

SANTA CLARA UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Date: June 10, 2021

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Chelsea Fernandes
Shreya Venkatesh
Aiyushi Kumar

ENTITLED

Alzheimer's Disease Diagnostic Support Tool

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTER ENGINEERING

Thesis Advisor

Thesis Advisor

Department Chair

Department Chair

Alzheimer's Disease Diagnostic Support Tool

by

Chelsea Fernandes
Shreya Venkatesh
Aiyushi Kumar

Submitted in partial fulfillment of the requirements
for the degree of
Bachelor of Science in Computer Engineering
School of Engineering
Santa Clara University

Santa Clara, California
June 10, 2021

Alzheimer's Disease Diagnostic Support Tool

Chelsea Fernandes
Shreya Venkatesh
Aiyushi Kumar

Department of Computer Science and Engineering
Santa Clara University
June 10, 2021

ABSTRACT

Alzheimer's Disease is the 6th leading cause of death overall and the most common cause of dementia in older people in the US. The prevalence of the disease is projected to increase in the next few decades and disproportionately impact low/middle income populations. Unfortunately, specialized doctors, such as neurologists, may not be present in situations where a diagnosis is necessary, resulting in the possibility of AD being overlooked at its early and most treatable stages. Our proposed application is a tool that can aid doctors in determining a probable AD diagnosis using an inputted combination of imaging data, biomarkers, patient medical history data, and cognitive and functional assessments into a random forest machine learning model. It has a user interface designed with a focus on accessibility and simplicity. Our trained decision tree achieves an accuracy of 90% for a binary classification between CN and AD patients, and an accuracy of 77% for a multi-class classification between CN, MCI, and AD patients.

Acknowledgements

We would like to thank our advisor Dr. Ahmed Amer from the Department of Computer Science and Engineering for his support throughout the project. We would also like to thank our advisor Dr. Julia Scott with the Bioinnovation and Design Lab at Santa Clara University for continuously guiding us through the technical aspects of our project, and for connecting us with the Cortechs.ai team and their representatives Dr. Renee George and Dr. Christine Swisher. Our completion and success with this project could not have been possible without their support.

Table of Contents

1	Introduction	1
1.1	Background	1
1.1.1	Definition	1
1.1.2	Diagnostic Process	1
1.1.3	Impact	1
1.2	Problem Statement	3
1.3	Solution	3
2	Objectives	4
2.1	Conceptual Model	4
2.2	Requirements	4
2.2.1	Functional Requirements	4
2.2.2	Non-Functional Requirements	4
2.2.3	Design Constraints	5
2.3	Use Cases	5
2.3.1	Use Case Diagram	5
2.4	Technologies Used	5
2.4.1	Backend	5
2.4.2	Frontend	6
2.4.3	Fullstack Integration	6
2.4.4	Other Technologies	6
3	User Analysis and Research	7
3.1	Interviews	7
3.1.1	UC Davis Alzheimer’s Disease Center	7
3.2	Key Findings	8
3.2.1	Diagnosis Process	8
3.2.2	Common indicators of AD	8
3.2.3	User Interface	8
4	Project Development	10
4.1	Dataset	10
4.1.1	Overview	10
4.1.2	Dataset Breakdown	10
4.1.3	Data Preprocessing	11
4.2	Feature Considerations	11
4.2.1	Imaging Data	11
4.2.2	Demographics	12
4.2.3	Cognitive and Functional Tests	12
4.2.4	Biomarkers	12
4.3	Feature Selection	12
4.3.1	Final Input List	14
4.4	Model Research	14

4.5	Testing	15
4.5.1	Initial Testing Plan	15
4.5.2	Final Testing Plan	15
4.5.3	Preliminary Testing	15
4.5.4	First Stage: All Features	16
4.5.5	Second Stage: Missing Data Handling	16
4.5.6	Third Stage: Image Normalization	16
5	The Final Application	18
5.1	The Machine Learning Model	18
5.1.1	Description	18
5.2	Front-End Prototype	18
5.3	Final User Interface Design	20
6	Results	22
6.1	Model Performance Metrics	22
6.1.1	Background	22
6.1.2	Model Accuracy	22
6.1.3	Precision, Recall, and F-1 Scores	23
7	Discussion	24
7.1	System Analysis	24
7.1.1	Performance Comparison	24
7.1.2	Model Feature Contribution	24
7.2	Project Assessment	26
7.2.1	Advantages	26
7.2.2	Limitations and Future Work	26
7.3	Challenges and Project Complexity	27
7.4	Societal Impact	27
7.4.1	Ethical	27
8	Conclusion	29
9	Appendix	30
9.1	Team Approach	30
9.2	Development Timeline	30

List of Figures

1.1	Infographic Summarizing Alzheimer’s Disease Facts and Figures, an annual report released by the Alzheimer’s Association [1]	2
2.1	Use Case Diagram for Diagnostic Support Tool	5
4.1	Histograms representing the distribution of data for the RAVLT and Cerebral Volume features	13
4.2	Boxplots representing the distribution of data for the ADAS11 and CSF Volume features for the three classes	13
5.1	Home Page Prototype	18
5.2	Input Form Page Prototype	19
5.3	Classification Report Web Prototype	19
5.4	UI design: preview page	20
5.5	UI design: Demographic Inputs	21
5.6	UI design: Cognitive Tests Input	21
7.1	Pie Chart of Model Features with the Highest Contribution	25
9.1	Development Timeline	30

List of Tables

4.1	Data Breakdown by Class	11
4.2	Initial Feature List	11
4.3	Final Feature List	14
4.4	Chosen Models for Testing	14
4.5	CN vs AD Binary Test Accuracies for Given Models	15
4.6	Test Accuracies for Given Models - All Features	16
4.7	Test Accuracies for Given Models - With Data Imputation	16
4.8	Test Accuracies for Given Models - with Image Normalization	16
6.1	Final Model Performance Metrics	23
7.1	Comparable Alzheimer's Disease model classifiers with their respective accuracies	24

Chapter 1

Introduction

1.1 Background

1.1.1 Definition

Alzheimer's Disease, hereafter referred to as AD, is the most common cause of dementia, a class of neurological diseases typical in older adults. AD is characterized by the decline in cognitive and behavioral functions and accounts for between 60%-80% of all dementia cases. Caused by protein accumulation in parts of the brain, the loss of grey matter, identifiable changes in white matter integrity, and alterations in functional connectivity of key brain networks, the disease is currently incurable. Early diagnosis leads to the best prognostic outcomes as treatment is predominantly symptom and comorbidity management.

1.1.2 Diagnostic Process

The process begins with symptom recognition and progresses through the diagnostic testing, typically taking anywhere from 6 months to 5 years [3]. Patients are asked to take a battery of invasive and non-invasive tests measuring a combination of neuro-behavioral symptoms, biomarkers, and brain imaging to come to their conclusions. Teams of general practitioners, geriatricians, neurologists, radiologists, and primary caregivers have to communicate back and forth which can cause critical diagnostic information to be lost in translation. Because of the importance of catching the disease early to provide the most effective treatment plan, any method of making the diagnosis more efficient could greatly benefit patient prognosis.

1.1.3 Impact

5-8% of the general global population above the age of 60 has dementia, of which Alzheimer's is the most common cause. The prevalence of the disease is projected to increase in the next few decades and disproportionately impact low/middle income nations. Dementia not only physically and psychologically impacts the patient suffering from the disease but also those around them, such as caregivers, families and society [2].

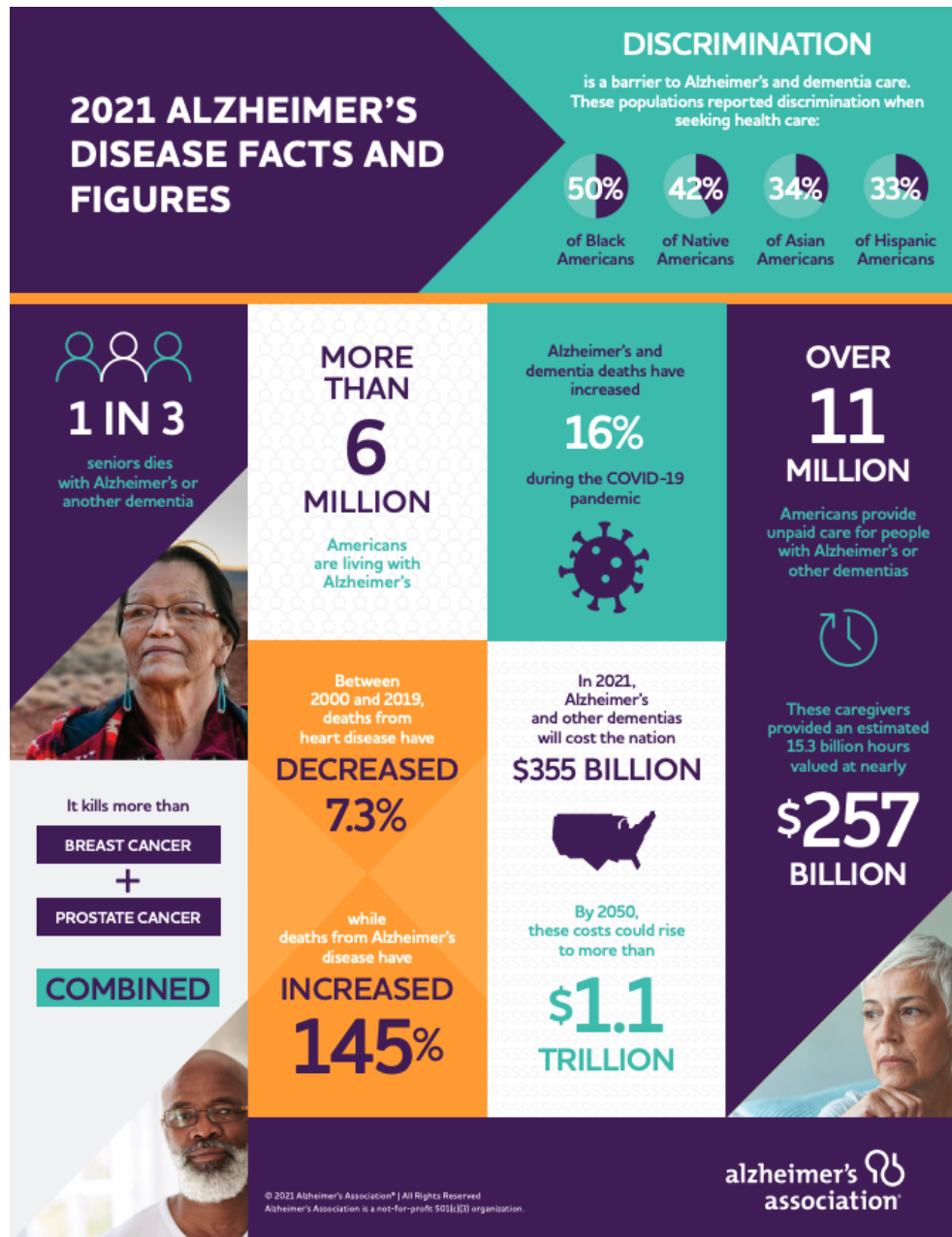


Figure 1.1: Infographic Summarizing Alzheimer's Disease Facts and Figures, an annual report released by the Alzheimer's Association [1]

1.2 Problem Statement

It can take years to diagnose a typical AD patient, at which point medical treatment options are limited, at best. In the current medical system, radiological imaging technologies and biomarkers, despite their many benefits, are underutilized in the AD diagnostic process. Much of the information available is lost in translation between the technology, radiologists, doctors, and patients because of time constraints and human limitations. Patients in areas significantly lacking medical resources have limited opportunities to raise concerns of developing neuro-degenerative diseases resulting in an even more delayed AD diagnosis. Unfortunately, specialized doctors, such as neurologists, may not be present in situations where diagnosis is needed, resulting in the possibility of AD being overlooked at its early stages. [2]

1.3 Solution

We propose an end to end technical solution to assist general practitioners and neurologists in processing patient data and diagnosing whether or not a patient has AD. We created a machine learning classification model that will reach a probable diagnosis and identify the stage of disease progression, if applicable. We used a combination of structured imaging data and biomarkers alongside more typical diagnostic criteria. The use of biomarkers in AD diagnosis has yet to be utilized to their fullest potential but its integration into the process can dramatically improve the likelihood of an early diagnosis.

Our target users are general practitioners who may not have access to specialists when attempting to make diagnoses for their patients. The tool is intended to simplify and expedite the diagnostic process. It provides a comprehensive list of tests to conduct along with symptoms and biomarkers to identify leading to a more accurate and efficient diagnosis.

For the machine learning element of the project, we developed a multi-class classification model that determines the stage of Alzheimer's disease a patient has (cognitively normal [CN], mild cognitive impairment [MCI], or Alzheimer's's Disease [AD]). As our inputs, we used imaging data, biomarkers, patient medical history data, and cognitive and functional assessments. These inputs are fed into various machine learning models to determine which is the most accurate and generalized. The goal is to store the classification output into a database, along with other metrics that can be useful for physicians, which will then be provided in a PDF document to view.

Chapter 2

Objectives

2.1 Conceptual Model

The objective of the system is to be able to predict the stage of Alzheimer's Disease a patient is in similar to what a neurologist would do. The best system would emulate the diagnosis process at least to a certain degree. The inputs into the system include general patient information and measurements, using a user-friendly web interface. The backend of the system would then process the inputs into a model-friendly format and feed the inputs into the model. Once the model has made the predicted classification, the patient's suggested diagnosis is outputted back to the physician. The physician can then use the classification output along with any reported information related to the classification to determine the best next step for the patient.

2.2 Requirements

2.2.1 Functional Requirements

- Allows a medical professional to input patient data
- Accepts patient data in some format.
- Classifies the patient as CN, MCI, or having AD
- Outputs the classification as a report to the user

2.2.2 Non-Functional Requirements

- Web-based tool
- User interface is designed for ease-of-use
- Model performs classification with an accuracy greater than 90%
- Classification is bounded by an associated level of uncertainty

- Conveys why the model chose a specific classification

2.2.3 Design Constraints

- User interface for manual data entry
- Training of model through publicly available datasets only
- Working within memory requirements of the given computer
- Model accepts heterogeneous data
- Model trains with a limited dataset

2.3 Use Cases

There is one main use case for our system, which is that a physician must be able to manually input patient information into a web page. Once submitted, a classification report is outputted, which the physician may then use to come to a diagnosis, and can be shared with the patient. Since both the physician and patient may be reviewing the classification report, it is important that the report is easy to read and is organized to effectively communicate the results of our machine learning model.

2.3.1 Use Case Diagram

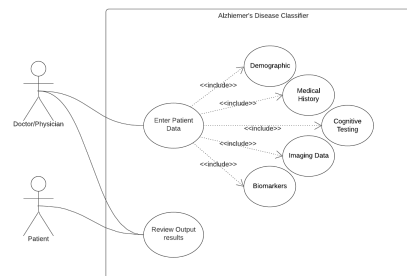


Figure 2.1: Use Case Diagram for Diagnostic Support Tool

The Use Case Diagram shows how the intended system connects the frontend with the backend.

2.4 Technologies Used

2.4.1 Backend

Python

Python is a high-level general-purpose programming language that supports the relevant data science-related libraries required for the development of our application. These libraries include Pandas, Numpy, and Scikit-learn (sklearn).

Jupyter Notebooks

Jupyter Notebooks is an open-source web application that allows users to develop and run live code and seamlessly create visualizations. The majority of our model development was completed in Jupyter Notebooks, from which the final model was transferred to a Python (.py) file for integration with our frontend.

Hardware Configuration

The computer used to train our model and build our frontend prototype was the Apple Macbook Air, 2015. It has a 1.6 GHz Dual-Core Intel Core i5 Processor with 8 GB of RAM.

2.4.2 Frontend

HTML

HTML (Hyper Text Markup Language) is a standard language for making web pages. This was the primary language used to create the basic prototype, along with CSS (Cascading Style Sheets) for styling.

2.4.3 Fullstack Integration

Flask

Flask a micro web framework written in Python that allowed for seamless integration between our front- and back-end technologies.

2.4.4 Other Technologies

Git and Github

We used Git and Github for tracking software -related changes and for storing important documentation and required files for our application.

LaTeX

LaTeX is a web application that was used for creating and formatting our documentation in a clear and readable format.

Chapter 3

User Analysis and Research

Before making any decisions on the design and implementation of our system, we conducted extensive research on potential data inputs and the best approach to our problem statement.

3.1 Interviews

The research process began with exploratory interviews to gather information about the current state of the field. We wanted to explore what day-to-day applications professionals were already using to identify successes and failures in the system. With a combination of a focus group and user interview, we hoped to design a tool that was intuitive and supportive of the clinicians we are providing for.

3.1.1 UC Davis Alzheimer's Disease Center

UC Davis Alzheimer's Disease Center is a group of interdisciplinary researchers, clinical staff, and support staff located in Sacramento and Walnut Creek. They work to advance our understanding of AD through clinical patient evaluation and other cutting edge research. In August, our team had the opportunity to interview Dr. DeCarli and other members of his staff about what AD is.

At the time of the interview, our goal was outlined by the following statement: A tool that doctors can use in a hospital setting to increase the efficiency and accuracy of early detection of Alzheimer's Disease diagnosis. We asked the experts at UC Davis to provide us with insight into the current diagnostic process, everything from the kind of clinical testing that is conducted on patients to the various steps of escalation doctors go through from the moment a patient first walks in the doors to when they have a confirmed diagnosis.

We asked the following questions to get expert advice on the current systems in place for producing a diagnosis.

- What is the timeline for a patient from when they first walk in with memory complaints to a result?
- How critical is integration with the current PACS systems?

We also wanted to begin the user interface and experience design process.

- Is there a strong preference for particular input styles (dropdown menus, radio buttons, checkbox multi selections, tables, etc)
- What kind of data visualizations and numerical outputs would be helpful (tables, error bars, graphs, text, etc)
- What is the importance or usability of a “patient facing” output with simplified information?

And finally, we knew from the start that we needed to build user confidence in the tool in order for it to be successful.

- How can we increase confidence level in the final diagnosis?
- How do we express a feeling of transparency of calculations between the input and the output?

3.2 Key Findings

We received significant insight into the AD diagnosis process and were able to use the insight to better plan our own next steps.

3.2.1 Diagnosis Process

It can take around 5 years into the disease progression before a patient is brought in to see a doctor. In fact, the first doctor may not even recognize the symptoms as a form of dementia. In many cases, the family of the patient notices the impairment, namely behavioral changes. The diagnosis timeline is highly subjective and differs from patient to patient, as does the severity of disease progression.

3.2.2 Common indicators of AD

A common indicator of dementia is memory loss, however it can be difficult to determine whether that is on the trajectory towards dementia or just an age related memory loss, which experiences much slower change over time. Memory can be tested using cognitive testing, of which there is data available via ADNI. It was recommended that we specifically look at functional cognitive tests. Regarding imaging data, MRI scans are much more common for patients to receive compared to PET scans, so incorporating MRI into the tool would likely help a larger population. We also learned to use either CSF or PET, but not both because they are highly correlated and highly expensive, making them inaccessible to many patients. It was also recommended to look at changes in white matter measurements could be helpful as well, as opposed to grey matter.

3.2.3 User Interface

For the user interface, it was recommended that we keep it as simple and user-friendly as possible. Providing confidence levels and data visualizations of how the inputs were used in the backend was a common suggestion. The

experts believed a tool that was straightforward and allowed both physicians and patients to understand their medical data was important.

Chapter 4

Project Development

4.1 Dataset

4.1.1 Overview

For our dataset, we sourced data from the Alzheimer’s Disease Neuroimaging Initiative - or ADNI - which is an open source dataset. The goal of this initiative is to identify markers for the disease, making it a good resource for developing a classification model. In ADNI, we used the ADNI GO and ADNI 2 studies since it was the most updated dataset and had patient information from all the other ADNI studies. We used baseline data and categorized our data into demographics, biomarkers, imaging data, and cognitive and behavioral test scores. From our discussion with the UC Davis ADC, we concluded that the following categories of clinical parameters should be included as features in our final dataset.

- Demographic Information
- Biomarkers
- Imaging Data (Brain Volumes)
- Cognitive and Behavioral Data

4.1.2 Dataset Breakdown

From ADNI, we were able to obtain a dataset consisting of 1,194 anonymous medical records (Table 4.1), which is considerably small for a classification problem like ours. This issue stems from ensuring that all the features we were considering were present in our set, which was difficult to obtain as the 2 groups of ADNI data we were sourcing from were not consistently linked, resulting in many records not having the majority of features we needed.

Cognitively Normal (CN)	Mild Cognitive Impairment (MCI)	Alzheimer's Disease (AD)
359	673	162

Table 4.1: Data Breakdown by Class

4.1.3 Data Preprocessing

Some preprocessing was required upon acquiring the data from ADNI. Since some patients were represented in both the ADNI GO and ADNI 2 datasets, we first merged the records based on the anonymous record number using a python script. We then removed duplicate records such that each patient carried equal weighting. Next, we recoded non-numerical categorical data to numerical values, to meet data format requirements for our model types. We then normalized all continuous variables by mapping values to within the interval [0,1]. This standardization technique appropriately rescaled values that differed widely so that they were comparable.

4.2 Feature Considerations

The features that we considered for the tool were split into 4 main categories - imaging data, demographic information, functional and cognitive tests, and biomarkers.

Imaging Data	Demographics	Cognitive Tests	Biomarkers
Hippocampus Vol.	Age	ADAS 11	APOE Gene
Ventricles Vol.	Gender (M/F)	ADAS 13	CSF ABETA values
Entorhinal Vol.	Racial Origin	RAVLT immediate	Tau
CSF Vol.	Education (Years)	MMSE	Ptau
Total Intracranial Vol.	Marital Status	FAQ TOTAL	
Cerebral Vol.		E-Cog Tests: MEM, LANG, VISSPAT	

Table 4.2: Initial Feature List

4.2.1 Imaging Data

Brain scans are the primary form of information showing tissue degeneration in the brain for patients with AD or other forms of dementia. Specific regional brain volumes are essential to identify loss of tissue. Hippocampal volume is specifically considered a strong indicator of brain atrophy over time for predicting disease progression [5]. Full brain volumes or the intracranial volume is essential for calculating the ratio of the tissue loss compared to the patient's actual brain volume. This ratio of regional brain volume to the full brain volume gives a better understanding of tissue loss due to dementia compared to the original brain volume of the patient. There is significant evidence showing the relation between loss in hippocampal, ventricular and entorhinal volumes and Alzheimer's Disease progression [9].

4.2.2 Demographics

Key demographic data include number of years of education, age, gender, marital status, and race. Research shows that people who are illiterate are more prone to AD compared to people having more advanced education [4]. Although the correlation between age and onset of dementia has not been completely proven, most of the patients in our study are above the age of 60. Research shows that about two thirds of persons diagnosed with AD dementia are women [6]. There is also concrete evidence suggesting the relationship between gender and diagnosing AD. Lastly, ethnic background [8] and marital status [10] have proven to be key demographic factors for diagnosing AD. In our dataset, we do, however, have most of our patient data consisting of American white and black populations. Including patients from the ADNI global dataset would be the next step to include more racial categories in the dataset.

4.2.3 Cognitive and Functional Tests

Our initial dataset includes the total score from the Functional Assessment Questionnaire (FAQ) test which is a scale that focuses on a person's ability to function and perform tasks of daily living. The score range for each task is (0-3) where a higher score indicates greater impairment. The total FAQ score is the sum of all the scores of all tasks listed in the questionnaire. For cognitive tests, our dataset includes the scores for the ADAS 11, 13 and MMSE tests. ADAS, also known as the Alzheimer's Disease Assessment Scale, is a two part scale designed to assess cognitive decline where scores range from 0 - 70, with higher scores indicating severe impairment. MMSE or the Mini Mental State Examination is a measure of cognitive function and is used as a preliminary screening tool. The scoring for this test ranges from a 0-30, with higher scores indicating severe impairment.

4.2.4 Biomarkers

The biomarkers we planned on using initially were the cerebrospinal fluid Beta Amyloid (CSF ABETA), APOE gene, CSF Tau and Plasma Tau values. We decided to rule out Tau values since they are obtained from PET scans, which was out of the imaging data scope that was originally decided upon. The CSF Abeta is a major indicator of Alzheimer's dementia and research proves that low CSF ABETA levels were indicators of early stages of dementia [7]. There is significant evidence proving the relation between brain plaque formations and Alzheimer's Disease. A patient with Alzheimer's develops abnormal levels of protein clumps known as plaques which collect between neurons and destroy brain function. The Apolipoprotein E or APOE gene is speculated to be involved in the formation of these plaques. Our dataset consists of patients with and without the presence of the APOE gene allele.

4.3 Feature Selection

For our feature selection process, we decided to manually analyze various visuals of the data to determine which features were most feasible in contributing to a classification. This was completed by plotting histograms and box

plots for each of the shortlisted features to determine how well the data separates the three classes and ultimately select features that improve the model performance. Both portrayals of data show the same distribution for each class, but highlight different characteristics.

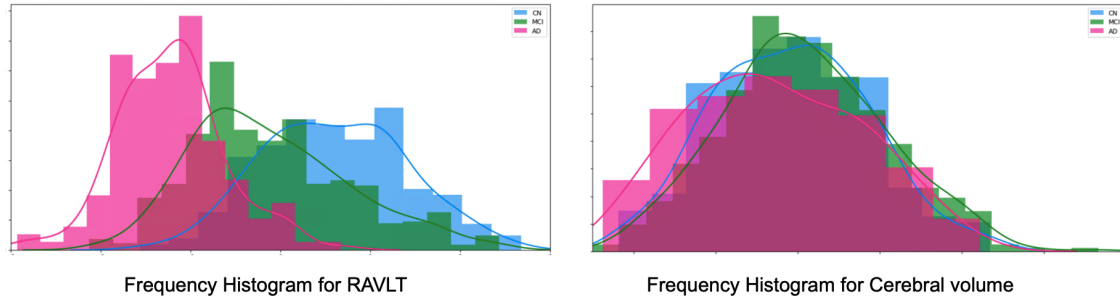


Figure 4.1: Histograms representing the distribution of data for the RAVLT and Cerebral Volume features

In the histograms 4.1, overlapping peaks of the three classes indicate that the feature has no discriminatory values between the classes. For example, the histogram on the left has minimally overlapping peaks and shows the RAVLT cognitive test can be used as a feature to separate the three classes. On the other hand, the histogram on the right completely overlaps the three classes showing that the Cerebrum values are not ideal for separating the classes. We used this thought process to analyze histograms for all of the features we were considering.

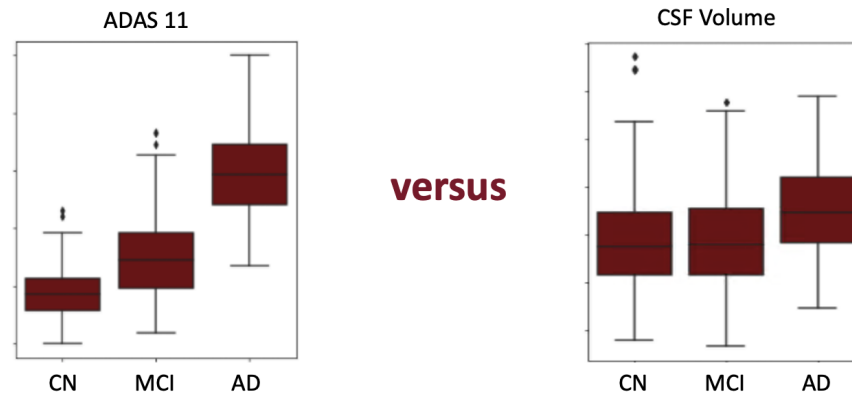


Figure 4.2: Boxplots representing the distribution of data for the ADAS11 and CSF Volume features for the three classes

Similar to the histograms, the box plots in figure 4.2 show overlapping data ranges between classes, noted by overlapping boxes. However, the box plots are better at revealing any outliers for each class of a given feature. Features that contain data with many outliers are undesirable as they can distort results and lower accuracy and precision of the model. In Figure 4.2, the box plot on the left shows a clear distinction in data between the 3 classes with minimal outliers. However, the box plot on the right shows overlapping data denoted by the intersecting interquartile ranges of the plots.

4.3.1 Final Input List

Imaging Data	Demographics	Cognitive Tests	Biomarkers
Hippocampus Vol.	Age	ADAS 11	APOE Gene
Ventricles Vol.	Gender (M/F)	ADAS 13	CSF ABETA values
Entorhinal Vol.	Racial Origin	RAVLT immediate	
CSF Vol.	Education (Years)	MMSE	
Total Intracranial Vol.	Marital Status	FAQ TOTAL	
Cerebral Vol.			

Table 4.3: Final Feature List

For our final input list, we kept most of the features that we initially considered, minus the Tau/Ptau values and E-Cog tests. This was based on our conversation the UC Davis ADC along with our analysis of histograms and boxplots.

4.4 Model Research

There are several important characteristics we considered when determining which models we wanted to test with our curated dataset. The most important characteristic was that the models could handle both continuous and categorical data, since the data we obtained from ADNI reflected those data types. Another important requirement of the models was that they should be suitable for a multi-class classification problem, as opposed to a binary one. This requirement meant we could no longer consider linear regression as one of our options since it performs best for binary classifiers. Another requirement was that our selected models accepted normalized data, which the majority of models we researched supported.

When initially discussing solutions, we considered neural networks as another potential solution due to its robustness and popularity as a deep learning technique. However, upon further research, it was determined that our curated dataset was not suitable for a neural network, as such a model requires thousands of instances of data. All the models we selected were better designed to handle a smaller dataset like the one we curated.

Model	Categorical + Continuous Data	Requires Normalization	Multiclass Classifier
K-Nearest Neighbor	X	X	X
Decision Tree	X		X
Support Vector Machine	X	X	X
Naïve Bayes Classifier			X

Table 4.4: Chosen Models for Testing

The models we ultimately used for testing are provided in Table 4.4, and were chosen based on their alignment to our requirements and their general performance as models for a classification problem. While the Naive Bayes

Classifier isn't designed to handle both continuous and categorical together, there are two versions of the model that can handle them separately. We decided to include one version of the model in our testing regardless as an experimental model, specifically the version that handles continuous data (Gaussian Naive Bayes). All of the chosen models were available through Python's scikit-learn library, which made implementation simple and efficient.

4.5 Testing

4.5.1 Initial Testing Plan

With our initial testing plan, we ran multiple test iterations to see how the accuracy was affected as we changed various parameters and formatting of the data. For each iteration, we ran the dataset through the four previously selected machine learning models (see Model Research). We also ran the data with 4 different classifiers: CN vs MCI vs AD, CN vs (MCI+AD), and MCI vs AD. The reasoning was that by doing so, we would have a better understanding of how the model performed in terms of differentiating one class from another, and whether or not the MCI class was more similar to the CN or AD class. It was later determined that this was unnecessary as each model report came with Precision, Recall, and F1-Scores, which explain how the data was classified using the model compared to the actual classifications.

4.5.2 Final Testing Plan

For our final testing plan, we had two levels of model evaluation. The first level was based on accuracy of each model. The models with the highest accuracy values were then analyzed using precision, recall and F1-score metrics per class. From there, we varied which features were part of the training set until we reached an optimal accuracy.

4.5.3 Preliminary Testing

To make sure that our initial data selection and preprocessing was performed sufficiently, we first ran our model as a binary classifier (as opposed to a multiclass one) to classify between Cognitively Normal and Alzheimer's Disease patients, which are the 2 extremes of patients in the spectrum. The results of this run was an accuracy of approximately 100% across the 4 different models (see Table 4.5), which confirmed that the features we selected were effective in classifying the 2 different classes.

Performance Metric	Decision Tree	KNN	SVM	Naïve Bayes
Accuracy	1	0.98	0.97	1

Table 4.5: CN vs AD Binary Test Accuracies for Given Models

4.5.4 First Stage: All Features

For our first stage of testing that involved all three classes, we ran the model using all 18 of our initially curated features without any further preprocessing. As noted in table 4.6, all models performed with an accuracy above 69%, which was far from ideal. Upon further analysis, we found that a majority of the data was getting removed before training the model due to missing features for each data instance. This unwanted removal of data was addressed in the second stage.

Performance Metric	Decision Tree	KNN	SVM	Naïve Bayes
Accuracy	0.69	0.74	0.77	0.72

Table 4.6: Test Accuracies for Given Models - All Features

4.5.5 Second Stage: Missing Data Handling

In order to maximize the amount of data we could use to train our model, we decided to impute, or fill in the missing data using the k-nearest neighbors (KNN) algorithm. This prevented the model from throwing away records that did not have all features. The KNN algorithm was chosen due to its ability to handle continuous, discrete and categorical data, which are all data types that our feature list consists of. While our accuracies improved drastically, we acknowledged that this was a result of data overfitting, meaning that while our model could accurately classify patients in our current dataset, this may not be translatable to records from other datasets.

Performance Metric	Decision Tree	KNN	SVM	Naïve Bayes
Accuracy	0.98	0.92	0.91	0.97

Table 4.7: Test Accuracies for Given Models - With Data Imputation

4.5.6 Third Stage: Image Normalization

To address the overfitting, we decided to only impute the data of features that was missing the largest proportion from our dataset in our third stage of model testing. From a quick analysis, we determined that the imaging data overall was missing the greatest amount of data out of all the records, which led us to only imputing the imaging data. While this means that some records would still have missing data and would subsequently be thrown away, we decided that this was much better than overfitting our model to our dataset. Upon greater analysis of our data, we realized that our

Performance Metric	Decision Tree	KNN	SVM	Naïve Bayes
Accuracy	0.77	0.75	0.72	0.75

Table 4.8: Test Accuracies for Given Models - with Image Normalization

initial method of normalization for the imaging data was not suitable given the type of data we had. Imaging data refers

to volumetric measurements of various parts of the brain (i.e. hippocampus, entorhinal, ventricles). These volumes can vary from patient to patient, so it was imperative that we normalized the data in a way that does not imply that a generally large or small brain leads to a specific diagnosis. This led us to revising our initial normalization process of mapping all imaging values to a value between the interval $[0,1]$ to normalizing each measurement by dividing by the total intracranial volume (ICV). The results from this stage were much more realistic. Upon changing our method of normalization for the imaging data, we began removing features to see if certain features played a larger or smaller role in the classification. All of our results from this stage remained relatively close to the results displayed in table 4.8. This led us to the conclusion that the accuracies generated from this stage were the

Chapter 5

The Final Application

5.1 The Machine Learning Model

5.1.1 Description

The model we ultimately integrated with our web form was the decision tree trained during the third stage of our testing phase. See *Final Model Performance Metrics* in the discussion section for an analysis of the integrated model's performance.

5.2 Front-End Prototype

A front-end prototype was created to show the functioning end-to-end flow of our system.

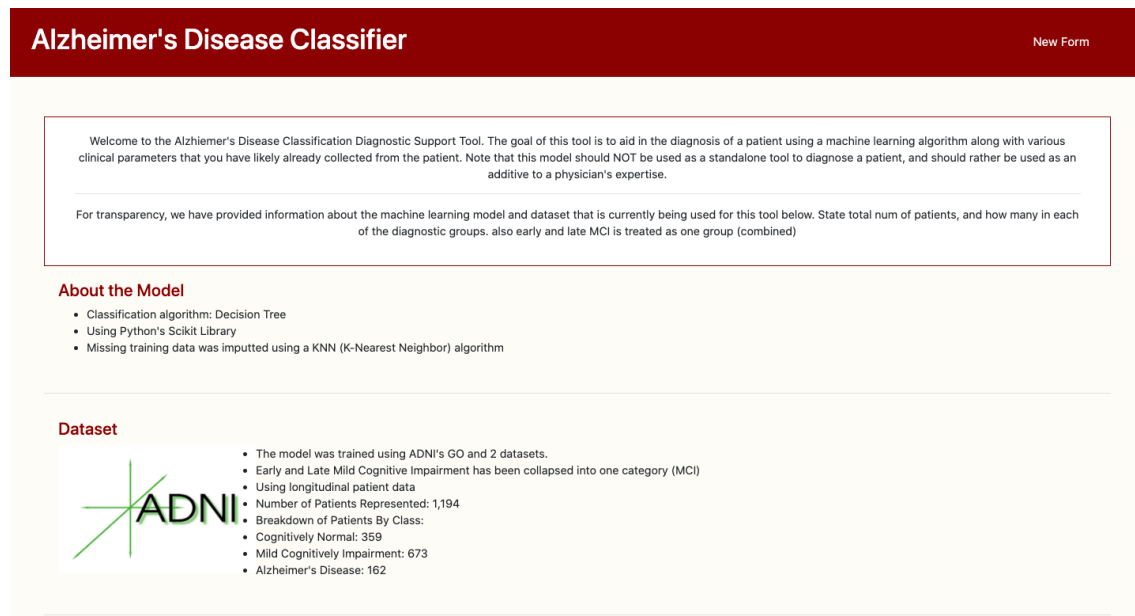
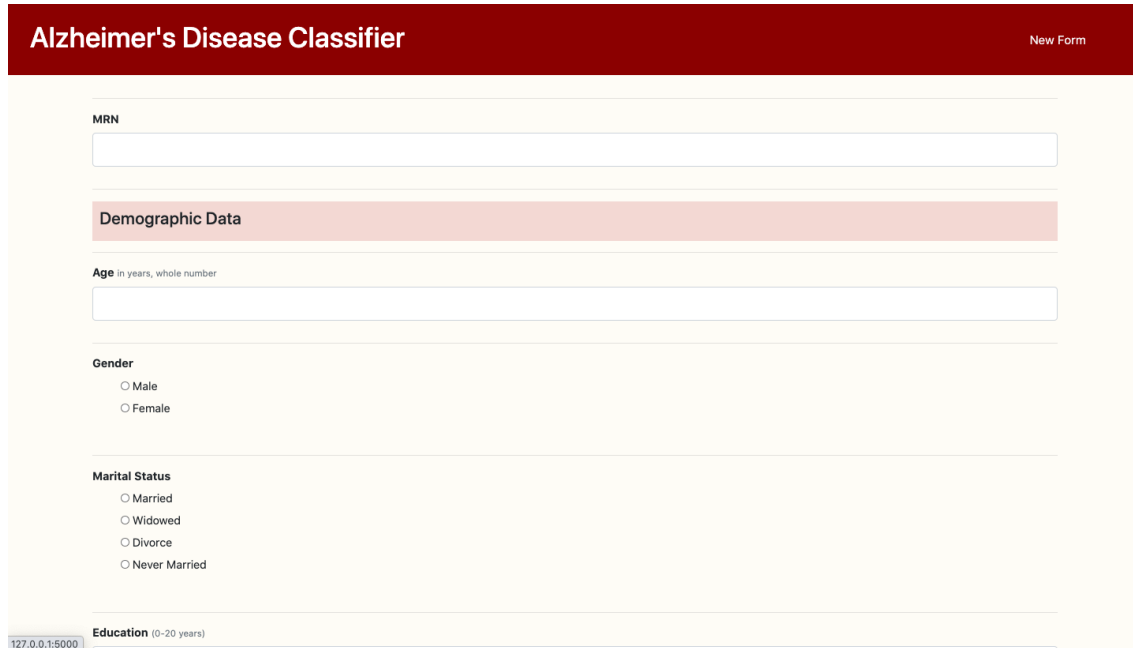


Figure 5.1: Home Page Prototype

The goal of the home page (Figure 5.1) was to provide information about the tool overall, including the ma-

chine learning model used, characteristics of the dataset, and key features used in the model. Additionally, there is a disclaimer advising that the tool be used in support of a physician's decision, and not as the final diagnosis alone. Providing this information in the form of a home page allowed us to maintain user transparency, which is one of our non-functional requirements.



The image shows a web form titled "Alzheimer's Disease Classifier" with a "New Form" button in the top right corner. The form is divided into several sections: "MRN" with a text input field; "Demographic Data" with a sub-section "Age in years, whole number" and a text input field; "Gender" with radio buttons for "Male" and "Female"; "Marital Status" with radio buttons for "Married", "Widowed", "Divorce", and "Never Married"; and "Education (0-20 years)" with a text input field. A small status bar at the bottom left shows "127.0.0.1:5000".

Figure 5.2: Input Form Page Prototype

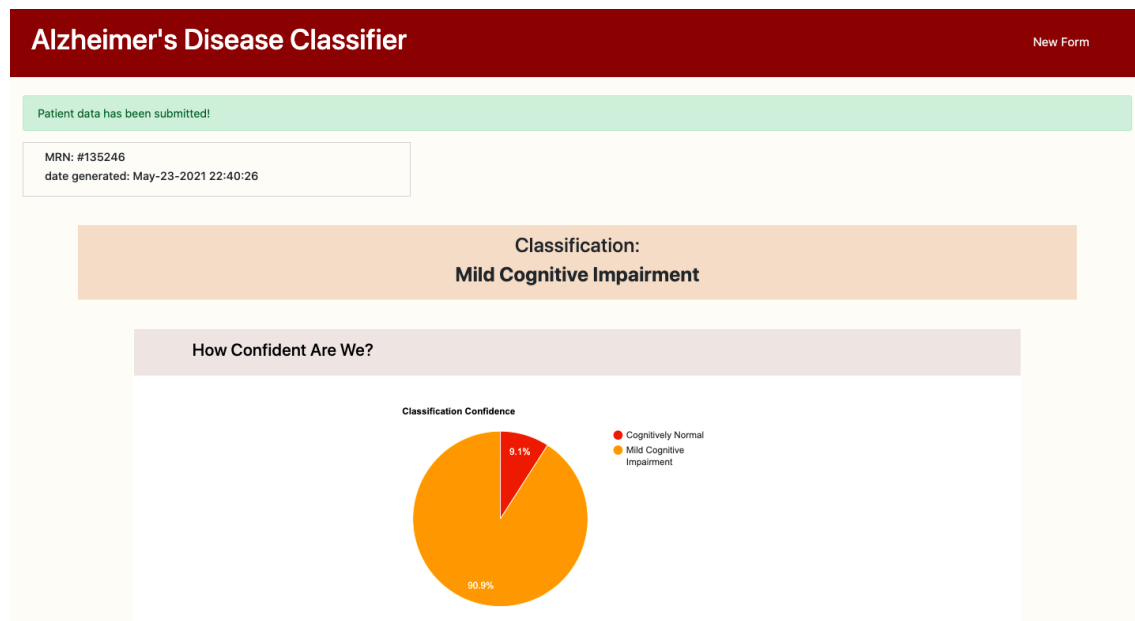


Figure 5.3: Classification Report Web Prototype

Since the physician is required to pull information from various parts of the patient’s medical record and manually input the data into our form, it was essential that we keep the form as simple to use as possible. The form is separated by feature category using headings, as shown by Demographic Data in Figure 5.2.

The classification report prototype (See Figure 5.3) provides the patient’s predicted diagnosis at the top of the page, a pie chart displaying the model’s level of confidence, and patient data fed through the model. Confidence level values were generated using the predict-proba function and passed to the frontend, where a javascript function generated the actual pie chart with the given values (see Figure 5.3). Both tables and forms were created using bootstrap, a common css library.

5.3 Final User Interface Design

The following are images of the User Interface that we would want to implement in the future. The interface better takes into account the suggestions made by the experts from the UC Davis ADC.

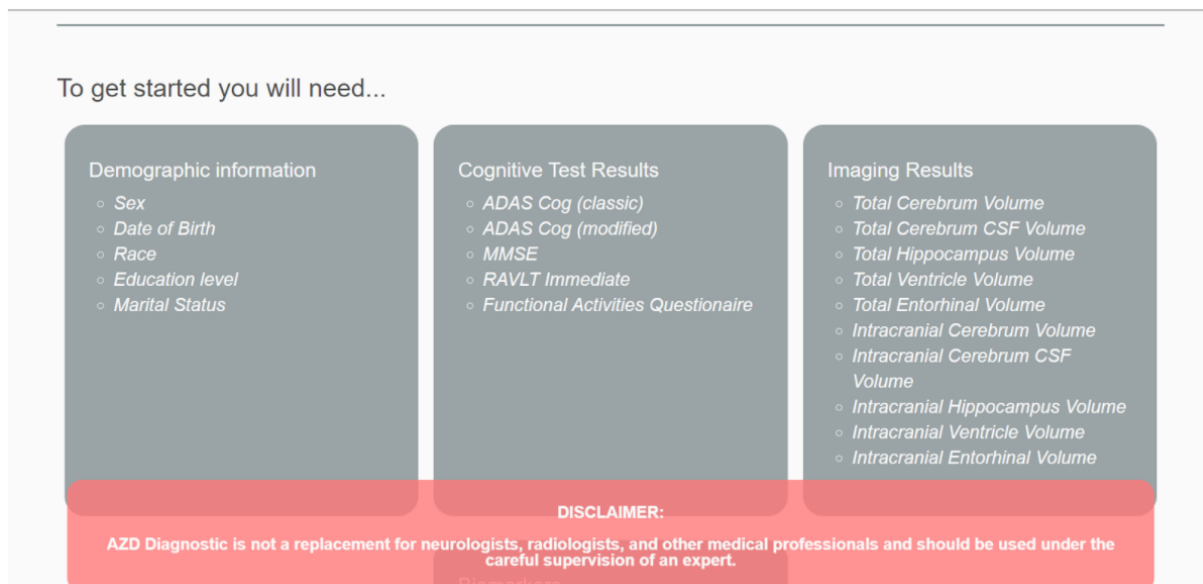



Figure 5.4: UI design: preview page



Home

Demographics

Cognitive Testing

Imaging

Biomarkers

Demographics

Sex

☐ Male ☐ Female

Date of Birth

MM/DD/YYYY

Race

American Indian or Alaskan Native

Education


0

Marital Status

Married

Next

Figure 5.5: UI design: Demographic Inputs



Home

Demographics

Cognitive Testing

Imaging

Biomarkers

Cognitive Testing

ADAS Cog - classic (11 items)

Range: 0 - 70

ADAS Cog - modified (13 items)

Range: 0 - 85

MMSE

Range: 0 - 30

RAVLT Immediate

Range: 0 - 100

Functional Activities Questionnaire

Range: 0 - 30

Previous

Next

Figure 5.6: UI design: Cognitive Tests Input

Chapter 6

Results

6.1 Model Performance Metrics

6.1.1 Background

Among the various metrics used to evaluate the performance of machine learning classification algorithms, we decided to use the precision, recall and F-1 scores. Since the tool is meant for a medical diagnosis application, it is important to measure the false positives and false negatives. The number of false positives and false negatives determines the usability of a tool such as ours in a real-life application.

Precision

Precision is the number of true positive results divided by the number of total positive results predicted by the classification model. This total includes the true positives and false positives.

Recall

Recall is the number of true positive results divided by the number of relevant results i.e. the total number of patients that should have been classified as positive. This total includes the true positives and false negatives.

F1-Score

F1 score is the mean between precision and recall. F1 scores typically range between 0-1. The greater the F1 Score, better is the model performance.

6.1.2 Model Accuracy

Our integrated model had a performance accuracy of 77.3%.

Class	Precision	Recall	F1-Score
AD	0.78	0.91	0.84
MCI	0.76	0.84	0.8
CN	0.8	0.56	0.66

Table 6.1: Final Model Performance Metrics

6.1.3 Precision, Recall, and F-1 Scores

Table 6.1 shows the precision recall and F1 score for our final tool. On analysing the table, one can note that the recall value for the cognitively normal class is very low whereas the recall value for Alzheimer’s Disease is high. This means that the Cognitively Normal class is not being rightly classified. Many of the Cognitively Normal patients were being classified as MCI patients. This is mainly due to the homogeneous nature of the MCI subgroup. The MCI class is split into early and late stages of MCI and it is extremely difficult to classify between the CN and early MCI patients as well as between the late MCI and AD patients due to overlapping symptoms and similar cognitive test scores. The MCI class also has a separate category of converts where patients who were in early MCI stages fell back to the CN stage, late MCIs converted to either early MCI or AD, or remained stagnant and did not progress to AD. The complexity of the MCI class was the main reason for lower accuracies in our model performance.

Chapter 7

Discussion

7.1 System Analysis

7.1.1 Performance Comparison

Author	Model Used	Accuracy
C. Jiménez-Mesa et al., 2020	SVM	67%
Katherine R. Grey et al. 2013	Random Forest	75%
Moore, 2019	Random Forest	79%
Our Model	Decision Tree	77.30%

Table 7.1: Comparable Alzheimer’s Disease model classifiers with their respective accuracies

While we were unable to achieve the targeted accuracy of above 90%, our solution is still reasonable based on the complexity of the problem as previously discussed. Compared to other other models that have been implemented for a similar problem statement, our accuracy of 77.3% falls within range of expected performance.

7.1.2 Model Feature Contribution

From the model, we were able to extract the features that contributed most to the classifications. From Figure 7.1, it appears that the features with the largest contribution were MMSE and FAQTOTAL, accounting for almost 75% of the classification decision. Both these features are cognitive tests, implying that cognitive tests play a large role in diagnosing a patient with Alzheimer’s Disease. This finding aligns well with our discussion with the UC Davis ADC as well as other experts in the field. Also notable is the only imaging data feature that had a large enough contribution to the classification was the CSF volume. This is interesting because we had expected more imaging features to have a strong contribution to the classifications based on related research. Nevertheless, it is clear that cognitive tests are important in classifying a patient using our specific model.

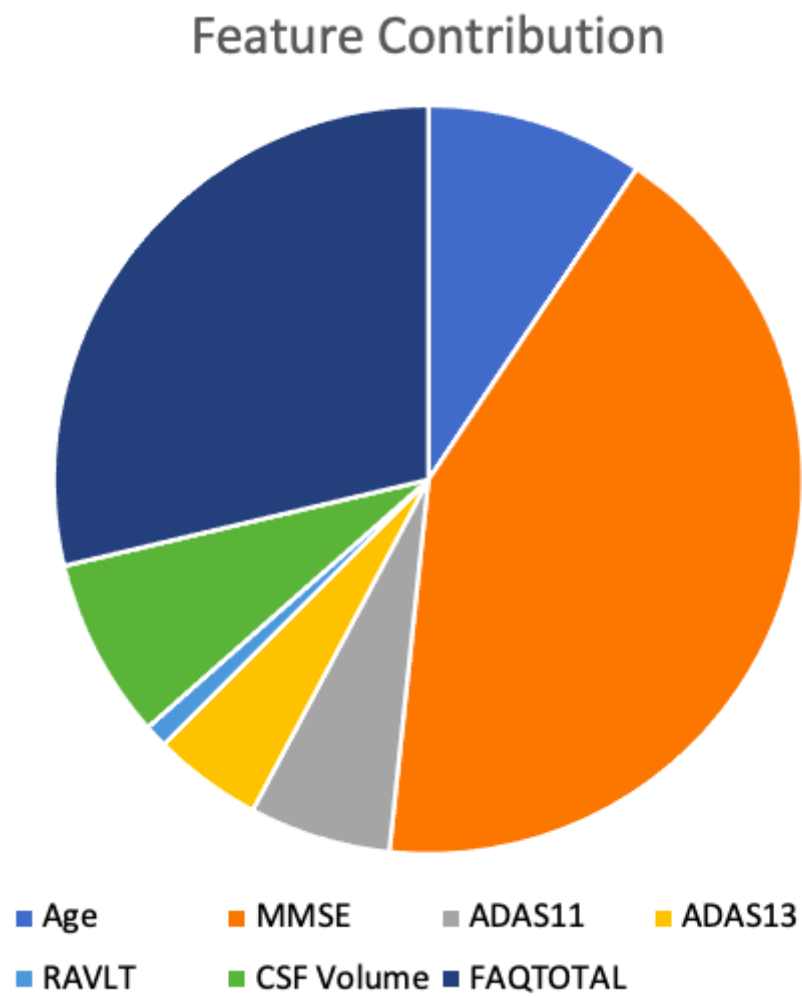


Figure 7.1: Pie Chart of Model Features with the Highest Contribution

7.2 Project Assessment

7.2.1 Advantages

In terms of functional requirements, we were successful in building an end-to-end system that accepts manual input from a physician and outputs the patient's probable diagnosis as a report. Our model was fully integrated into our system and is able to classify if a patient has Alzheimer's Disease or not with a certain level of confidence. From the model, we were able to extract the level of confidence the model has in classifying the patient, along with the features used to do so. Based on our preliminary testing using a binary classifier, it is evident that our model is able to very accurately predict if a patient is Cognitively Normal or has Alzheimer's Disease.

7.2.2 Limitations and Future Work

Although we were successful in building a complete system, there are still some limitations to our tool. The first, and most notable is that our model accuracy is not at the 90% threshold that we had outlined at the beginning of this project. While the model successfully classifies between CN and AD patients, the classification accuracy between MCI and AD is far from perfect, and is the biggest reason for our overall low accuracy rate. As discussed in Performance Comparison, our obtained accuracy is reasonable based on the complexity of the problem and similar attempts to create a model.

Another limitation is that our model is unable to handle missing data as input into the model. During our first stages of testing, we found that our model did not accept missing data, and would throw out any records that contained missing data, which led us to imputing our dataset. This is sub-optimal because in a real-world scenario, it is likely that a patient needing to use the tool will not have all the required features present with them. In future iterations, the project team will need to come up with a solution to accepting missing data, such as imputing the patient data before feeding it through the model.

One other disadvantage to our system is that not enough feature testing was completed to rule out features that are of lesser importance. From our analysis on feature importance, it was determined that the imaging data had a limited contribution to the classification. This can be seen as a positive since it means that cognitive tests such as FAQTOTAL and MMSE are more relevant to our model, which are likely accessible since general physicians themselves are able to conduct the tests as questionnaires. Imaging data on the other hand may not be as easy to obtain since the methods to retrieve such data can be costly. Nevertheless, more testing on the features themselves need to be done in future iterations of the project.

7.3 Challenges and Project Complexity

The diagnosis of Alzheimer's disease is not a simple classification problem. From a clinical perspective, diagnosis of an Alzheimer's related dementia falls under a very grey area. In our case, despite researching the clinical causes and factors involved in the diagnosis of Alzheimer's disease, many of the inputs were redundant when it came to diagnosing the disease in the machine learning models we implemented. This is mainly because of the homogenous nature and overlapping symptoms for the MCI category of patients. In the ADNI database, the MCI category was divided into two subsections - Early and Late MCI. To put things into perspective, mild cognitive impairment is a situation where the patient has cognitive decline enough to not be categorised as cognitively normal, but the decline is not severe enough to be classified as Alzheimer's Disease.

The reason our machine learning models did not perform very well is due to the fact that early MCI patients had overlapping symptoms with cognitively normal patients and the late MCI patients had overlapping symptoms with AD patients. To add to this, the MCI cohort in ADNI also included converts, which refers to patients that converted from MCI to CN or AD over prolonged periods of time. This "converts" group became very difficult for us to incorporate into our dataset mainly because we were using baseline patient data (data from the first visit) as inputs in our model.

Overall, even though our model performance stood up to the available standards of diagnosis, we feel there is a lot of scope to dig deeper to analyse specific parameters - both clinical and technical, that would improve the performance of the model. The disease in itself is very complex and has multiple clinical factors that are minute and difficult to compile together as one textbook definition of Alzheimer's diagnosis.

7.4 Societal Impact

7.4.1 Ethical

Given the domain of this project, it was important that we considered ethical implications associated with a machine learning tool to diagnose patients. Since the goal of the tool was to accurately diagnose a patient, it was necessary that we communicated any reasoning behind a specific classification to the users of the tool. The confidence level conveyed through the pie chart and the list of important features used to determine the classification were the two ways we decided to promote transparency. It was also important that we communicated that the tool itself is not enough to diagnose a patient with Alzheimer's Disease. Rather, the tool's intended use is to support a general physician in his/her determination of a potential diagnosis. The home page in our prototype was used as a dedicated location to convey this information. For future iterations, a pop-up window before and after use of the tool could be another way to inform the tool's intended use to prevent conflicts regarding responsibility.

While curating our dataset and developing our model, we were also aware of the need to limit any biases that could be associated with a specific parameter. Model design was the focus of the project, which meant that we prioritized

working with data that was the most comparable. The North American consort of ADNI ultimately had the most complete data available, which is why we chose not to incorporate data from other global consorts. Fortunately, the advantage of machine learning models is that they can continue to learn from experience. In the future, our model can definitely scale and learn to consider more diverse populations as we incorporate studies from Japan, Australia, and other countries around the world.

Chapter 8

Conclusion

Throughout this project, our team learned new technical and interpersonal skills to design and build a solution to support the diagnosis process of Alzheimer's Disease in areas where neurologists may not be readily available. At the beginning of this project, the whole team had minimal knowledge of the initial signs of AD in patients, as well as the biological progression of AD. Through extensive research of machine learning concepts and user interface design, we were eventually able to develop a viable first-step solution for diagnosing patients with AD using machine learning techniques. While we are able to build an end-to-end solution with a user interface, there are still many revisions and iterations that can be made to our existing tool. We have just completed step one - future teams have the potential to develop a tool that is much more accurate and potentially deployable for user testing. Although there were some communication challenges due to the COVID-19 pandemic, we are nevertheless proud of what we have achieved this past year and for completing our project.

Chapter 9

Appendix

9.1 Team Approach

To account for gaps in knowledge regarding Alzheimer’s Disease and biological factors related to the disease, the students gained knowledge in June and July 2020 to understand and meet project requirements. All members of the team met periodically in the beginning stages of the project to ensure foundational understanding of the disease and requirements of the project. Once requirements and model features were finalized, the group split in January between backend and frontend to individually work on the project requirements. Throughout the project, the project management tool BaseCamp was used to streamline communication.

9.2 Development Timeline

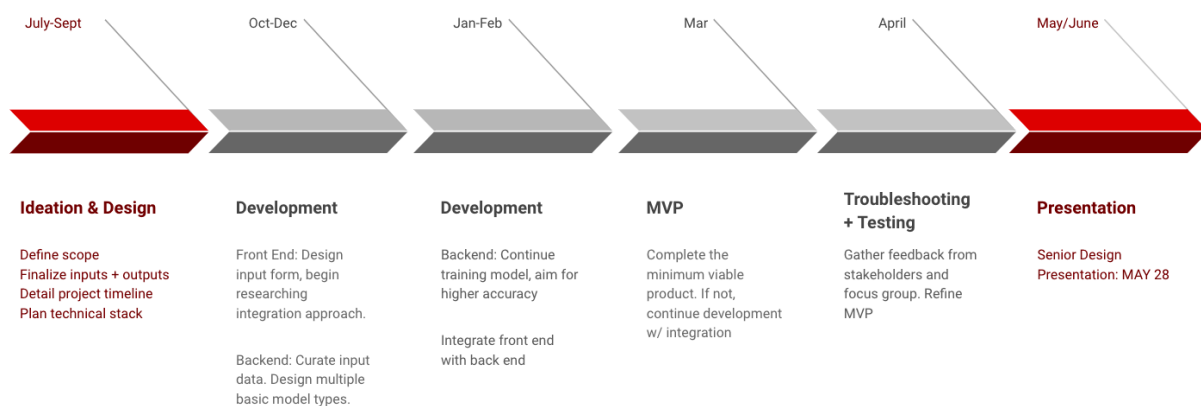


Figure 9.1: Development Timeline

The figure above is an estimated timeline for Alzheimer’s Disease Support Tool Team. Project starting date in July of 2020, extending to the beginning of June 2021.

Bibliography

- [1] Alzheimer's Association 2021 alzheimer's disease facts and figures. <https://www.alz.org/media/Documents/alzheimers-facts-and-figures-infographic.pdf>. Accessed: 2021-05-24.
- [2] World Health Organization dementia. <https://www.who.int/news-room/fact-sheets/detail/dementia>. Accessed: 2021-05-24.
- [3] Aiyushi Kumar Charles DeCarli. User experience research expert interview with uc davis alzheimer's disease center. personal communication, August 2020.
- [4] Margaret Gatz, Pia Svedberg, Nancy L. Pedersen, James A. Mortimer, Stig Berg, and Boo Johansson. Education and the risk of alzheimer's disease: Findings from the study of dementia in swedish twins. *The Journals of Gerontology: Series B*, 59:P34, 2004.
- [5] K. McRae-McKee, S. Evans, C. Hadjichrysanthou, and et al. Combining hippocampal volume metrics to better understand alzheimer's disease progression in at-risk individuals. *Scientific Reports*, 9(7499), 2019.
- [6] Michelle M. Mielke. Sex and gender differences in alzheimer's disease dementia. *The Psychiatric times*, 35(11):14–17, 2018.
- [7] Andreasen N, Hesse C, Davidsson P, and et al. Cerebrospinal fluid -amyloid. *Alzheimer Disease: Differences Between Early- and Late-Onset Alzheimer Disease and Stability During the Course of Disease. Arch Neurol*, 56(6):673–680, 1999.
- [8] Anderson NB, Bulatao RA, Cohen B, and editors. Ethnic differences in dementia and alzheimer's disease. *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, 4, 2004.
- [9] W. van der Flier and P. Scheltens. Hippocampal volume loss and alzheimer disease progression. *Nat Rev Neurol*, 5:361–362, 2009.
- [10] Zhenmei Zhang and et al. Marital status and risk of dementia: Does race matter? *Innovation in Aging*, 4:745–746, 2020.