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

Jay Fu, Watchung Hills Regional High School

Introduction

Alzheimer's disease is ranked as the fourth leading cause of death in US, with 65,800 fatalities attributable to the disease each year. Currently over 50 million people worldwide were diagnosed with Alzheimer's. The disease is caused by the degeneration and eventual death of a large number of neurons in several areas of the brain. Alzheimer's is an irreversible disease in which patients progressively lose their memory and thinking skills. Early detection and treatment can slow down the disease progression. The objective of this research is to build sophisticated machine learning models incorporating MRI imaging data to detect Alzheimer's with high accuracy.

Research Objective

The research aims to use available MRI imaging data to build various predictive machine learning models including Logistic Regression (Logit), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) , and Neural Network (NN), and to compare accuracy of each model.

Data and Method

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. Each individual had at least two visits, separated by at least a year, during which MRI and clinical data were obtained. In each visit, participants were classified as demented nondemented, or converted (from nondemented to demented).

Method

Exploratory data analysis were conducted to remove covariates that are highly correlated. The OASIS data set were randomly split into training set (60%) and testing set (40%). Five models including logistic regression, K nearest neighbor, support vector machine, random forest, and artificial neutral network, were built using training set. The testing set was used to assess the accuracy of prediction for each model. Key risk factors were identified, and various models were compared to come forward with the best prediction model. The concordance between the models were also computed to assess consistency of each model prediction.

Results

- All of the machine learning models provides good accuracy in the range of 83% to 91%, in which the random forest model has the highest accuracy, followed by the support vector machine, logistic regression, neural network, and KNN model.
- Mini Mental State Examination (MMSE) is the most important variable by a considerable margin, followed by normalized Whole Brain Volume (nWBV), gender, age, estimated Total Intracranial Volume (eTIV), and years of education.
- The logistic regression and neural network models have the highest concordance rate of 97.26%. The rest of the models have fairly high concordance ranging 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%.

Table 1 Variables used in this study

Sex	1: male 2: female
Age	60 to 96 years old
EDUC (years of education)	6 to 23 years
eTIV (estimated total intracranial volume)	1106 to 2004 mm^3
nWBV (normalized whole-brain volume)	Proportion of tissue in brain volume, value from 0 to 1
MMSE (Mini Mental State Examination)	Score between 1 to 30

Table 2 Logistic Regression

Predictor	Est.	Std. Error	*P*	Odds Ratio	Risk increase
Intercept	80.129	15.246	<0.001	NA	NA
nWBV	-31.135	9.858	0.0016	3e-14	100%
eTIV	-0.0038	0.0018	0.0341	0.996	-0.38%
Age	-0.12	0.045	0.0079	0.887	11.31%
MMSE	-1.387	0.239	<0.001	0.25	75.03%
EDUC	-0.259	0.096	0.0068	0.772	22.85%
Gender	1.499	0.604	0.013	4.479	347.85%

Table 3 Pairwise Concordance Rate (in percent)

	RF	KNN	SVM	NN
Logit	91.78	87.67	93.15	97.26
RF		93.15	93.15	91.78
KNN			89.04	87.67
SVM				90.41

Figure 1 Comparison of Model Accuracies

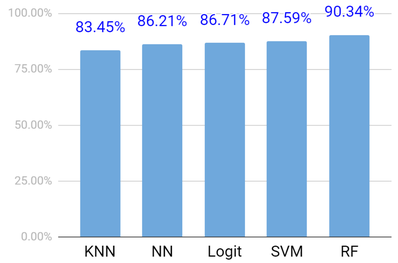


Figure 2 Comparison of Key Risk Factors

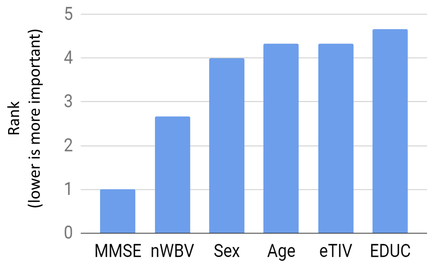


Figure 3 Accuracy Plot of KNN Algorithm

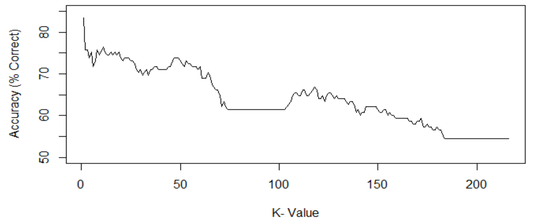


Figure 4 Neural Network with 2 Nodes

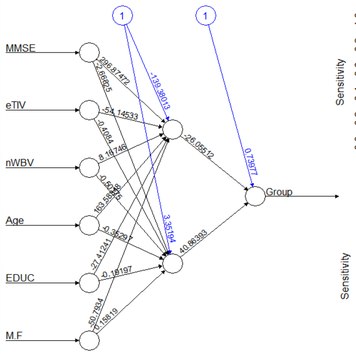


Figure 5 ROC Plots

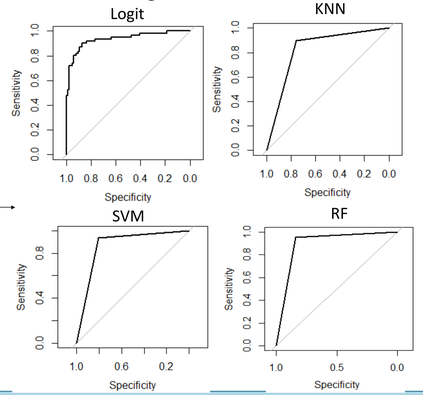


Figure 6 Accuracy of SVM

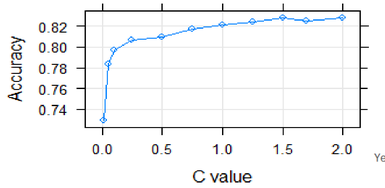


Figure 7 OOB and Misclassification Error Plot

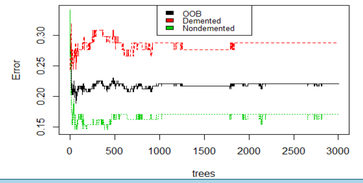
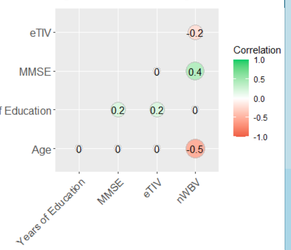
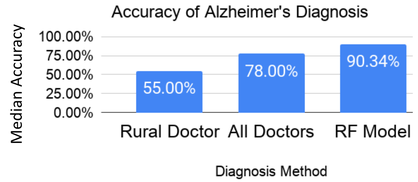


Figure 8 Correlogram



Significance and Conclusions

- Five models including logistic regression, K-nearest neighbor, support vector machine, random forest, and neural network were used in this research.
- Machine learning models incorporating MRI imaging data produce good prediction accuracy of Alzheimer's disease with accuracy ranging from 83.45% (KNN model) to 90.34% (Random forest model).
- The Random Forest model detects Alzheimer's at an accuracy about 50% higher than that of American rural doctors and 15% higher than that of all American doctors.



- Mini Mental State Examination (MMSE) and normalized Whole Brain volume (nWBV) are the two most influential factors in predicting Alzheimer's disease.
- Random Forest model produces the highest prediction accuracy (90.34%) among the five models considered.
- The concordance (a measure of agreement) between each pair of models is high, ranges from 88% to 97%.
- The developed models allow an early detection of Alzheimer's and potential early treatment/intervention to slow down the disease progression and prolong survival.

Acknowledgement

I'd like to thank my mentor Tommy Fu for his help reviewing and giving recommendations for my research.

Reference

1. Cobb, B. R (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. 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
3. 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