITY	0.2060	0.1948	0.1828	0.3071	0.1235
AUC	0.5929	0.5845	0.5860	0.6352	0.5560

The best performing model was AdaBoost on all data. It returned the highest negative predictive value, accuracy, and specificity for all models and all years. The next highest performing model was the year 3 Random Forest model, followed by the year 1 Random Forest model. All models in Table 3 show strong negative predictive value, accuracy, and specificity but are not as strong in sensitivity or AUC metrics.

Negative predictive values and specificity are metrics that identify a model's ability to correctly detect negative values. Negative predictive value (NPV) is a ratio between true negatives and true negatives plus false negatives, while specificity is a ratio of true negatives over true negatives plus false positives. Since the test set is representative of the whole dataset, many of the records are negative for Alzheimer's disease or Dementia. The models for each year show an exceptional ability to detect true negatives in the data.

Sensitivity is the ability for a model to detect true positives for all classified positive cases. All models performed poorly in this area as the scores were not able to surpass 0.2060. The sensitivity scores were low due to preferential fitting towards the highest NPV possible in order to reduce unnecessary testing.

Correctly framing the "context of use" for the rule out screening is critical for determining success. Models investigated here have an exceptional ability to screen patients and identify those who does not have Alzheimer's disease or Dementia with an NPV between 97.78 and 96.35% over a 4-year time span. The median number of doctor visits for all patients during this time span was 6-7 visits; indicating the rule out screen would likely have been able to screen all patients 65 years and older, who saw their doctor.

Sensitivity and specificity metrics are usually associated with gold standard tests (CSF and PET). Performance characteristics for rule out screening, such as NPV and PPV, are used to determine how well predictions were made considering prevalence.

The rule out screening demonstrated very accurate detection rate of 75% for individuals that truly do not have disease (NPV ≈ 98.86% overall). Unfortunately, after accounting for the prevalence rate, the rule out screen has a high false positive rate and does not reliably identify Dementia and Alzheimer's disease in the overall population. Rule out screening only correctly identified ≈ 15% of patients with Dementia and Alzheimer's disease (PPV ≈ 18 %). Future comparison of rule out screening to gold standards tests would help to support findings from this study.

Natural Language Processing generated several clusters that ranked high in Random Forest feature importance. Out of nearly 1,200 features, 13 NLP-engineered features were in the top 50, and 3 were in the top 15. The use of NLP features with traditional Machine Learning techniques has been infrequently mentioned in previous research and shows promising results as they contribute to models in a significant manner.

Accuracy scores for all models had values ranging from 0.8352 to 0.9081. Accuracy is a ratio of true positives and true negatives over the total samples. Since negative values account for over 85% of the data much of the accuracy score is due to each model's strong ability to detect negative cases. However, each model's poor performance in detecting positive cases prevents the accuracy score from being stronger and closer to the model NPV.

AUC also performed poorly for all models as AUC scores depend on a balance of specificity and sensitivity scores. Due to each model's poor performance in sensitivity, the resulting AUC scores are also low.

Table 4: Random Forest Feature Importance for All Data Model – Top 15 Features.

Features	*Gini Importance*
enc_AgeAtEnc	0.09897
vit_weight_lb	0.05520
vit_BMI_calc	0.04997
vit_height_cm	0.03457
vit_pulse_rate	0.03366
vit_bp_systolic	0.03214
vit_bp_diastolic	0.03214
asmt_topic_8	0.02968
vit_temp_deg_F	0.02867
ccsr_Nervous system signs & symptoms	0.02048
vit_respiration_rate	0.01946
visit_number	0.01274
asmt_kmeans_13	0.01043
ccsr_Depressive disorders	0.00927
rfv_cluster_2	0.00866

Random Forest can also be used to determine features that contribute most to model, or feature importance, using Gini Importance measures in which more important features have larger values and less important values have smaller values. Gini importance is calculated by adding decreases in node impurity and averaging the differences for all trees. Using these values, the top n features can be subset and used to trim insignificant features from the model in order to avoid overfitting. Feature importance for both future modeling and model understanding was derived by fitting the best performing Random Forest model on the whole data set. Once the model was fit, the feature importance attribute was applied to the Random Forest object, where a data frame was generated and sorted.

The NLP features took 11 out of the top 25 most important features using Recursive Feature Elimination with the Random Forest algorithm, indicating the data within doctors' assessments is useful in detecting the disease. Several noticeable features identified as significant by Random Forest feature importance were NLP-engineered features: asmt_topic_8, asmt_kmeans_13, and rfv_cluster_2. Physician notes were used to generate asmt_topic_8 and asmt_kmeans_13 and patient's reported reason for visit was used for topic rfv_cluster_2. The NLP features were named based upon the table they originated from and the topic or k-means cluster they were placed in, respectively.

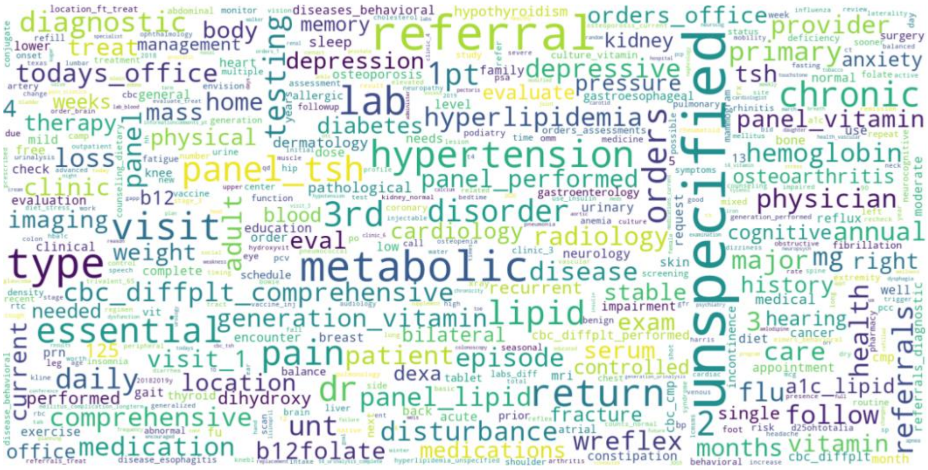

Figure 15: Visualization of word frequency within asmt_kmeans_13 cluster Larger text indicates more frequency.

Rfv_cluster_2 included primarily follow-up appointment visits, consisting of numeric week follow-up frequencies, physical therapy appointments, and prescription refill timelines among other reasons. The top NLP-engineered feature, asmi_topic_8, was generated utilizing Latent Derelict Allocation topic modeling.

Table 5: asmt_topic_8 – LDA derived tri-gram phrases provide insight on the contents of this topic.

Feature	Tri-gram Phrases
asmt_topic_8	10_mg_daily, metformin_1000_mg, type_2_diabetes, 2_diabetes_mellitus, encounter_medication_review, low_blood_sugar, per_ada_guidelines, medication_review_counseling, mg_daily_continue, essential_primary_hypertension

The topics provided by this cluster, along with the asmt_kmeans_13 cluster, show trigrams that are strongly associated with Alzheimer's disease and Dementia as co-morbidities. The direct mention of diabetes mellitus, diabetes, low blood sugar, or drugs like metformin used to treat diabetes were important in generating the NLP feature. Additionally, hypertension was mentioned in this cluster demonstrating a strong association between the contents of this cluster, and Alzheimer's disease and Dementia.

6 Ethics

Three ethical areas were identified and were required to be maintained throughout the course of study.

The first ethical standard was regulation of data storage and controlled access to Patient Health Information (PHI). Access to the EHR records required approval from the UNT HSC IRB - Project Title#: 1660705-1, Reference#: 2020-11 and compliance to the Healthcare Insurance Portability and Accountability Act (HIPAA) required all EHR records were scrubbed of PHI and anonymized to prevent identification of patients. User access was provided by UNT HSC IT (Information and Technology) security, VPNs and unique login credentials to a private server were issued. The secure private server ensured ethical handling and storage of the data and provided an ideal location for cleaning, manipulating, modeling, and visualizing data and results within the UNT HSC domain.

The second ethical standard was related to the Machine Learning process. The use of powerful computers and large data sets combined with algorithms can solve complex problems not directly visible to humans. This research has shown promise in utilizing opportunities previously untapped. The ethical conflicts that arise from Machine Learning algorithms are lack of transparency, reproducibility, ethics, and effectiveness. The article "*Machine Learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness*" was used as framework of questions that were directly investigated during the creation of the Alzheimer's disease and Dementia rule out screening [19]. Throughout this study the following 6 topics were repeatedly considered and evaluated: Inception, Study, Statistical methods, Reproducibility, Impact evaluation, Implementation.

The third ethical standard was related to the current lack of predictive screening creating a situation where patients go unscreened or diagnosed until it is too late. The Alzheimer's disease and Dementia rule out screen reports metrics showing real benefit could be provided to patients 65 and older when implemented. Ethically, it is in the best interest of public health to continue this research further defining limitations and implications. Some statistical concerns surrounding low PPV, false negatives and false positives are present. Test results need to be clearly explained to patients and doctors, so the correct interpretations are easily understood. The rule out screen only indicates no disease is present. Ethically, the interpretation of the Alzheimer's disease and Dementia rule out screen can be simply stated as: the rule out screen can only detect absence of disease; for all other findings, patients are recommended to seek additional medical care and possible escalation to CSF or PET screening.

7 Discussion

Rule out screening quickly provides patients with information about their neurocognitive health. This is vastly different than current assessment practices where no predictive information is available. This new paradigm allows for patients to act based on information provided from the rule out screen.

Gold standard screening methods historically fall short when a patient or family member has memory concerns. If they are interested in diagnosis/treatment, there few alternatives. Screening currently consists of non-predictive options and typically costs between one thousand and several thousand dollars. Traditional screening methods also suffer from a national shortage of Neuropsychologists, which may introduce several months delay to neurological assessments. The shortcomings of gold standard screenings are further exemplified when more than half of patients go undiagnosed.

Rule out screening can be done every visit reliably, letting patients know they do not have any disease with very high certainty. Conversely, rule out screening cannot accurately confirm patients are truly positive for the disease. The test only provides patients that do not screen negative with the strong recommendation to consult a doctor for additional, higher-level testing. Multi-tiered approaches have successfully been used; most recently in 2020 rapid IgM/IgG testing to assess Coronavirus disease in emergency rooms 1-2 days before gold standard screening [30]. Breast cancer detection highlights the role of mammography for early-stage detection and then relies on biomarker and MRI when cases are escalated [31].

This research has extended previous Machine Learning approaches that aim to facilitate detection and screening of Alzheimer's disease and Dementia by utilizing NLP applied to doctors' notes. There is substantial evidence that NLP has positively improved the prediction performance in screening for Alzheimer's disease and Dementia. Identifying phrases that contribute towards feature importance score can be accomplished using term frequencies and word clouds. Several key phrases associated with Alzheimer’s disease and Dementia like hypertension, depressive, depression, and disturbance are known co-morbidities.

In addition to the NLP features, several previously discovered correlated features such as: age, blood pressure, under-weight, and blood sugar levels were further validated as significant in this research. Utilizing more complex algorithms in the NLP space will likely yield marginally better performance, and evidence suggests that future work may benefit from a deeper application of these techniques and the use of a medical specific lexicon.

Despite NLP-engineered features contributing towards a large portion of model predictive ability improvements for detections of positive cases is needed to address poorly performing positive predictive power, sensitivity, and AUC. This suggests that more future engineering specific to positive cases needs to be done, or incorporation of more features used to help diagnose Alzheimer’s disease or Dementia like blood biomarkers, or cognitive tests should be included in future work.

Other improvements can be made to data storage and models by migrating data to a graph data base. Data for this research was relational in nature and during processing data required aggregation of fields to flatten data to an encounter basis. In doing so a record was able to provide information on whether a feature occurred but was no longer able to provide information about the order in which features occurred within an encounter. Utilizing a graph database allows for these relationships to be preserved in a form that can be utilized by models and could be advantageous in future iterations of this work. Additional studies should be done that independently rule out Dementia and Alzheimer’s disease separately. Stratification by disease, generation of new features space and creation of new models which are better tuned to each disease should produce improved performance metrics.

Any person over 65 or with a family with elderly members needs to consider the current state of healthcare: the lack of predictive tests, overly expensive gold standard testing that sadly ends with a terminal disease which has been historically underdiagnosed for years. Considering the previous statement, it only makes sense that rule out screening for undiagnosed Alzheimer’s disease and Dementia using an EHR based machine learning be considered as the first line of defense in the fight against these diseases. Ethically, the test does not put patients at-risk because a diagnosis is not given. Only the absence of disease is determined, therefore patients with a non-negative result would need to seek additional care for effective diagnosis.

8 Conclusion

The social and economic impact of Alzheimer’s disease and Dementia are incredible compared to heart disease or even cancer. Considering only limited treatment options are available for neurodegenerative disease being able to effectively use health care records as an early rule out tool has significant implications as well as cost-savings. Rule out screening described within this study uses health care systems already in place that more efficiently guide patient care and treatment.

Rule out screen has the potential to save billions of dollars for patients, healthcare, and insurance companies. In 2020, the Alzheimer's Association reported staggering numbers for payments made by Medicare and Medicaid combined with other payments total nearly 300 billion dollars for the 5.8 million people age 65 and older living with Alzheimer's disease and Dementia. The rule out screen as it is currently configured today could provide significant improvement to families and patients by providing knowledge where none has existed. If a gold standard testing available today screened everyone age 65 and older tomorrow, there would be millions of patients that would need to be screened. In 2020, 16.9 percent of the U.S. population were 65 and older; that equates to 55.9 million seniors. Screening this entire elderly population would cost somewhere between 55 to 385 billion dollars depending which gold standard test was used. If the Alzheimer's disease and Dementia rule out screening was used as the first line screening tool the number of gold standards tests could be reduced 75 percent. The estimate savings would be 41.2 to 288.8 billion dollars compared to only using gold standard screenings. The rule out screening described with this study will considerably reduce screening costs, increase accessibility, improve healthcare options for our elderly population. It is important to acknowledge that the findings of this research have shown great potential for use of NLP in detection of Alzheimer's disease; however, more research should be performed until conclusive decisions can be made as to whether an individual has Alzheimer's disease.

Acknowledgments. Jacquelyn Cheun, Ph.D.$_1$, and Faizan Javed, Ph.D.$_1$, – Capstone Professor

References

1. Deborah E Barnes, Jing Zhou, Rod L Walker, Eric B Larson, Sei J Lee,W John Boscardin, Zachary A Marcum, and Sascha Dublin. Development and validation of eradar: A tool using ehr data to detect unrecognized dementia. Journal of the American Geriatrics Society, 68(1):103–111, 2020.
2. Randall J Bateman, Chengjie Xiong, Tammie LS Benzinger, Anne M Fagan, Alison Goate, Nick C Fox, Daniel S Marcus, Nigel J Cairns, Xianyun Xie, Tyler M Blazey, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med, 367:795–804, 2012.
3. Heiko Braak, Dietmar R Thal, Estifanos Ghebremedhin, and Kelly Del Tredici. Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. Journal of Neuropathology & Experimental Neurology, 70(11):960–969, 2011.
4. Gandhi, R. (2018, July 05). Support Vector Machine - introduction to Machine Learning algorithms. Retrieved March 30, 2021, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
5. Joseph E Gaugler, Lisa J Bain, Lauren Mitchell, Jessica Finlay, Sam Fazio, Eric Jutkowitz, Sube Banerjee, Kim Butrum, Joseph Gaugler, Laura Gitlin, et al. Reconsidering frameworks of Alzheimer's dementia when assessing psychosocial outcomes. Alzheimer's & Dementia: Translational Research& Clinical Interventions, 5:388–397, 2019.
6. Brian A Gordon, Tyler M Blazey, Yi Su, Amrita Hari-Raj, Aylin Dincer, Shaney Flores, Jon Christensen, Eric McDade, Guoqiao Wang, Chengjie Xiong, et al. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. The Lancet Neurology, 17(3):241–250, 2018.

7. Gupta, P. (2017, November 12). Decision trees in Machine Learning. Retrieved March 30, 2021, from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
8. Peter Harrington. Machine Learning in action. Manning Publications Co.,2012.
9. Liesi E Hebert, Jennifer Weuve, Paul A Scherr, and Denis A Evans. Alzheimer disease in the united states (2010–2050) estimated using the2010 census. Neurology, 80(19):1778–1783, 2013.
10. Michael D Hurd, Paco Martorell, Adeline Delavande, Kathleen J Mullen, and Kenneth M Langa. Monetary costs of dementia in the united states. New England Journal of Medicine, 368(14):1326–1334, 2013.
11. Clifford R Jack Jr, Val J Lowe, Stephen D Weigand, Heather J Wiste, Matthew L Senjem, David S Knopman, Maria M Shiung, Jeffrey L Gunter, Bradley F Boeve, Bradley J Kemp, et al. Serial pib and mri in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. Brain, 132(5):1355–1365, 2009.
12. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning (vol. 112, p. 18), 2013.
13. Leigh A Johnson, Melissa Edwards, Adriana Gamboa, James Hall, Michelle Robinson, and Sid E O'Bryant. Depression, inflammation, and memory loss among mexican americans: analysis of the hable cohort. International Psychogeriatrics, 29(10):1693–1699, 2017.
14. Eric Jutkowitz, Robert L Kane, Joseph E Gaugler, Richard F MacLehose, Bryan Dowd, and Karen M Kuntz. Societal and family lifetime cost of dementia: implications for policy. Journal of the American Geriatrics Society,65(10):2169–2175, 2017.
15. Max Kuhn, Kjell Johnson, et al. Applied predictive modeling, volume 26. Springer, 2013.
16. Jodi Liu, Jakub Hlávka, Richard John Hillestad, and Soeren Mattke. Assessing the preparedness of the US health care system infrastructure for an Alzheimer's treatment. RAND, 2017.
17. Thomas H McCoy Jr, Sheng Yu, Kamber L Hart, Victor M Castro, Han-nah E Brown, James N Rosenquist, Alysa E Doyle, Pieter J Vuijk, Tianxi Cai, and Roy H Perlis. High throughput phenotyping for dimensional psychopathology in Electronic Health Records. Biological psychiatry, 83(12):997–1004, 2018.
18. Thomas H McCoy Jr, Larry Han, Amelia M Pellegrini, Rudolph E Tanzi, Sabina Berretta, and Roy H Perlis. Stratifying risk for dementia onset using large-scale Electronic Health Record data: a retrospective cohort study. Alzheimer's & Dementia, 2019.
19. Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ... & Hemingway, H. (2020). Machine Learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, *368*, 2020.
20. George A Miller. Wordnet-about us. wordnet. princeton university. 2009, 2010.
21. The Office of the National Coordinator for Health Information Technology. 2018 report to congress, annual update on the adoption of a nationwide system for electronic use and exchange of health information, 2018, from https://www.healthit.gov/sites/default/files/page/2018-12/2018-HITECH-report-to-congress.pdf.
22. National Institute on Aging. Basics of Alzheimer's disease and dementia: What is Alzheimer's disease? 2017. URLhttps://www.nia.nih.gov/health/what-alzheimers-disease.
23. Ji Hwan Park, Han Eol Cho, Jong Hun Kim, Melanie M Wall, Yaakov Stern, Hyunsun Lim, Shinjae Yoo, Hyoung Seop Kim, and Jiook Cha. Machine Learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. NPJ digital medicine, 3(1):1–7, 2020.
24. Xiaogang Peng and Ben Choi. Document classifications based on word semantic hierarchies. In Artificial Intelligence and Applications, volume 5, pages 362–367. Citeseer, 2005.

25. Fred Ramsey and Daniel Schafer. The statistical sleuth: a course in methods of data analysis. Cengage Learning, 2012.
26. https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp#overdiagnoses
27. Eric M Reiman, Yakeel T Quiroz, Adam S Fleisher, Kewei Chen, Carlos Velez-Pardo, Marlene Jimenez-Del-Rio, Anne M Fagan, Aarti R Shah, Sergio Alvarez, Andrés Arbelaez, et al. Brain abnormalities in young adults at genetic risk for autosomal dominant Alzheimer's disease: a cross-sectional study. The Lancet. Neurology, 11(12):1048, 2012.
28. Dipanjan Sarkar. Text analytics with python: A practical real-world approach to gaining actionable insights from your data. New York: Apress;2016.
29. Yiu, T. (2019, August 14). Understanding Random Forest. Retrieved March 30, 2021, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2
30. Pulia, M. S., O'Brien, T. P., Hou, P. C., Schuman, A., & Sambursky, R. Multi-tiered screening and diagnosis strategy for COVID-19: a model for sustainable testing capacity in response to pandemic. *Annals of medicine*, *52*(5), 207-214, 2020.
31. Esserman, L. J., Shieh, Y., Park, J. W., & Ozanne, E. M. A role for biomarkers in the screening and diagnosis of breast cancer in younger women. *Expert review of molecular diagnostics*, *7*(5), 533-544, 2007.