Mining the Electronic Health Records for detecting early onset of Alzheimer's Disease

TABLE OF CONTENTS

I. ABSTRACT 3

II. INTRODUCTION 3

III. PROBLEM 4

a. PROBLEM STATEMENT 4

b. MOTIVATION TO STUDY 5

IV. GOALS 5

V. OBJECTIVES 6

VI. LITERATURE REVIEW 6

VII. METHODOLOGY 8

VIII. TECHNOLOGIES 10

IX. SIGNIFICANCE/IMPACT 10

X. TIMELINE 11

XI. DELIVERABLES 11

XII. CONCLUSION 11

XIII. BIBLIOGRAPHY 12

# ABSTRACT

Alzheimer is the most common form of dementia. As the population around the globe is increasing day by day, the number of people having Alzheimer’s disease is also increasing. Even though we are in 21st century, we have been unsuccessful to find complete cure for this disease. Because of this failure to cure Alzheimer’s, the tests related to it are time consuming as well as expensive. Early detection of this disease is critical in order to postpone the necessary symptoms and to prepare both health-care providers and families for patients’ future needs. The current methods that are being used for detection of AD may therefore also benefit from these early markers, which can thereby save time and money. This project involves applying the various data mining techniques like data cleaning, missing value prediction, dimensionality reduction on the Alzheimer's disease Neuroimaging Initiative (ADNI) database for selecting the useful features from a number of features and then classifying the patients into the different stages of dementia. The aim of this study is not only to provide insight into which features qualified for early markers of AD but also highlights the various machine learning algorithms that were used for missing value prediction, feature selection process that eventually helped to classify the patients based on these new early markers.

# INTRODUCTION

Alzheimer’s disease (AD) is usually found among older population. It is mostly expected to begin after the age of 60. Alzheimer's disease is the 6th leading cause of death in the United States. More or less 500,000 individuals die each year in light of the fact that they have Alzheimer’s.An estimated 5.2 million Americans have Alzheimer's disease in 2014[2]. People with dementia may not even be able to perform basic activities like eating or getting dressed up. People may not recognize family members or have trouble speaking, reading or writing. No treatment can stop the disease [3].

This proposal is organized as follows. The problem that I would be addressing is precisely described under the “Problem Statement” section, which you can find, on page 4. The section “Literature Review” identifies the existing research that was significant. It also addresses the strengths and weaknesses of these existing researches. This section will be followed by the “Goals and Objectives” that explains the desired outcomes from my project and how I plan to accomplish these goals. The “Methodology” section describes the entire process right from the data collection phase to the integration of the model with iOS platform, which is followed by the “Technologies” section that enlists the technologies that have been used in the entire process. Finally, the proposal concludes by detailing the significance of this study under the “Significance” section and also provides the necessary timeline for the tasks under the “Timeline” section.

# PROBLEM

## PROBLEM STATEMENT

Part A: The Ideal

Alzheimer’s disease (AD) is the most common form of dementia. As such there is no treatment to stop or cure this disease. A number of tests are associated with the patient’s who begin complaining about memory issues [4]. Data from patients’ EHRs is thus expert-filtered and is a rich source of medical information that highlights a patient's status over time. The inability to cure these conditions results in prolonged and expensive medical care. Early detection is critical that would potentially postpone symptoms and could prepare both health-care providers and families for subjects' future needs or care.

Part B: Reality

Presently we do not have accurate early indicators or tools, which can be used to flag patients who are at risk of developing dementia, prior to the disease’s ultimate diagnosis [5]. The cognitive tests involved in the process are not only costly in terms of time & money but also tiresome for the patients, as it requires years of monitoring. Hence it's not feasible for all the patients. Some of the tests are not conducted until the patient starts showing the symptoms of Alzheimer’s or sometimes they might be even conducted when the patient has already transitioned to the final stage of dementia.

Part C: The Consequences

Indeed, there is a need for a computational automated model to predict Alzheimer’s. My proposal will therefore be focusing on building this automated screening system based on the patients’ EHR’s data that could be used to identify and flag potentially problematic patients only for further assessment by clinicians, eventually saving time and money. Additionally, identification of useful features from classification experiments may improve understanding of important markers in early dementia and Alzheimer's Disease detection that would help postpone the necessary symptoms and to prepare both health-care providers and families for patients’ future needs. The current methods that are being used for detection of AD may also benefit from these early markers, which can thereby save time and money of patients.

## MOTIVATION TO STUDY

It’s true that as such there is no treatment for Alzheimer’s disease, so are there ways to detect this disease at an early stage so that proper care can be taken with regards to the Alzheimer’s patient?

There are numerous tests associated with Alzheimer’s disease like medical history, physical exam, neurological exam, mental status tests and brain imaging. A lot of data is being warehoused over a period of time related to these tests. Analysis of these tests can be costly in terms of time and money. Therefore, what if we can explore certain features in a machine-learning context, which can help us identify patients' dementia status? [4]

# GOALS

Goal 1:

The primary goal of my capstone project is to build a model with the help of potential early indicators/ markers of Alzheimer’s that could flag only the potentially problematic patients for further assessment by clinicians, eventually saving time and money of the patients and caretakers.

Goal 2:

Finally, this proposal also aims at developing an iOS application based on this computational, automated model that can predict Alzheimer’s.

# OBJECTIVES

Objectives for Goal 1:

1. Identify the source of data such that the data should be sufficient enough for analysis.
2. If required preprocess the data obtained from step 1 using data mining techniques.
3. Build a model with the help of machine learning algorithms. To compare various algorithms and select the one with highest accuracy.

Objectives for Goal 2:

1. Decide whether the application will be a web application or iOS based.
2. Identify methods to integrate the model obtained from Goal 1 with the selected platform for the application.
3. Assess the usefulness of this application; provide a demo of the application.

# LITERATURE REVIEW

Over a period of time, the electronic health records have been well acknowledged by the society for data mining and machine learning techniques as these EHR’s provide high-dimensional, multivariate data sets for research purpose [6]. Notably, deducing some beneficial neurodegenerative patterns from this humongous structured data collection has been the basic approach in the prediction of dementia. Till date, many data mining methodologies and machine learning algorithms have been proposed for early diagnosis of Alzheimer’s disease.

David B. Hogan and Erika M. Ebly in the year 2000 experimented in a cohort of Canadian Seniors to predict who will develop dementia. Although they determined that the basic three variables viz. Mini Mental State Exam (MMSE), age and report of the presence of memory problems could be useful in the prediction of dementia, the specificity measure was very low, only 56%. Moreover, the tests in this paper failed to consider laboratory, radiological, and neuropsychological data and were not focused on a particular type of dementia [7]. Similar experiments were performed in conjunction with other possible predictors like neuropsychological markers that are specified in [8,11] which resulted in better specificity and sensitivity than simply using age and MMSE for prediction.

In [10], Gomar et al. aimed at using ADNI database for determining the effects of cognitive markers, biomarkers and other markers on the conversions from Mild Cognitive Impairment (MCI) to Alzheimer's disease (AD) over a period of 2 years. An interesting outcome of applying a series of logistic regression models along with size analysis was that the Functional Assessment Questionnaire itself accounted for about 50% of the prediction of conversion from mild cognitive impairment to Alzheimer disease, thus concluding that the cognitive markers were strong predictors of conversion than rest of the markers and biomarkers. But these cognitive tests are not only costly in terms of time & money consuming but also tiresome for the patients, as it requires years of monitoring. Hence it's not feasible for all the patients. Biomarkers from cerebrospinal fluid (CSF), as well as brain volume measurements from magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) together have been useful in predicting conversion from Mild Cognitive Impairment (MCI) to Alzheimer's disease (AD) within the ADNI dataset thus reducing the misclassification rate from 41.3% to 28.4% as per [12]. However, such tests are not typically conducted until the patient starts showing the symptoms of MCI or sometimes they may be even conducted when the patient has already transitioned to the final stage of dementia. D. AL-Dlaeen and A. Alashqur in [13] describes the methodology of data mining technique called classification used for training a model that can predict whether a patient has AD or not. The classification tree that was obtained was based only on 17 different samples comprising of attributes like gender, age, genetic causes, brain injury, and vascular disease.

My capstone will focus the classification of Alzheimer’s patient based on both versions: one including the cognitive markers and second without the cognitive markers. It also intends to use a number of useful numerical attributes related to MRI, CSF, PET scans and then building a model by comparing and choosing the best classifier amid a set of classifiers like KNN, SVM, Decision Tree, RandomForest, LDA, QDA that can help predicting Alzheimer’s more accurately which are missing in papers [9,13].

# METHODOLOGY

To meet the drafted objectives, I plan to follow the stages, which are usually involved in any Data Mining process.

Figure 1: The process

*Objective 1: To identify the source of data such that the data should be sufficient enough for analysis.*

1. Data Collection

Legally request the Alzheimer Disease Neuroimaging Initiative (ADNI) for the datasets that includes information regarding the tests conducted on the Alzheimer patients. Carefully read the conditions documented in ADNI Data Use Agreement [14].

*Objective 2: If required preprocess the data obtained from step 1 using data mining techniques.*

1. Dataset Preparation

It may happen that data related to a patient may be scattered in different datasets. To perform an analysis, it is necessary to have all the data in one file, which may be possible by merging various datasets based on the patients’ id and on the month of their visits’.

1. Data Pre-processing

Data pre-processing is an important stage in the data mining process.

* 1. Missing Value Prediction

During the data preparation phase it may happen that final dataset obtained after merging the datasets may contain irrelevant or missing data. Depending upon the percentage of missing data either use the mean method if the percentage is small else use the imputation algorithm available in Amelia II package in R [15] to find these missing values.

* 1. Feature Selection

Try to find only the important features from a set of features by using feature selection algorithms like L1 Logistic Regression (Lasso). Tune the L1 algorithm with different values of C. Create a different dataset with these selected attributes, which will be the early indicators of Alzheimer’s.

*Objective 3: Build a model with the help of machine learning algorithms. Compare various algorithms and select the one with highest accuracy.*

1. Data Classification

Using Scikit-learn in python and the selected features from step III b, choose the classifier amid a set of classifiers like KNN, SVM, Decision Tree, RandomForest, LDA, QDA, etc that provides highest accuracy with the 10-fold cross validation. Check accuracies on training and testing datasets.

1. Comparison of Results

Compare the accuracy, precision and recall of both the datasets one including the cognitive markers and second without the cognitive markers.

*Objective 4: Identify methods to integrate the model obtained from Goal 1 with the selected platform for the application*

1. Integration of model with iOS

With the help of Xcode and Objective C programming language try to create an iOS application that will accept values for the selected features in step III b, and outputs that individual’s Alzheimer stage status.

# TECHNOLOGIES

* **R, R Studio**: These technologies would be used for data preparation and missing value prediction.
* **scikit learn, Spyder**: These technologies would be used for Feature selection and data classification process.
* **Xcode and related frameworks**: These would be used for integration of the model with iOS.

# SIGNIFICANCE/IMPACT

Once we have this automated screening system, it can be used to identify and flag potentially problematic patients only for further assessment by clinicians, eventually saving time and money. Additionally, identification of useful features from classification experiments may improve understanding of important markers in early dementia and Alzheimer's Disease detection that would help postpone the necessary symptoms and to prepare both health-care providers and families for patients’ future needs. The current methods that are being used for detection of AD may also benefit from these early markers, which can thereby save time and money of patients.

# TIMELINE

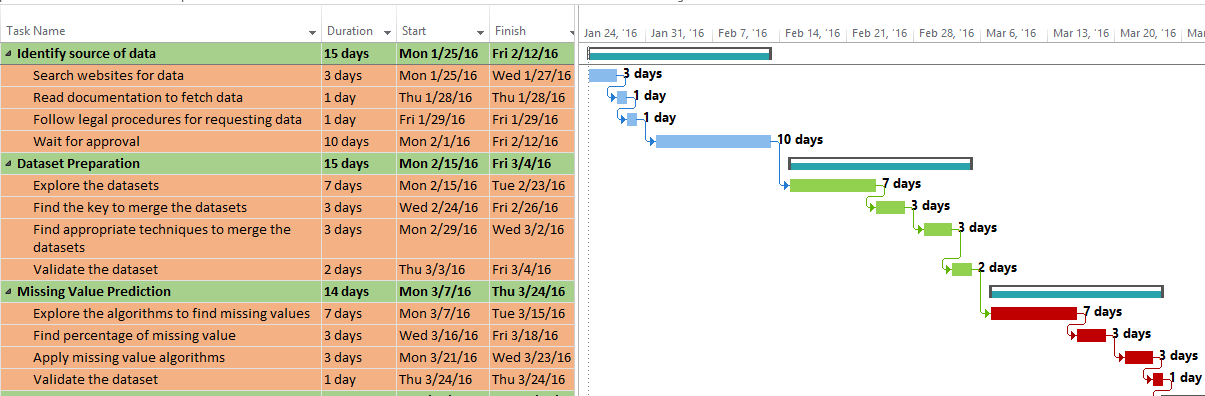


Figure 2(a): Timeline

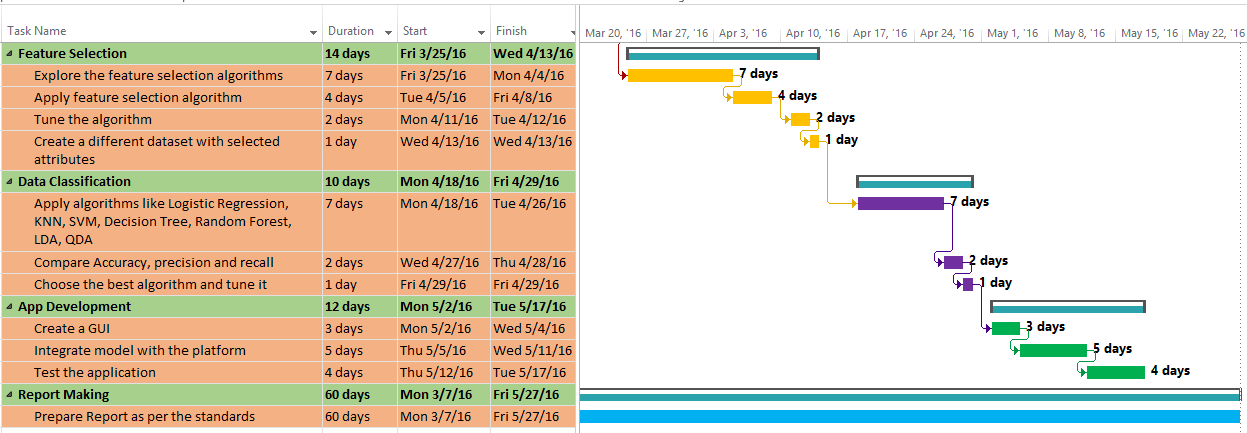


Figure 2(b): Timeline

# DELIVERABLES

At the end of data classification phase, a model that has been well trained using machine learning and data mining techniques will be delivered that can be then used to integrate with the iOS platform. An application will be developed for iOS that will use this model to predict the Alzheimer’s state of a particular subject after the subject inputs the readings from the report.

# CONCLUSION

This proposal aimed at exploring the structured features of subject records by using some machine learning approaches and data mining techniques. The subjects’ data used in this proposal was collected as a part of Alzheimer’s Disease Neuroimaging Initiative (ADNI). By identifying the potential early markers, it is possible to build a machine learning model that can act as an automated screening system to identify and flag potentially problematic patients only for further assessment by clinicians, eventually saving time and money of the patients and their caretakers.

# BIBLIOGRAPHY

[1] Bullard, J., Murde, R., Yu, Q., Alm, C. O., & Proano, R. Inference from Structured and Unstructured Electronic Medical Data for Dementia Detection.

[2] Alzheimer's Association. 2014 Alzheimer's Disease Facts and Figures. Alzheimer's & Dementia, 10, 2014.

[3] Alzheimer's Disease. (2014, September 12). Retrieved September 14, 2014, from <http://www.nlm.nih.gov/medlineplus/alzheimersdisease.html>

[4] Tests for Alzheimer's Disease and Dementia. (2014). Retrieved September 21, 2014, from <http://www.alz.org/alzheimers_disease_steps_to_diagnosis.asp>

[5] “Alzheimer’s & Dementia Testing Advances | Research Center,” *Alzheimer’s Association*. [Online]. Available: <http://www.alz.org/research/science/earlier_alzheimers_diagnosis.asp>.

[Accessed: 20-Oct-2014].

[6] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” Nat Rev Genet, vol. 13, no. 6, pp. 395–405, Jun. 2012.

[7] D. B. Hogan and E. M. Ebly, “Predicting Who Will Develop Dementia in a Cohort of Canadian Seniors,” The Canadian Journal of Neurological Sciences, vol. 27, no. 1, pp. 18–24, Feb. 2000.

[8] D. P. Devanand, X. Liu, M. H. Tabert, G. Pradhaban, K. Cuasay, K. Bell, M. J. de Leon, R. L. Doty, Y. Stern, and G. H. Pelton, “Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer’s disease,” Biol Psychiatry, vol. 64, no. 10, pp. 871–879, Nov. 2008.

[9] X. Qu, B. Yuan, and W. Liu, “A Predictive Model for Identifying Possible MCI to AD Conversions in the ADNI Database,” in Second International Symposium on Knowledge Acquisition and Modeling, 2009. KAM ’09, 2009, vol. 3, pp. 102–105.

[10] Gomar JJ, Bobes-Bascaran MT, Conejero-Goldberg C, Davies P, Goldberg TE, and Alzheimer’s Disease Neuroimaging Initiative f, “UTility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to alzheimer disease in patients in the alzheimer’s disease neuroimaging initiative,” Arch Gen Psychiatry, vol. 68, no. 9, pp. 961–969, Sep. 2011.

[11] D. P. Devanand, K. S. Michaels-Marston, X. Liu, G. H. Pelton, M. Padilla, K. Marder, K. Bell, Y. Stern, and R. Mayeux, “Olfactory Deficits in Patients With Mild Cognitive Impairment Predict Alzheimer’s Disease at Follow-Up,” Am J Psychiatry, vol. 157, no. 9, pp. 1399–1405, Sep. 2000.

[12] J. L. Shaffer, J. R. Petrella, F. C. Sheldon, K. R. Choudhury, V. D. Calhoun, R. E. Coleman, and P. M. Doraiswamy, “Predicting Cognitive Decline in Subjects at Risk for Alzheimer Disease by Using Combined Cerebrospinal Fluid, MR Imaging, and PET Biomarkers,” Radiology, vol. 266, no. 2, pp. 583–591, Feb. 2013.

[13] D. AL-Dlaeen and A. Alashqur, “Using decision tree classification to assist in the prediction of Alzheimer’s disease,” in 2014 6th International Conference on Computer Science and Information Technology (CSIT), 2014, pp. 122–126.

[14]“Amelia II: A Program for Missing Data.” [Online]. Available: http://gking.harvard.edu/amelia. [Accessed: 13-Nov-2014].

[15] ” ADNI | Access Data. ” [Online]. Available: http://adni.loni.usc.edu/data-samples/access-data/. [Accessed: 13-Nov-2014].