Predicting Alzheimer's Disease: Development and Validation of Machine Learning Models

Abstract

Patients with Alzheimer's disease progressively lose their memory and thinking skills and, eventually, the ability to carry out simple daily tasks. The disease is irreversible, but early detection and treatment can slow down the disease progression. The purpose of this research is to use publicly available MRI data and demographic data from 373 MRI imaging sessions to build models to predict Alzheimer's in patients. Various machine learning models, including Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Neural Network, were developed. Data was divided into training and testing sets where training sets were used to build the predictive model, and testing sets were used to assess the accuracy of prediction. Key risk factors were identified, and various models were compared to come forward with the best prediction model. Among these models, the Random Forest model appeared to be the best model with an accuracy of 90.34%. This accuracy rate is higher than that of the average clinical diagnosis rate of Alzheimer's among all rural US doctors, 50-60%, and among all US doctors, 78%. MMSE, nWBV, and gender were the three most important contributing factors to detection of Alzheimer's. Among all the models used, the percent in which at least 4 of the 5 models shared the same diagnosis for a testing input was 90.42%. These machine learning models allow early detection of Alzheimer's with good accuracy, which ultimately leads to early treatment of these patients.

1. Introduction

Alzheimer's disease is currently ranked as the fourth leading cause of death in the United States, with approximately 65,800 fatalities attributable to the disease each year.[1] In fact, by early 2017, over 5.5 million people in the United States were diagnosed with Alzheimer's.[2] In the coming years, as more baby boomers reach and pass through old age, the number of Alzheimer's disease cases is expected to grow substantially.[3] In 2050, the number of Alzheimer's patients worldwide is expected to triple from 50 million (in 2018) to 152 million. [4] Furthermore, Alzheimer's disease is the most common form of dementia, which refers to the sustained deterioration of intellectual functions. About 60 to 80 percent of all dementia cases can be attributed to Alzheimer's.[2]

Alzheimer's disease is caused by the degeneration and eventual death of a large number of neurons in several areas of the brain. Alzheimer's is a very gradual disease, and it starts with short-term memory loss, followed by the progressive loss of memory and cognitive and intellectual functions. This eventually leads to deterioration of physical functioning and incapacitation.[1]

Despite the prevalence of Alzheimer's, especially among the elderly population, diagnosis for Alzheimer's remains a major challenge. Diagnosing Alzheimer's conclusively requires either an autopsy or brain biopsy.[3] However, autopsies can only be done after a patient's death, while brain biopsies are generally regarded as a procedure of last resort since they are costly, difficult to execute, and lengthy. Thus, most patients diagnosed with Alzheimer's are diagnosed through clinical diagnosis.[7] Studies have shown that community doctors in rural areas are only about 50 to 60 percent accurate in clinically

diagnosing Alzheimer's.[5] According to researchers from Keenan Research Center for Biomedical Science at St. Michael's Hospital, among more than 1,000 people listed in the National Alzheimer's Coordinating Center database, only 78% of the patients were accurately diagnosed by doctors.[6] One of the main reasons for inaccuracy of clinical diagnosis for Alzheimer's disease is that there are many other diseases with similar symptoms, like Parkinson's disease, diffuse white matter disease, and alcohol-associated dementia.[1]

Although Alzheimer's disease is irreversible, an early, accurate diagnosis can help doctors devise effective strategies to manage symptoms and plan for long-term care. Since treatment and care can influence how long one survives with Alzheimer's, patients who get diagnosed and treated at an early stage can potentially live longer and experience slower memory deterioration than those who begin later treatment.[1] In addition, an early diagnosis can significantly reduce treatment cost, since patients who begin treatment earlier routinely have access to more affordable options.[4]

The goal of this study is to create a machine learning model that can accurately diagnose Alzheimer's based on a patient's demographic and clinical data including magnetic resonance imaging (MRI) data. This model would allow physicians or nurses to diagnose patients accurately, without a brain biopsy. A machine learning model for diagnosing Alzheimer's will help patients at the early stages of Alzheimer's get better care and timely treatment. As a result, an effective model could potentially allow patients to live longer with slower memory impairment.

2. Data

2.1 Overview of Data

The Open Access Series of Imaging Studies (OASIS) data set of Longitudinal MRI in Nondemented and Demented Older Adults consists of 150 distinct individuals, ranging from 60 to 96 years old, as well as data from 373 MRI imaging sessions. These individuals were selected from a larger pool who participated in MRI studies at Washington University. Each individual had at least two visits, separated by at least a year, during which MRI and clinical data were obtained. Based on the Clinical Dementia Rating (CDR) scale, participants were classified as nondemented, demented, or converted (from nondemented to demented).

CDR is a dementia-staging tool that uses six domains to rate subjects for cognitive impairment. The six domains include memory, orientation, judgment and problem solving, role in community affairs, home and hobbies, and personal care. The CDR scale is from 0 to 3. The numbers equate to the following diagnoses: non-dementia (0), very mild/questionable dementia (0.5), mild dementia (1), moderate dementia (2), and severe dementia (3). Individuals who had a CDR of 0 for all visits were categorized as nondemented. Individuals who had a CDR of 0.5 or higher for all visits were categorized as demented. Individuals who had a CDR of 0 on the initial visit and a CDR of at least 0.5 on any subsequent visits were categorized as converted. Of the 150 individuals in the data set, only 14 individuals were categorized as converted.

For each visit, MRI biomarkers, clinical data, and demographic data were

collected, including estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), atlas scaling factor (ASF), Mini Mental State Examination (MMSE) result, age, gender, socioeconomic status, and years of education received. This study uses machine learning models to predict either demented (indicating the patient has Alzheimer's) or nondemented (indicating that the patient does not have Alzheimer's) based on the MRI, clinical, and demographic data obtained on each visit.[8]

2.2 MRI Data

MRI produces detailed images of the brain by using radio waves and a strong magnetic field. Although it does not provide a definitive diagnosis, MRI-based measures of the brain have been regarded as valid markers indicating Alzheimer's progression. MRI can also help diagnose diseases that are commonly mistaken for Alzheimer's due similarity in symptoms.[9] Rather than using the raw image data from the MRI scan, the machine learning models in this study use biomarkers from the MRI data, such as eTIV, nWBV, and ASF.

The estimated total intracranial volume (eTIV) is the volume of the cranial cavity (the space inside the skull) taken from an MRI. It is a standard measure that is used to correct for head size variation across subjects in brain studies, including Alzheimer's-related studies.[10] In the data set used in this study, eTIV was estimated using the software FreeSurfer and was calculated by dividing a predetermined constant by the atlas scaling factor (ASF), the value that the MRI image is scaled by to align to the MNI305 head atlas, a brain mapping template for MRI scans. Thus, ASF is directly proportional to eTIV.[8] Since they are directly proportional to each other (high collinearity), only eTIV will be used in this study

to train and test the machine learning models.

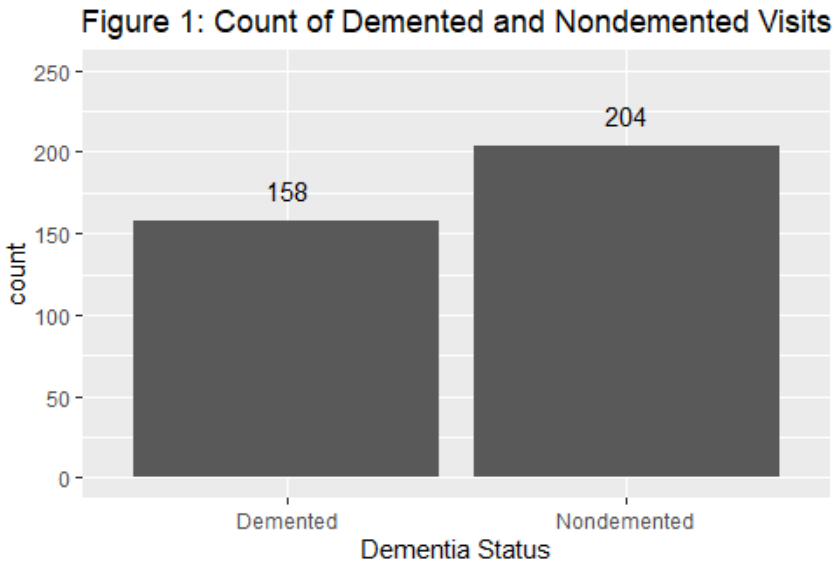
Normalized whole-brain volume (nWBV) refers to the proportion of tissue in the brain volume. In the data set, nWBV was evaluated using the FAST program. First, the MRI image was segmented to classify brain tissue as cerebral spinal fluid, gray matter, or white matter. The segmentation procedure assigned voxels (which are essentially 3D pixels) to the tissue classes using the Markov random field model. The nWBV was then evaluated as the proportion of all voxels categorized as tissue. Previous studies have shown that nWBV atrophies at a much greater rate in patients with early Alzheimer's.[13]

2.3 Cognitive Assessment Data

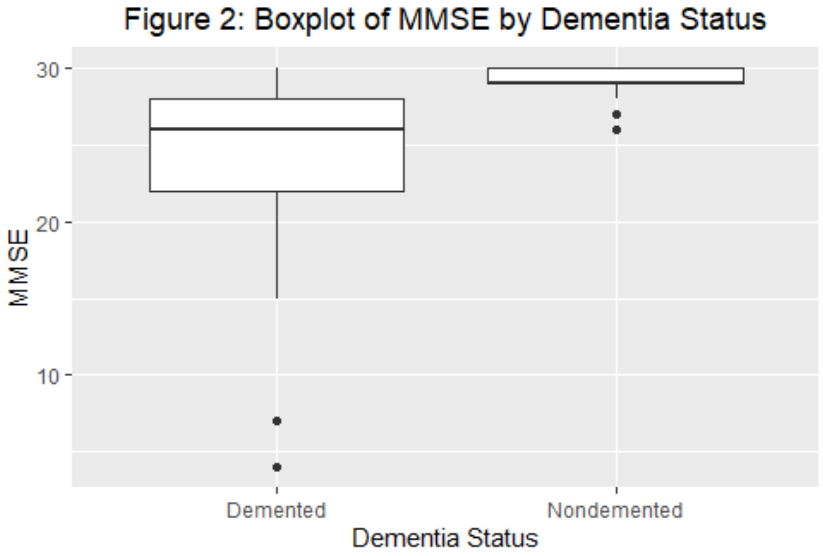
Each patient in this study received a Mini Mental State Examination (MMSE). MMSE is the most widely used assessment for the evaluation of an individual's cognitive state. MMSE is a questionnaire that takes about 5 to 10 minutes to complete, which makes it a fast and easy way to evaluate cognitive state. MMSE is based on a 30-point score. A lower score indicates a greater degree of cognitive impairment. A score between 20 to 24 is associated with mild dementia, 13 to 20 is associated with moderate dementia, and below 12 is associated with severe dementia. MMSE tests six different aspects of cognitive ability applicable to Alzheimer's, including orientation of time and place, short-term memory recall, immediate recall, language, simple math calculation ability, and the ability to create a simple figure. In other words, MMSE tests various, everyday mental skills.[11]

2.4 Exploratory Data Analysis

In the OASIS data set of Longitudinal MRI in Nondemented and Demented Older Adults, there are 150 individual participants who are categorized as demented, nondemented, or converted. Since the machine learning models in this study will be designed to predict either demented or nondemented (binary outcome), the first visit of each converted participant will be classified as nondemented and the last visit of each converted patient will be classified as demented. For converted patients, intermediate visits (only 9 visits out of 373) are disregarded from the data. In addition, there are two subjects with two visits having missing MMSE data. These 2 visits will also be disregarded from the data in this study. Figure 1 provides a histogram representing the number of visits that are categorized as demented and nondemented. There are 158 visits categorized as demented and 204 visits categorized as nondemented.

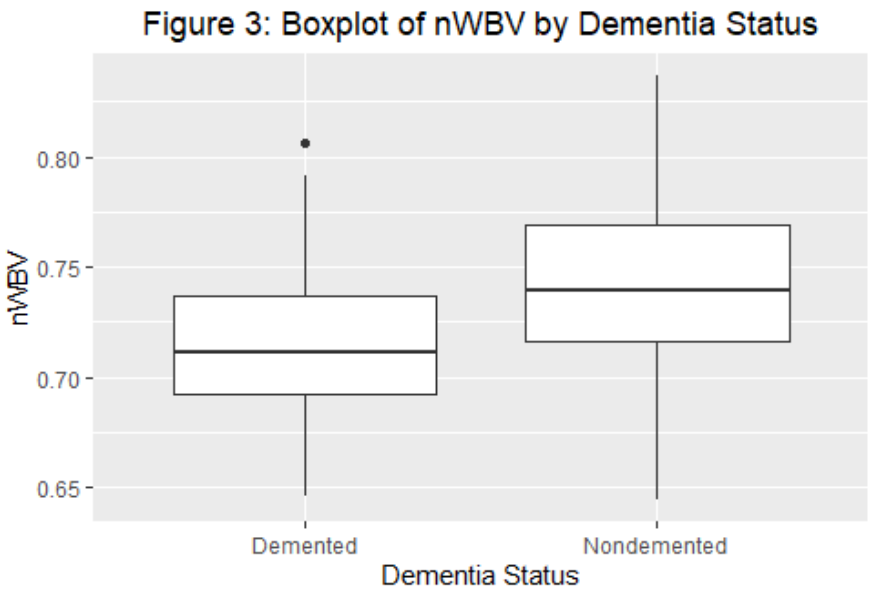


MMSE is a cognitive assessment that is used to measure an individual's cognitive impairment. Lower scores of MMSE indicate higher degrees of cognitive impairment. Figure 2 clearly shows that demented individuals tend to have lower MMSE scores than nondemented individuals. The median MMSE for nondemented individuals, 29, is approximately 11.5% higher than the median MMSE for demented individuals, 26. Furthermore, the first quartile MMSE for nondemented individuals, 29, is approximately 32% higher than the first quartile MMSE for dementia individuals, 22. Demented subjects have a large variation in MMSE scores compared with non-demented subjects in Figure 2.

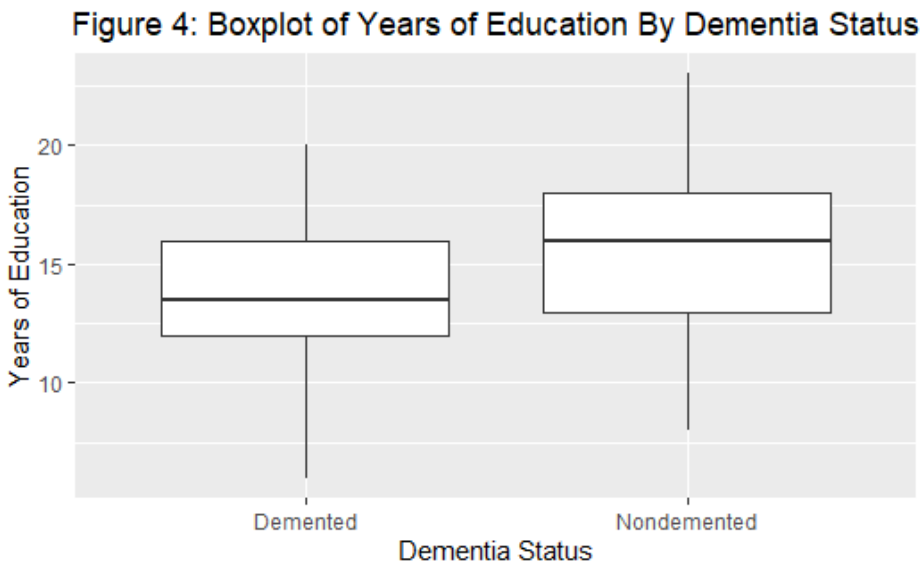


nWBV is calculated as the proportion of tissue in the brain volume. In this data set, nondemented individuals had a median nWBV of 0.7390, which is approximately 3% higher than the median nWBV for demented individuals, 0.711.

As seen in Figure 3, nondemented individuals generally had higher nWBV than demented individuals.

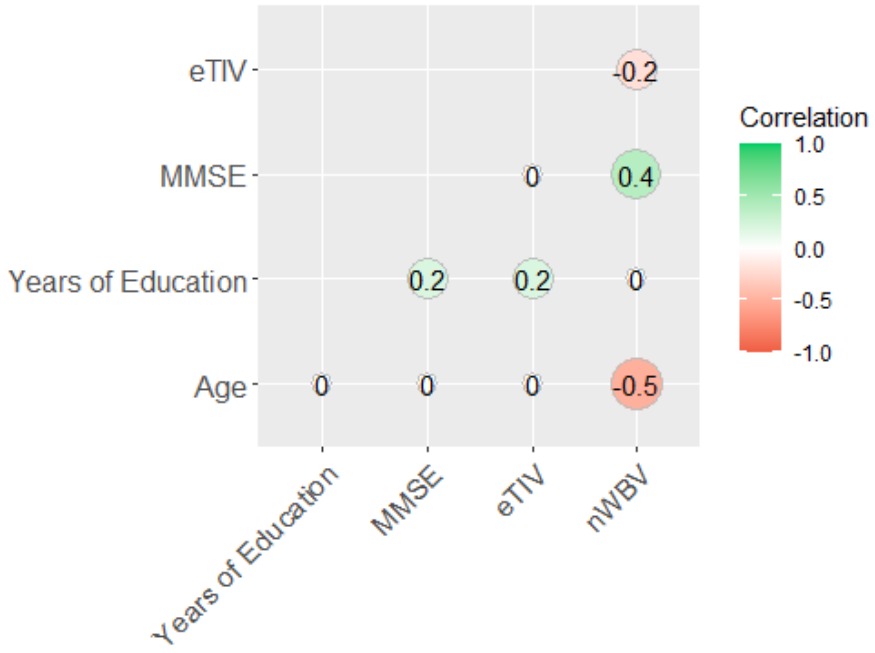


In addition to clinical data, the machine learning models in this study also use demographic data to help diagnose Alzheimer's. In the data set used in this study, years of education correlate with the status of dementia. Individuals with more years of educations were less likely to be categorized as demented. As shown in Figure 4, the median years of education for nondemented individuals, 16, was 19% higher than the median years of education for demented individuals, 13.5.



Among all the numerical continuous variables used in this study, a correlogram was created, as shown in Figure 5. A correlation of 0 indicates that there is no correlation between the variables. Correlations greater than 0 indicate positive correlations, while correlations closer to 1 indicate higher degrees of positive correlation. Correlations less than 0 indicate negative correlations, while correlations closer to -1 indicate higher degrees of negative correlation. In this data set, nWBV and MMSE have a positive correlation of 0.4, and Age and nWBV have a negative correlation of -0.5. MMSE and Years of Education, as well as eTIV and Years of Education both exhibit slight positive correlations of 0.2. Finally, nWBV and eTIV have a slight negative correlation of 0.2. Other relationships between the continuous variables have little correlation in Figure 5.

Figure 5: Correlogram of Continuous Variables



3. Machine Learning Models

3.1 Logistic Regression

Logistic regression is one of the most widely used machine learning algorithms for solving a classification problem. It is used to predict the probability of a particular outcome, given a set of independent variables, which can be continuous, discrete, dichotomous, or a combination of these. In this study, the dependent variable used in the logistic regression model is the probability of dementia, and the independent variables include years of education, age, gender, eTIV, nWBV, and MMSE. Logistic regression is modeled by the equation below, where P is the probability of dementia, $\boldsymbol{\beta}_0$ is the intercept, $\boldsymbol{\beta}_1$ through $\boldsymbol{\beta}_n$ are the regression coefficients, and x_1 through x_n are the independent variables.

Equation 1

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

The left side of the equation is referred to as "the logit". The interpretation of the coefficients describes the independent variable's effect on the logit, rather than directly on the probability *P*. To facilitate interpretation, e^{β_n}, a transformation of the original regression coefficient β_n, can be derived and interpreted as follows:

If $e^{\beta_n} > 1$, $P / (1 - P)$ increases.

If $e^{\beta_n} < 1$, $P / (1 - P)$ decreases.

If $e^{\beta_n} = 1$, $P / (1 - P)$ stays the same.

To build the logistic regression model for this study, 60% of the data were randomly chosen for model development (204 observations), while the remaining 40% of the data (158 observations) were used to test the model. The results of the logistic regression analyses are displayed in Table 1.

Table 1: Logistic Regression Analyses of Clinical MRI Data

Predictor	**Estimate (β)**	**Standard Error**	***P***	**Odds Ratio (Exp(β))**
Intercept	80.1293	15.2464	<0.001	NA
nWBV	-31.1353	9.8578	0.0016	3.007e-14
eTIV	-0.0038	0.0018	0.0341	0.9962
Age	-0.1200	0.0451	0.0079	0.8869
MMSE	-1.3874	0.2386	<0.001	0.2497

Years of Education	-0.2594	0.0959	0.0068	0.7715
Gender	1.4993	0.6039	0.013	4.4785

Above, the "Estimate" refers to the coefficient $\boldsymbol{\beta}$. *P* values under 0.05 indicate that the predictor is statistically significant; lower *P* values indicate higher statistical significance. Thus, in this particular model, every predictors is considered statistically significant. The Odds Ratio is obtained by evaluating the natural exponential of $\boldsymbol{\beta}$. According to the logistic analysis results listed in Figure 7, at a significance level of 0.05, the predictive model for Alzheimer's diagnosis is:

Predicted logit of dementia = 80.1293 − 31.1353 × nWBV − 0.0038 × eTIV − 0.1200 × Age − 1.3874 × MMSE − 0.2594 × Years of Education + 1.4993 × Gender

The coefficients of the parameters were interpreted as follows. At the significance level of 0.05:

- On average, controlling other variables, for 1 unit increase in nWBV, the odds of being demented is decreased by nearly 100%.
- On average, controlling other variables, for 1 unit increase in eTIV, the odds of being demented is decreased by 0.38%.
- On average, controlling other variables, for 1 year increase in age, the odds of being demented is decreased by 11.31%. (Note that all ages are between 60 and 96.)

- On average, controlling other variables, for 1 unit increase in MMSE, the odds of being demented is decreased by 75.03%.
- On average, controlling other variables, for 1 year increase in Years of Education, the odds of being demented is decreased by 22.85%.
- On average, controlling other variables, female subjects are 347.85% more likely to be demented.

The model indicates that nWBV, eTIV, age, MMSE, years of education, and gender are significant predictors of Alzheimer's. Being female increases the odds of being demented, whereas having a higher eTIV, nWBV, a higher age, a higher MMSE score, and more years of education decreases the odds of being demented. In this particular model, it seems that an increase in age decreases the odds of being demented, which is contradictory to the general association that being older is correlated with a higher risk of Alzheimer's. This could be just a coincidence of the individuals who participated in the data set.

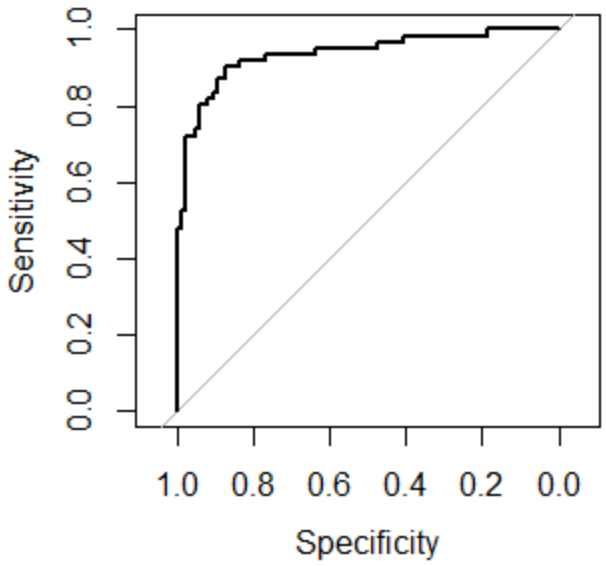
After the logistic regression model was analyzed, the model was tested with the testing data, which contained 158 observations. Of the 158 observations, 114 observations were of nondemented individuals, and 54 observations were of demented individuals. As shown in the confusion matrix (Table 2), 43 of the 54 observations of demented individuals were accurately predicted, resulting in a sensitivity of 79.63%. 94 of the 114 observations of nondemented individuals were accurately predicted, resulting in a 90.38% specificity. The overall accuracy rate of the logistic regression model is 86.71%.

Table 2: Confusion Matrix of Logistic Regression Model Test

Actual Diagnosis	Predicted as Demented	Predicted as Nondemented	% Correct
Demented	43	11	79.63%
Nondemented	10	94	90.38%
		Overall % Correct:	**86.71%**

In addition to a confusion matrix, an ROC curve plot (Figure 6) was created to visualize the performance of the logistic regression model on the testing data. The X axis shows the specificity or the proportion of observations that are actually demented that are correctly predicted as demented. The Y axis shows the sensitivity, or the proportion of observations that are actually nondemented that are correctly predicted as nondemented.

Figure 6: ROC Curve for Logistic Regression Model



3.2 K Nearest Neighbor

K Nearest Neighbor (KNN) is a supervised machine learning algorithm that classifies a new data point based on its neighboring data points' features. KNN is a lazy algorithm, which means it memorizes the training data set rather than learning a discriminative function from the training data. It is also a non-parametric model, which means that it doesn't make any assumptions about the data set, thus making it more effective at handling real world data.

In this study, a KNN algorithm is used to determine whether an individual is demented based on independent variables, including the individual's years of education, age, gender, eTIV, nWBV, and MMSE. Each visit in the data set is considered a data point, and each data point is plotted in *n*-dimensional space (where *n* is the number of independent variables), with the value of each independent variable being the value of a particular coordinate. Each data point is also categorized in a class, either demented or nondemented.

A KNN algorithm determines what class a new data point is by finding what class the majority of the *K* nearest data points are. The proximity between data points is calculated by using the Euclidean distance formula. As shown below, in the Euclidean distance formula, q_1 through q_n are the independent variables for data point *q*. Likewise, p_1 through p_n are the independent variables for data point *p*. The distance between data points *q* and *p* is given by Equation 2 below.

Equation 2

$$d = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

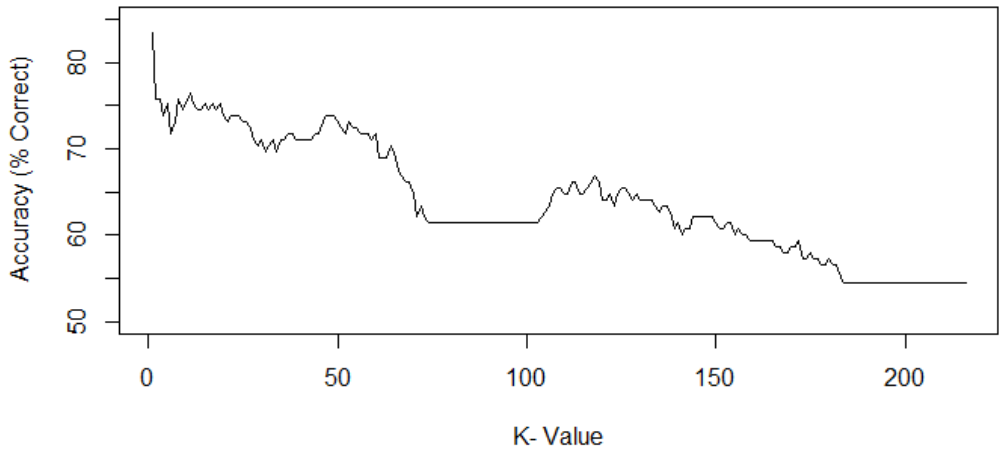
The independent variables in the data set used in this study have different magnitudes. For example, MMSE is calculated as a whole number between 1 and 30, whereas nWBV is a decimal number between 0 and 1. Thus, the data is normalized in order to create a level playing field for all the variables in the data set. Below is the formula used for normalizing values in the data set, where *v* is the original value, *max* is the maximum of the values, *min* is the minimum of the values, and *t* is the normalized value given by Equation 3 below.

Equation 3

$$t = \frac{v - min}{\max - min}$$

After normalization, all the transformed values were between 0 and 1. 60% of the data set of transformed values (217 observations) was randomly chosen for model development, and the remaining 40% of the data set (145 observations) was used to test the model. In order to optimize the accuracy of the KNN model, all possible values of K (from 1 to 216) were tested. Below is an accuracy plot showing the accuracy of the KNN model for values of K from 1 to 216.

Figure 7: Accuracy Plot of KNN Algorithm



As shown in Figure 7, the *K* value with the highest accuracy is 1. Thus, obtaining a new data point's class based on the class of its nearest data point yields the highest accuracy rate. A confusion matrix of the testing data tested with a KNN model with *K* value of 1 is shown in Table 3. Of the 145 observations in the testing data, 87 observations were of nondemented individuals, and 58 observations were of demented individuals. As shown below, 50 of the 58 observations of demented individuals were accurately predicted, resulting in a 86.21% sensitivity. 71 of the 87 observations of nondemented individuals were accurately predicted, resulting in a 91.61% specificity. The overall accuracy rate of the KNN model with *K* value of 1 is 83.45%.

Table 3: Confusion Matrix of KNN Model with K Value of 1

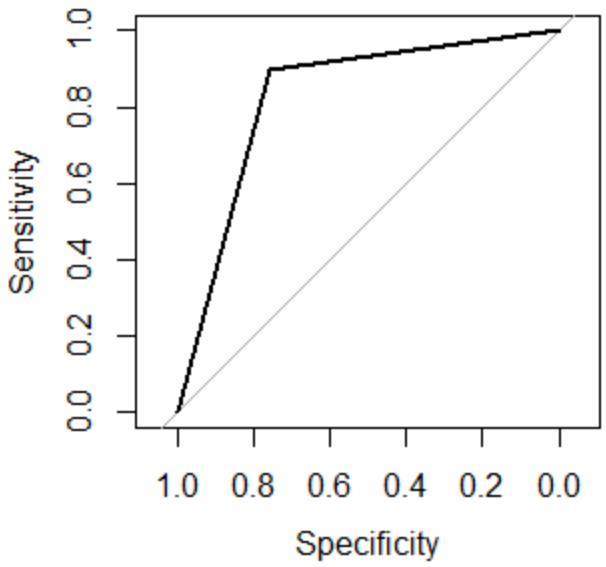
Actual Class	Predicted Demented	Predicted Nondemented	% Correct

Demented	50	8	86.21%
Nondemented	16	71	81.61%

Overall % Correct is **83.45%**

Like the logistic regression model, the performance of the KNN Model with *K* value of 1 can be visualized in a ROC curve plot, in which the x axis shows specificity and the y axis shows sensitivity, as shown in Figure 8.

Figure 8: ROC Curve of KNN Model with K Value of 1



3.3 Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm that is commonly used in classification models. In this study, SVM is

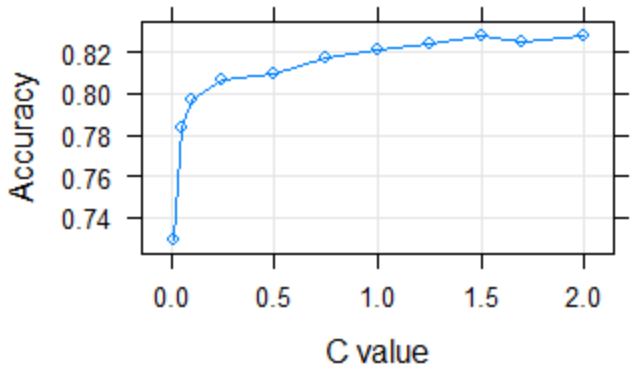
used to classify individuals as either demented or nondemented based on independent variables, including the individual's years of education, age, gender, eTIV, nWBV, and MMSE. Each observation in the data set is plotted as a point in *n*-dimensional space (where *n* is the number of independent variables) ,with the value of each variable being the value of a particular coordinate. Each data point is also categorized as a class (which is the dependent variable being predicted), demented or nondemented. Classification is then performed by finding the hyperplane (a subspace whose dimension is *n* - 1) that segregates the classes (demented and nondemented) in the best possible way.

If the training data is linearly separable, a hyperplane can be selected that best separates the two classes of data so that the distance (calculated by the Euclidean distance formula) between the hyperplane's two nearest distinct data points (known as the margin) is as large as possible. This is known as "hard-margin classification", in which no data points are allowed inside the margin. This type of classification is too stringent and sensitive to outliers. On the other hand, "soft-margin classification" allows certain data points to be inside the margin. The *c* value is used to tune how many data points are allowed inside the margin. If there are too many data points in the margin, the margin is too simple and does not adequately capture the underlying structure of the data. However, if there are too few or no data points in the margin, the separation may be influenced too greatly by the noise of the training data. Although the separation might be optimal from the training data, it would generalize poorly. As a result, the separation would be suboptimal for unseen data (eg. the testing data). The *c* value is defined as the weight of how much the samples inside the margin contribute to the overall value. With a low *c* value, samples inside the margins are penalized less than with a

higher *c*. In other words, a lower *c* value allows for more data points in the margin than a higher *c* value. With a *c* value of 0, samples inside the margin are not penalized at all. An infinite *c* value is essentially hard margin classification, where no data points are allowed in the margin.

Like in previous models, 60% of the data (217 observations) was chosen for model development, and the remaining 40% of the data (145 observations) was used for testing the model. Multiple SVM models were created using the training data, with the following *c* values: 0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.7, and 2. The accuracy of each c value was evaluated using 10 fold cross-sampling validation, in which 90% of the training data was randomly selected and used to train the model, while the remaining 10% of the training data was used to validate the model. Note that this is not the final accuracy of the model, which will later be evaluated based on the model's performance on the testing data (40% of the original data). Figure 9 shows the accuracy for each *c* value.

Figure 9: Accuracy of SVM Based on C Value



As shown in Figure 9, the *c* value with the highest accuracy is 1.5. The SVM model with *c* value 1.5 was then tested with the testing data (40% of the

original data). Table 4 below provides a confusion matrix indicating the performance of the model on the testing data. Of the 67 demented individuals, 54 were correctly predicted as demented, resulting in a sensitivity of 80.60%. Of the 78 nondemented individuals, 73 were correctly predicted as nondemented, resulting in a sensitivity of 93.59%. Overall, there was 87.59% accuracy when the SVM model with *c* value of 1.5 was evaluated using the testing data.

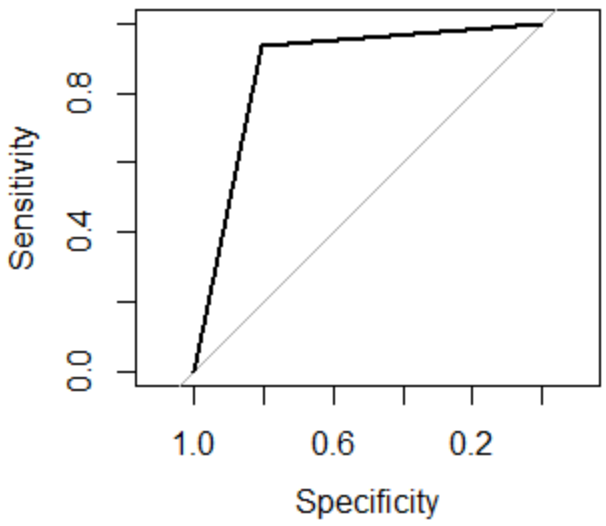
Table 4: Confusion Matrix of SVM Model with *c* value of 1.5

Actual Class	Predicted Demented	Predicted Nondemented	% Correct
Demented	54	13	80.60%
Nondemented	5	73	93.59%

Overall Accuracy: **87.59%**

Like previous models, the performance of the SVM model with *c* value of 1 can be visualized in a ROC curve plot. The x axis shows specificity, and the y axis shows sensitivity (Figure 10).

Figure 10: ROC Plot of SVM Model with *c* value of 1.5



3.4 Random Forest

The random forest algorithm is a supervised machine learning algorithm primarily used for classification. In this study, the random forest algorithm is used to classify individuals as demented or nondemented based on indepedent variables, including the individual's years of education, age, gender, eTIV, nWBV, and MMSE. Essentially, the random forest algorithm builds multiple decision trees (called a forest) and combines them to produce an accurate and stable prediction.

A decision tree is a tree-like model of decisions and their possible consequences. Each node in a decision tree represents a "test" on an independent variable (eg. a person's MMSE score is under 27), and each branch represents the result of the test. Each branch connects from the parent node (the node containing the test) to one of its child nodes, which either contains another test or is a leaf node. The leaf nodes (nodes without child nodes) represent the final classification result (in this study, demented or nondemented). In a decision tree, a decision is made by starting from the root node (the node without parent nodes) and

descending until a leaf node, which contains the final decision, is reached. A decision tree can also be referred to as a "tree."

The random forest algorithm contains a large number of decision trees that operate as an ensemble. Each individual tree generates a decision containing the class prediction (in this study, demented or nondemented), and the most popular class prediction among all the trees is used as the final prediction of the algorithm.

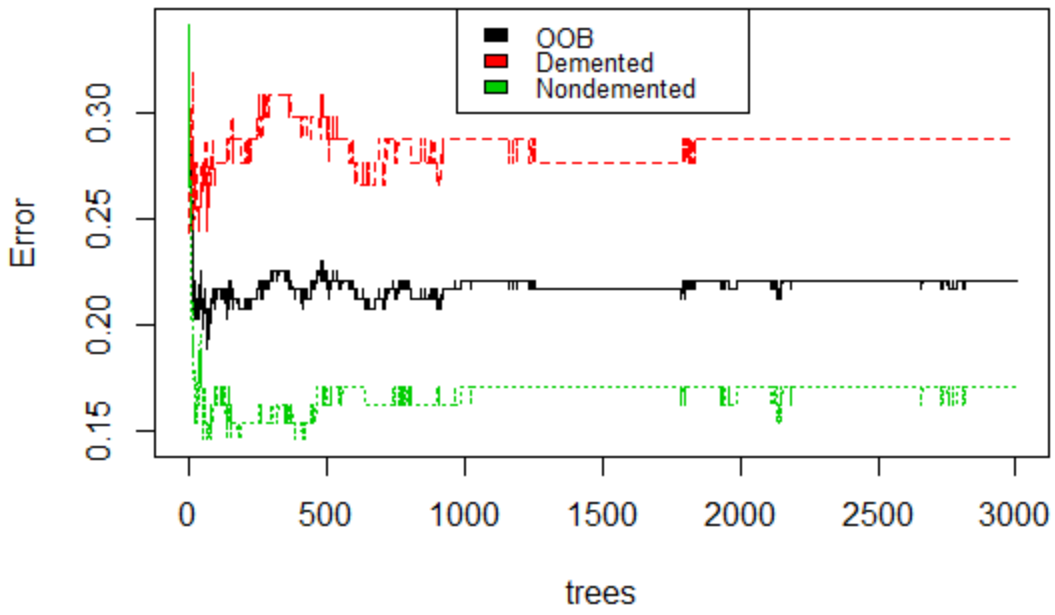
Before creating the random forest model, 60% of the data (217 observations) was chosen for model development (referred to as training data), and the remaining 40% of the data (145 observations) was used for testing the model.

The first step in creating a random forest model is to create a bootstrapped data set. To create a bootstrapped data set, observations from the training data set were randomly selected. Note that observations in the bootstrapped data set can be repeated if selected multiple times. The second step is to build a decision tree based on the bootstrapped data set. Steps 1 and 2 are then repeated for each decision tree in the random forest model. Note that one third of the training data are left out of the bootstrap samples and are therefore not used to construct the trees. In this study, exactly 2,001 decision trees were built for the random forest model. Each decision tree was constructed using two independent variables, which were selected at random from the total of six independent variables.

Figure 11 shows the Out-of-bag (OOB) error and misclassification error rates for each independent variable based on the number of trees in the random forest model. The OOB error is a way of validating the random forest model. A higher OOB error indicates a higher prediction error. The misclassification error rate refers to the proportion of trees that misclassify a particular class. The OOB error and misclassification error rates are estimated using the one-third of the

training data not used in the bootstrapping samples to test the model. The black color represents the OOB error, the red color represents the misclassification error for demented individuals, and the green color represents the misclassification error for nondemented individuals.

Figure 11: OOB and Misclassification Error Based on Number of Trees in Random Forest Model

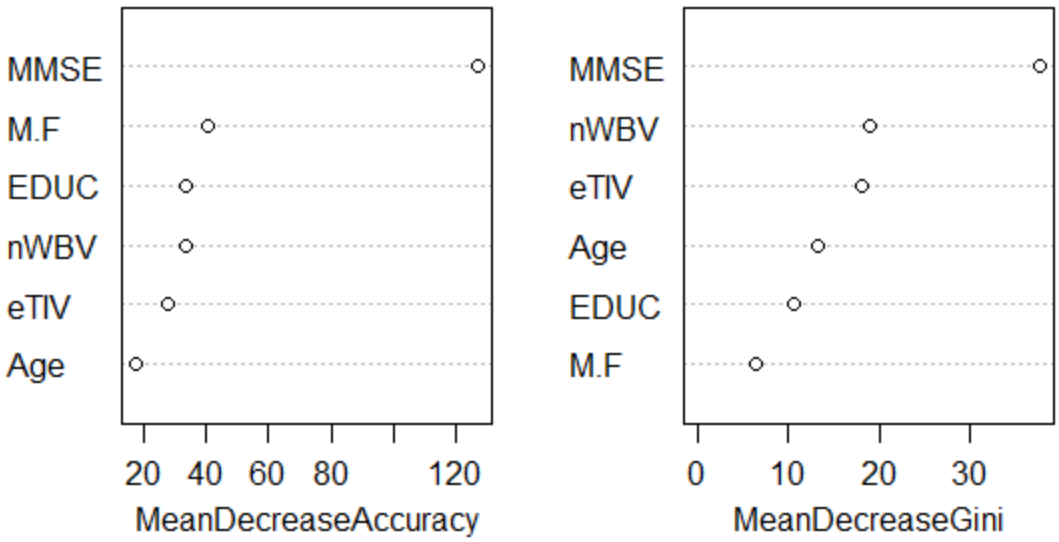


As shown in Figure 11, once roughly 1,000 trees have been generated, the OOB and misclassification error rates stay relatively constant. In other words, once there are 1,000 trees in the random forest model, increasing the number of trees in the model does little to reduce the error rates. Figure 12 shows the mean decrease accuracy and mean decrease GINI Impurity (GINI) for each of the

independent variables. The mean decrease accuracy estimates the loss in prediction performance when a particular variable is dropped from the training data. A higher mean decrease accuracy indicates greater loss in prediction performance when a variable is omitted.

The mean decrease GINI measures the variable's importance for estimating whether an individual is demented or nondemented. An independent variable's mean decrease GINI is based on the average decrease of the impurity of nodes that test the independent variable. A node's impurity is the probability of obtaining two different decisions in each of the node's subtrees (trees that are children of the node). For example, if each subtree both have a 50-50 chance of outputting demented or nondemented, this indicates that the node has high impurity. However, if one subtree mostly outputs demented while the other subtree mostly outputs nondemented, the node has low impurity. Variables with lower mean impurity among nodes testing that variable have a higher mean decrease GINI.

Figure 12: Mean Decrease Accuracy and Mean Decrease GINI of Independent Variables in Random Forest Model



Mean decrease accuracy and mean decrease GINI both measure the importance of each variable to the model. Higher values of mean decrease accuracy and mean decrease GINI both indicate that the variable is more important to the model. As shown in Figure 12, MMSE was overwhelmingly the most important independent variable to the model in both the mean decrease accuracy and mean decrease GINI plots. Gender (referred to as M.F in Figure 12) has the second-highest mean decrease accuracy but the lowest mean decrease GINI. Years of Education (referred to as EDUC in Figure 12) has the third highest mean decrease accuracy but the second lowest mean decrease GINI. nWBV has the fourth highest mean decrease accuracy and the second highest mean decrease GINI. eTIV has the second lowest mean decrease accuracy but the third highest mean decrease GINI. Lastly, based on mean decrease accuracy, age is the least important variable, but based on mean decrease GINI, it is still more important

than years of education and gender. Overall, although the importance of gender, years of education, nWBV, eTIV, and age varies based on mean decrease accuracy and mean decrease GINI, MMSE remains by far the model's most important variable.

Once the random forest model was developed, it was then evaluated using the testing data set. A confusion matrix (see Table 5) was created to analyze the performance of the model. Of the 64 demented individuals, 54 were correctly predicted as demented, resulting in a sensitivity of 84.38%. Of the 81 nondemented individuals, 77 were correctly predicted as nondemented, resulting in a sensitivity of 95.06%. Overall, there was 90.34% accuracy when the random forest model was evaluated with the testing data.

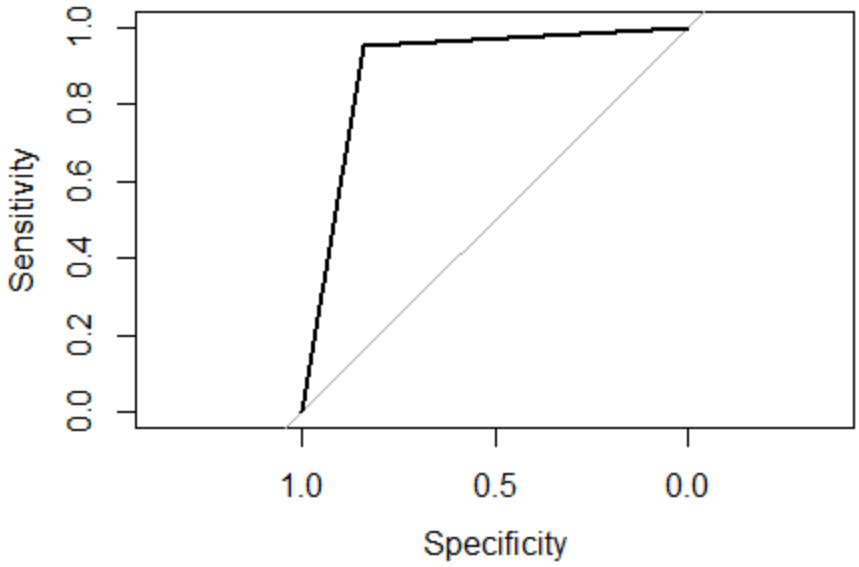
Table 5: Confusion Matrix of Random Forest Model

Actual Class	Predicted Demented	Predicted Nondemented	% Correct
Demented	54	10	84.38%
Nondemented	4	77	95.06%

Overall Accuracy: **90.34%**

Like previous models, the performance of the random forest model can be visualized in a ROC curve plot. The x axis shows specificity, and the y axis shows sensitivity (Figure 13).

Figure 13: ROC Plot of Random Forest Model



3.5 Neural Network

Finally, the last machine learning method used in this study is a neural network. In this study, neural networks are multi-layer networks of neurons that are used to classify demented or nondemented individuals based on clinical and demographic independent variables. Every neural network has an input layer, at least one hidden layer, and an output layer. The input layer, the first layer, takes inputs based on the existing data. In this study, the input layer takes in the attributes for these independent variables: years of education, age, gender, eTIV, nWBV, and MMSE. The nodes in the hidden layer receive inputs from the input layer nodes, perform some computation, and then provide the output to the output layer. The output layer node contains the prediction of whether the individual is demented or nondemented based on the original input attributes.

Each of the nodes of each layer links to the nodes of the next layer through connections. Each connection has a weight, and each neuron (another term for node) has a bias and an activation function. The output of each node in the hidden layers and output layer can be represented by the following method. Let x_1 through x_i represent the input values of each node in the input layer. Let w_1 through w_i represent the weights of each connection going to the current node. Let m represent the number of connections that go from the activated nodes of the prior layer to the current node (or the number of nodes in the input layer if the prior layer is the input layer). Let b represent the bias of the current node. Below is the equation to obtain the result z for a particular node. Note that z is not the final output of the node.

Equation 4

$$z = \sum_{i=1}^{m} w_i x_i + b$$

After obtaining z, the final output of the node is simply determined by the activation function. In this study, the sigmoid function is used as the activation function. Below is the sigmoid function, where $f(z)$is the final output of the node with result z.

Equation 5

$$f(z) = \frac{1}{1+e^{-z}}$$

The output layer node has an output $f(z)$that is a value between 0 and 1. If $f(z) \geq 0.5$,the final output of the neural network is 1, indicating that the individual is demented. If $f(z) < 0.5$, the final output of the neural network is 0, indicating that the individual is nondemented.

As in previous models, the data was randomly split into a training data set (containing 217 observations) and a testing data set (containing 145 observations). Before creating the neural network, the data was normalized in order to create a level playing field for all variables. Without data normalization, an independent variable may have a large impact on the dependent variable because of its scale, rather than its actual importance. The min-max normalization technique was used to normalize the data in this study, which was described in section 3.2.

Initially, each connection in the neural network was assigned random weights. For each observation, the input values were inputted into the input layer of the neural network and the final output was calculated. This is known as forward propagation. For each observation, after the forward propagation, backward propagation (which is essentially reinforcement learning) is used to modify the weights and biases in order to minimize the cost function. Backward propagation starts from the output layers and moves through the model until it reaches the starting layer. In backward propagation, the error attributable to each neuron (as determined by the cost function), is calculated, starting from the layer closest to the output all the way back to the starting layer of the model. The cost function used in this study is cross-entropy, as shown below. Note that the cost function is calculated for each neuron. C represents the cost, in which the closer C is to 0, the closer the neuron's output is to the desired output for the neural network. X is the sum of all the input values. N is the total number of input nodes. y is the desired output for the particular neuron. a is the actual final output of the particular neuron.

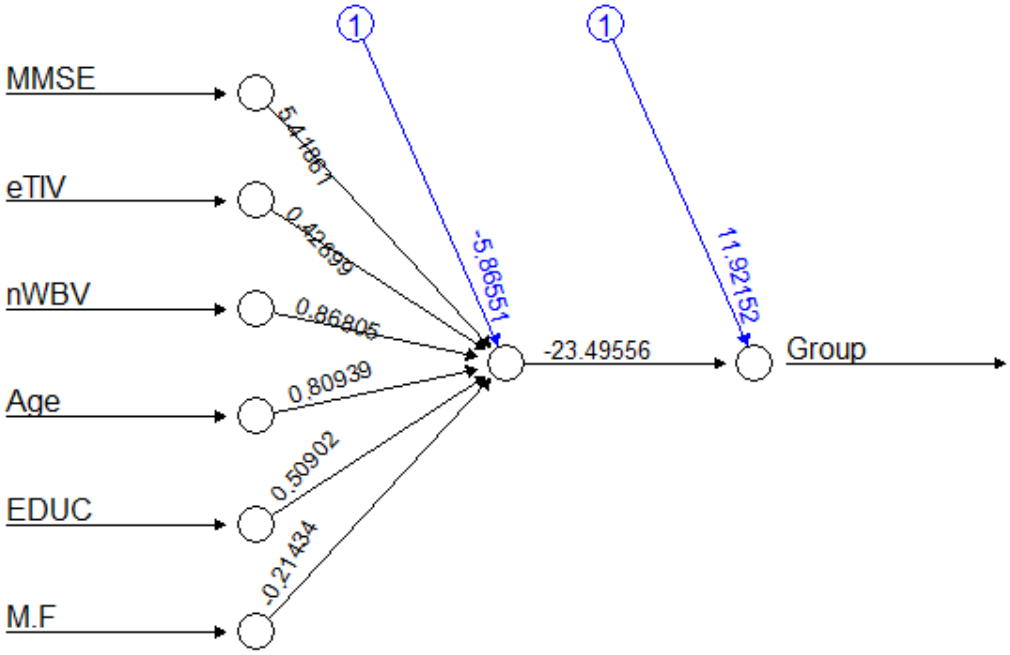
Equation 6:

$$C = -\frac{1}{n} X[y \ln a + (1-y)\ln(1-a)]$$

Once the error attributable to each neuron is calculated, the bias and weight of the activated neurons are tweaked by the backpropagation process in a way that minimizes the overall cost function for each neuron.

Once forward propagation and backward propagation is completed for each of the observations in the training data, the neural network training is complete. In this study, neural networks of one hidden layer neuron and two hidden layer neurons were created. Below is a diagram representing the first neural network created, which contained only one hidden layer neuron.

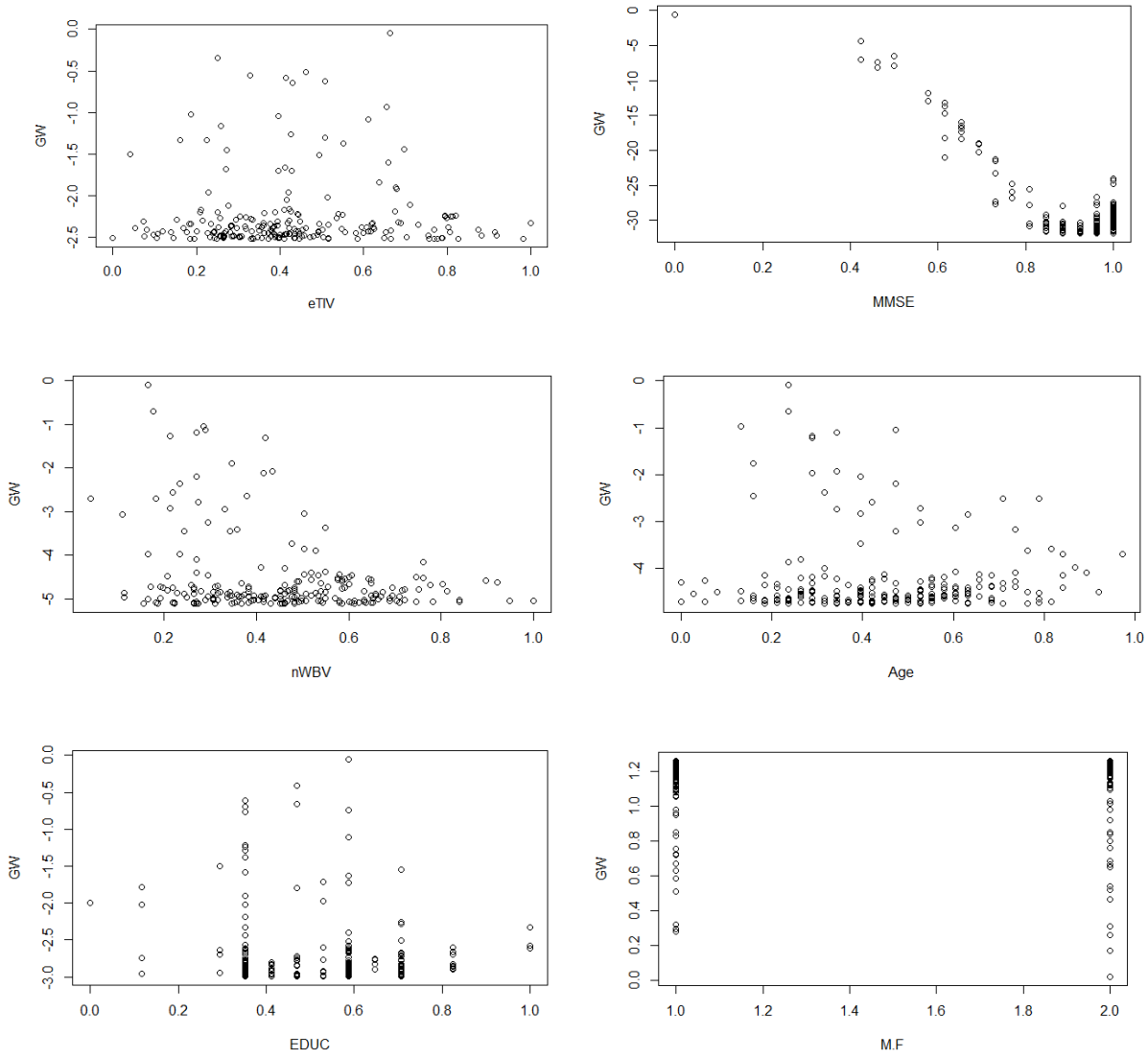
Figure 14: Plot of Neural Network with 1 Hidden Layer Node



Above, the black lines represent the connections between the nodes. A number, representing the weight, is assigned to each connection. The blue numbers indicate the bias of the particular neuron. It took the neural network

3,437 steps to converge, or reach a state in which the neural network has learned to properly predict an individual's dementia state with some margin of error. Below is a plot of the generalized weights with respect to each independent variable for the neural network with 1 hidden layer node.

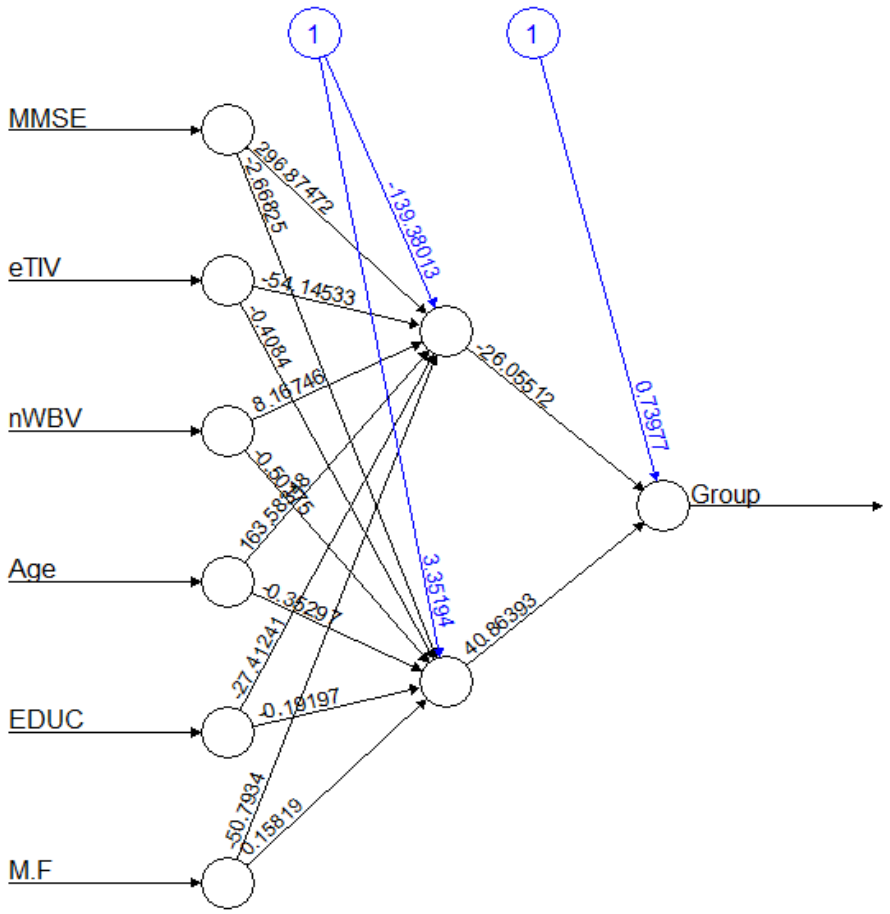
Figure 15: Plot of Generalized Weights with Respect to each Independent Variable for Neural Network with 1 Hidden Layer Node



"EDUC" and "M.F" refer to years of education and gender respectively.

In Figure 15, the distribution of generalized weights suggests that all the independent variables appear to have a nonlinear effect since the variance of their generalized weights is overall greater than one. The second neural network created in this study uses two hidden layer nodes. It uses the same activation function and cost function as the first neural network. Figure 16 shows a diagram representing the weights and biases of the neural network with two hidden layer nodes.

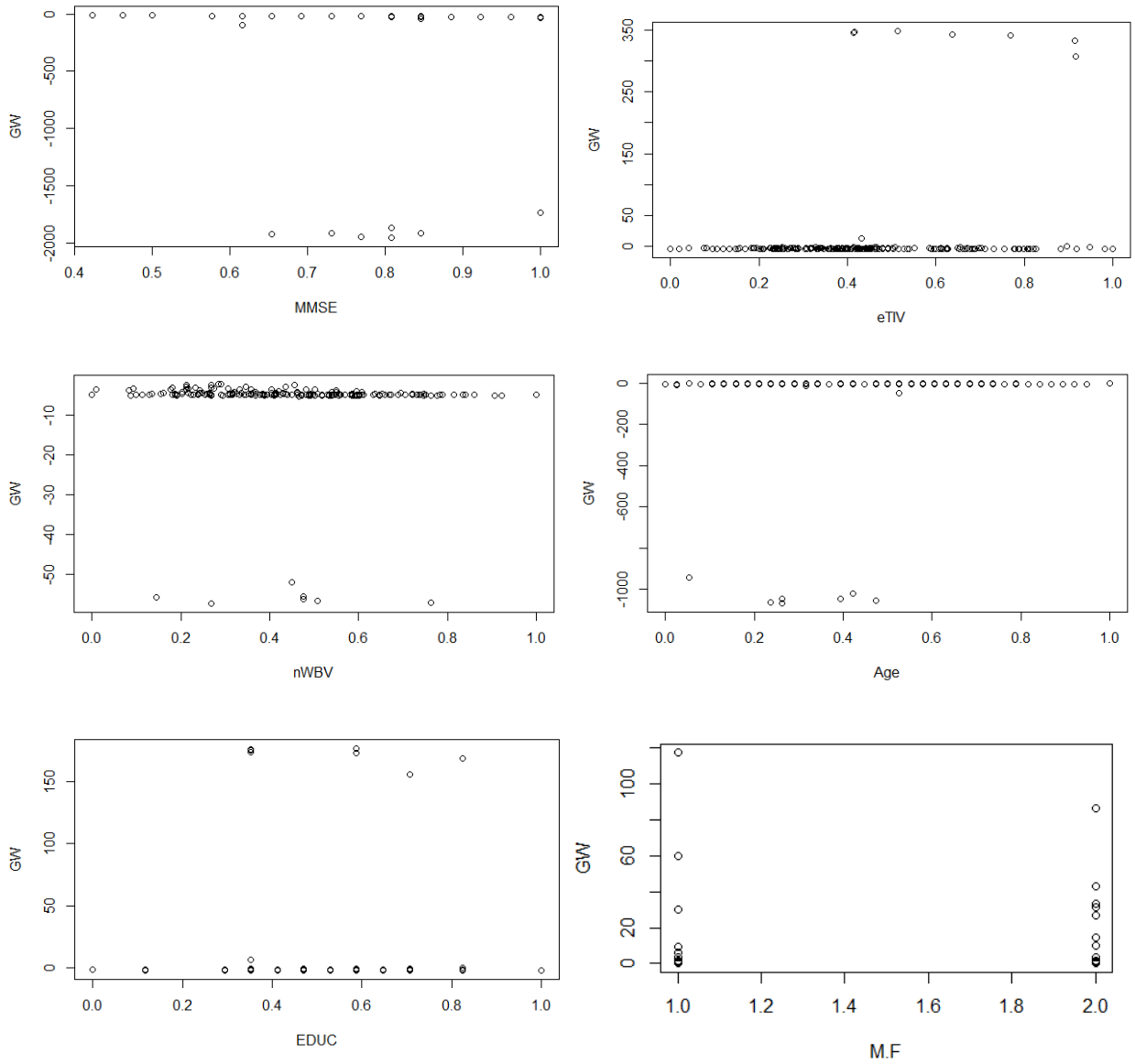
Figure 16: Plot of Neural Network with 2 Hidden Layer Nodes



Like in Figure 14, the black lines represent the connections between the nodes, whereas the blue numbers indicate the bias of a particular node. Each of the

black lines is assigned a numerical value for the weight. It took the neural network 58,078 steps to converge. Thus, the neural network with two hidden nodes required approximately 16.8 times the amount of steps to converge than the neural network with one hidden node, which indicates the training of the neural network with two hidden nodes involved significantly more backpropagation than the neural network with only one hidden node. Below is a plot of the generalized weights with respect to each independent variable for the neural network with two hidden layer nodes.

Figure 17: Plot of Generalized Weights with Respect to each Independent Variable for Neural Network with 2 Hidden Layer Nodes



In Figure 17, the distribution of generalized weights suggests that all the independent variables appear to have a nonlinear effect since the variance of their generalized weights is overall greater than one. Although, at a first glance, MMSE

and Age look like they have a generalized weight of 0, the scale of the plot suggests that the points near 0 are actually above 1. Once the number of steps and distribution of generalized weights was calculated and generated for each of the neural networks, each of the neural networks was tested with the testing data, which contained 145 observations. Confusion matrices were created to evaluate the performance of each of the neural networks when tested using the testing data set.

Table 6: Confusion Matrix of Neural Network with 1 Hidden Node

Actual Class	Predicted Demented	Predicted Nondemented	% Correct
Demented	53	11	82.82%
Nondemented	13	68	83.95%

Overall Accuracy: **83.44%**

Table 7: Confusion Matrix of Neural Network with 2 Hidden Nodes

Actual Class	Predicted Demented	Predicted Nondemented	% Correct
Demented	52	14	78.79%
Nondemented	6	73	92.41%

Overall Accuracy: **86.21%**

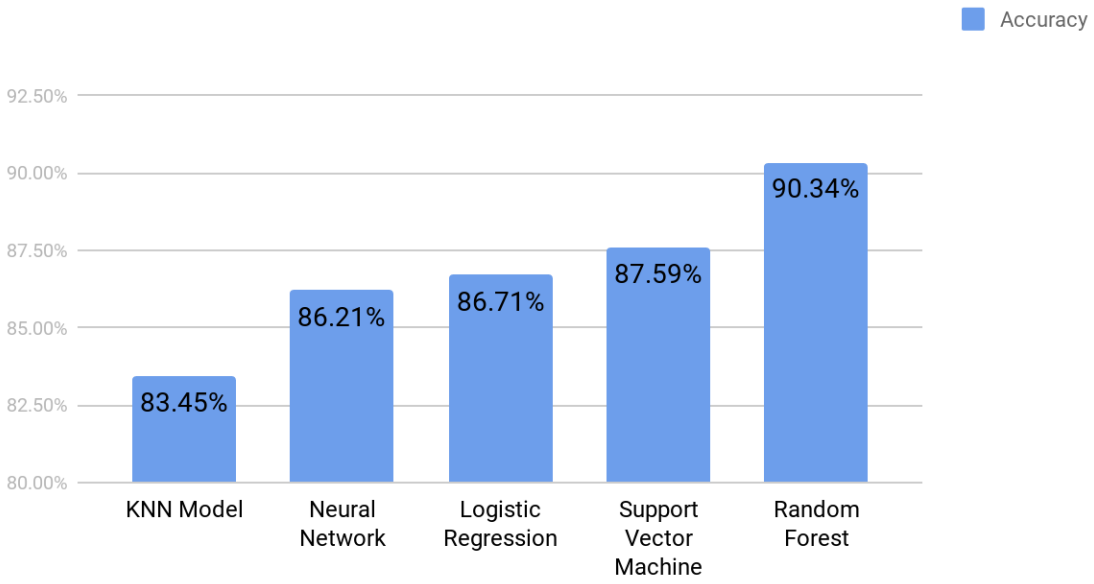
Of the 145 observations in the testing data, 64 observations were of demented individuals while 81 observations were of nondemented individuals. For the neural network with one hidden layer node, 53 of the 64 observations of demented individuals were accurately predicted, resulting in a sensitivity of 82.82%. 68 of the 81 observations of nondemented individuals were accurately predicted, resulting in a specificity of 83.95%. The overall accuracy rate of the first neural network model is 83.44%. For the neural network with two hidden layer nodes, the training and testing data were split differently; the testing data for the neural network with two hidden nodes had 66 observations of demented individuals and 79 observations of nondemented individuals. 52 of the 66 demented individuals were accurately predicted, resulting in a sensitivity of 78.79%. 73 of the 79 nondemented individuals were accurately predicted, resulting in a specificity of 92.41%. The overall accuracy rate of the second neural network with 86.21%.

4. Result

In this study, five types of machine learning models were developed to diagnose Alzheimer's based on clinical and demographic variables, including years of education, age, gender, MMSE, eTIV, and nWBV. In Section 3, the results of each machine learning model were evaluated and analyzed using a confusion matrix, and 4 of the 5 models were also evaluated using an ROC curve plot. Below

is a plot comparing the highest accuracies of each of five types of machine learning models developed in this study.

Figure 18: Comparison of Machine Learning Model Accuracies



As shown in Figure 18, all of the machine learning models are in the range of 83% to 91% accuracy. The random forest model has the highest accuracy, followed by the support vector machine, logistic regression, neural network, and KNN model in order of decreasing accuracy.

In addition to assess accuracy rate for each model, dataset were divided into the same training and testing sets to assess concordance (agreement on predictions) between various models. The concordance rate between each pair of models was provided in Table 8.

Table 8: Concordance Values Between Each Model

	Random Forest	KNN Model	SVM Model	Neural Network
Logistic Regression	91.78%	87.67%	93.15%	97.26%
Random Forest		93.15%	93.15%	91.78%
KNN Model			89.04%	87.67%
SVM Model				90.41%

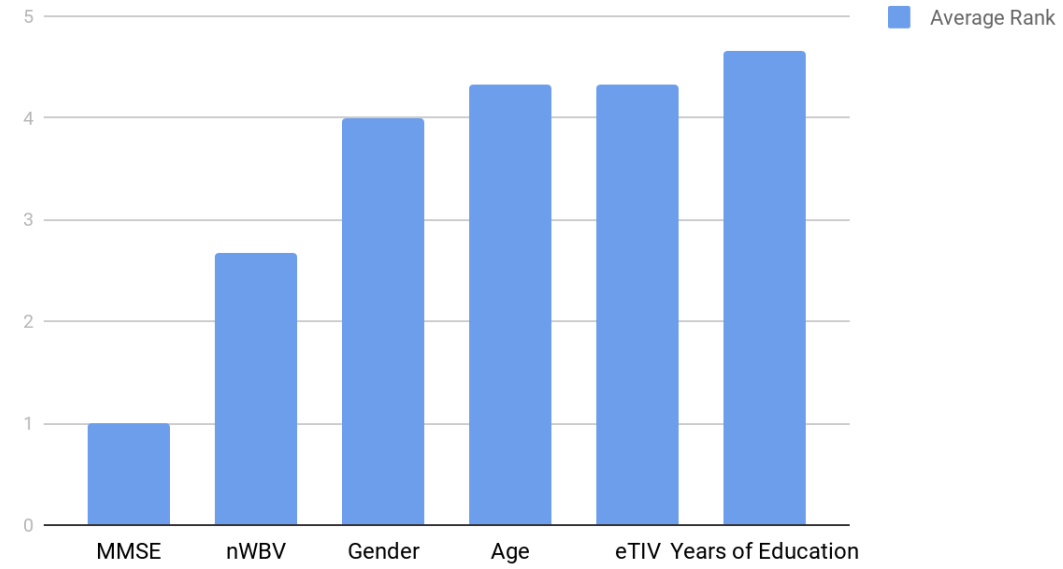
Overall, the concordance rates are pretty good. The logistic regression and neural network models have the highest concordance value of 97.26%. The rest of the models have fairly high concordance values between 87.67% and 93.15%. Among all the models, the percent in which at least 4 of the 5 models shared the same diagnosis for a testing input was 90.42%.

Another model was created that outputted the result that the majority of the 5 models outputted. This model reached an accuracy rate of 89.04%, which is lower than the accuracy of the random forest model but higher than the accuracy of the other four models.

When creating the logistic regression and random forest models, the importance of variables were evaluated. In logistic regression, the *p* value was used to evaluate a variable's statistical significance, and in the random forest models, mean decrease accuracy and mean decrease GINI were used to evaluate the model's dependent on each variable. Below is a plot indicating the average rank

(in terms of importance) among the different ways to evaluate variable importance for each variable. A rank of "1" indicates the greatest importance, whereas a rank of "6" indicates the least importance. A lower numerical value for rank indicates greater importance.

Figure 19: Rank of Importance of Each Variable



As shown in Figure 19, MMSE is clearly the most important variable by a considerable margin. In decreasing order of importance, MMSE is followed by nWBV, gender, age, eTIV, and years of education. The last four variables all seem to have similar levels of importance, whereas there is a considerable margin between the first three variables.

5. Discussion

The purpose of this research is to create a machine learning model to predict Alzheimer's using publicly available MRI data. There were 373 MRI imaging

sessions in the data, with 150 distinct individuals among those sessions. In each session, the individual's CDR, eTIV, nWBV, age, gender, and years of education were collected and used to train the machine learning models. ASF and socioeconomic status were both part of the original data set but were omitted from the study because ASF was directly correlated with eTIV and because roughly 13% of all the imaging sessions did not have a recorded socioeconomic status.

In order to assess sensitivity and robustness of our findings, multiple methods were used to build a model that can predict Alzheimer's. These methods include Logistic Regression, K Nearest Neighbor, Support Vector Machine, Random Forest, and Neural Network. These methods all had fairly consistent accuracy rates, as all the accuracy rates were within an 8% range. MMSE was consistently identified as the most important predictor, while other variables varied in importance.

In another similar study conducted by the French National Institute of Health and Medical Research, researchers presented and evaluated a new method based on SVM to diagnose patients with Alzheimer's using MR images. Using their new method based on SVM, the researchers were able to achieve a 94.5% accuracy in detecting Alzheimer's by bootstrap resampling a dataset containing 16 individuals with AD and 22 controls. In this study, rather than directly using raw MR images, MRI biomarkers along with cognitive and demographic data were used to classify individuals as either AD or non-AD. The SVM method in this study achieved a 87.59% accuracy, which is 7% lower than the accuracy of the new SVM method created by the researchers. The best accuracy in this study, 90.34%, is still roughly 4% lower than the new method created by the researchers. However, it's important

to note that the sample size of the data set used by the researchers is roughly 10 times smaller than the data set used in this study.[13]

Compared with the clinical diagnosis rate of Alzheimer's among rural doctors in the United States, all the models presented in the study are significantly better. The accuracy rate of diagnosing Alzheimer's among community doctors in rural areas is about 50 to 60 percent.[5] The model with the lowest accuracy presented in this study, KNN model, still does about 40% better than rural doctors in terms of accuracy. The model with the highest accuracy, Random Forest, does about 50% better than rural doctors. Moreover, the clinical diagnosis rate of Alzheimer's is estimated to be 78% among all doctors in The United States.[6] The worst model in the study still does about 7% better in terms of accuracy whereas the best model does about 15% better in terms of accuracy. Thus, the models created in this study have the potential to aid doctors to create more accurate clinical diagnoses of Alzheimer's.

This study can be improved by using a greater sample size of data and utilizing more MRI biomarkers and other independent variables. In addition, patients can have longer and more frequent follow-up MRI imaging sessions. The models can be improved by also accounting for patients' changes over time, rather than from one-time imaging sessions.

6. Conclusion

In conclusion, a variety of models were utilized to predict Alzheimer's using MRI biomarkers, cognitive assessment data, and demographic data from 371 imaging sessions. All models provide reasonable accuracy ranging from (83% to 91%) in predicting dementia. Among all the variables tested, MMSE was the most

influential variable, while other variables also influenced the detection of Alzheimer's at a lesser degree. Using the machine learning models in the study, physicians or nurses can potentially more accurately diagnose Alzheimer's in patients. An accurate diagnosis of Alzheimer's is important because it allows patients to better devise strategies to manage symptoms and plan for long-term care.

References

1. Cobb, B. R., Wells, K. R., & Cataldo, L. J. (2012). Alzheimer's Disease. In K. Key (Ed.), The Gale Encyclopedia of Mental Health (3rd ed., Vol. 1, pp. 59-73). Detroit, MI: Gale. Retrieved from https://link.gale.com/apps/doc/CX4013200025/GPS?u=watchunghrhs&sid=GPS&xid=9f274526
2. Martone, R. L., & Piotrowski, N. A., PhD. (2019). Alzheimer's disease. Magill's Medical Guide (Online Edition).
3. Alzheimer's Disease. (2018). Funk & Wagnalls New World Encyclopedia, 1;
4. Yiğit, A., & Işik, Z. (2020). Applying deep learning models to structural MRI for stage prediction of Alzheimer's disease. Turkish Journal Of Electrical Engineering & Computer Sciences, 28(1), 196-210. doi:10.3906/elk-1904-172
5. Kolata, G. (2019, August 01). A Blood Test for Alzheimer's? It's Coming, Scientists Report. Retrieved from https://www.nytimes.com/2019/08/01/health/alzheimers-blood-test.html
6. Reinberg, S. (2016, July 26). 2 in 10 Alzheimer's Cases May Be Misdiagnosed. Retrieved from https://www.webmd.com/alzheimers/news/20160726/2-in-10-alzheimers-cases-may-be-misdiagnosed
7. J. D. Warren, J. M. Schott, N. C. Fox, M. Thom, T. Revesz, J. L. Holton, F. Scaravilli, D. G. T. Thomas, G. T. Plant, P. Rudge, M. N. Rossor, Brain biopsy in dementia, Brain, Volume 128, Issue 9, September 2005, Pages 2016–2025, https://doi.org/10.1093/brain/awh543

8. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open Access Series of Imaging Studies: Longitudinal MRI Data in non-demented and Demented Older Adults. Journal of Cognitive Neuroscience, 22(12), 2677-2684. doi:10.1162/jocn.2009.21407
9. Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology, 6(2), 67-77. doi:10.1038/nrneurol.2009.215
10. Sargolzaei, S., Sargolzaei, A., Cabrerizo, M., Chen, G., Goryawala, M., Noei, S., . . . Adjouadi, M. (2015). A practical guideline for intracranial volume estimation in patients with Alzheimers disease. *BMC Bioinformatics, 16*(S7). doi:10.1186/1471-2105-16-s7-s8
11. Khan, T. (2016). *Biomarkers in Alzheimers disease*. Amsterdam: Academic Press.
12. Medical Tests. (n.d.). Retrieved from https://www.alz.org/alzheimers-dementia/diagnosis/ medical_tests
13. Magnin, B., Mesrob, L., Kinkingnéhun, S. et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83 (2009). https://doi.org/10.1007/s00234-008-0463-x