Evaluation of Machine Learning Algorithms Performance in Predicting Alzheimer’s Disease

Nikki Hessner, Clément Karumuhinzi, Marie Aimée Karumuhinzi,

Gayatri Soni, Yin Yin Tan
Team 01, DS 620 Machine Learning/Deep Learning,

School of Technology and Computing,

City University of Seattle

hessnernicole@cityuniversity.edu, karumuhinziclement@cityuniversity.edu, karumuhinzimarieaim@cityuniversity.edu, sonigayatri@cityuniversity.edu, tanyinyin@cityu.edu

Abstract

This project will evaluate the performance of machine learning algorithms in predicting Alzheimer’s disease. The goal is to find the best algorithm with high accuracy score in detecting the disease so that proper prevention measures can be implemented. Alzheimer’s disease is one of the major causes of dementia and detecting it early with the help of machine learning will provide timely medical intervention. Machine learning allows the use of mathematical and statistical models to achieve learning and adaptation without too many explicitly coded instructions. The algorithm learns gradually, continuously and improves its performance over time. In this paper we will evaluate the performance of different machine learning techniques that could assist in early detection of Alzheimer's disease to provide early treatment.

Keywords: Artificial Intelligence, Machine Learning, Performance Metrics, Prediction, Alzheimer Disease

1. INTRODUCTION

Alzheimer’s disease is a neurological disorder that causes the shrinking of the brain and dying of the brain cells and affects about 5.8 million people in the United States. (Mayo Foundation for Medical Education and Research, 2022). The disease impacts a person’s thinking, behavioral, and social skills. Therefore, it is the most common cause of dementia. Alzheimer’s disease impacts about 60% - 70% of the 50 million people worldwide with dementia (Mayo Foundation for Medical Education and Research, 2022). Unfortunately, there is no treatment to cure this disease.

With medical advances, images of the brain can help doctors diagnose Alzheimer’s disease in patients. Two of the imaging technologies used are Magnetic Resonance Imaging (MRI) and

Computerized Tomography (CT). In this report, we will focus on the MRI images of patients and use machine learning algorithms to detect early Alzheimer's disease in patients.

Machine learning is a branch of Artificial Intelligence that focuses on data and algorithms uses. It is a key component of data science, because it will improve accuracy and make better predictions using the dataset(s) that are being analyzed (IBM Cloud Education, n.d.). In this report, we will use several machine learning algorithms with the hope of finding the best model for detecting Alzheimer's disease and provide a better understanding of the disease as well as its effects on patients.

We will use five machine learning algorithms, and then compare the results to see which one performs better on the chosen dataset. The dataset used in this report is from Kaggle, Detecting Early Alzheimer's Disease (Choi, Song, & Parikh).

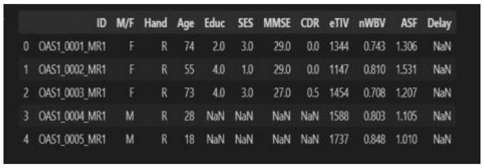
First, we will use K-Nearest Neighbors algorithm, followed by Logistic Regression algorithm, Super Vector Machine algorithm, Decision Tree algorithm, and Random Forest algorithm to analyze the dataset. We will compare the results of these five algorithms in a later section of the report.

This report consists of six sections, starting with an Introduction, Data Collection, Machine Learning Methods, Results and Discussion, Conclusion and finally, References.

2. DATA COLLECTION

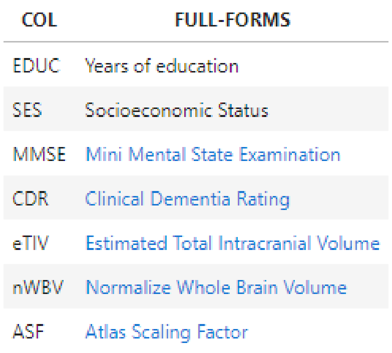
The two .csv files downloaded from the Kaggle site are oasis_cross-sectional.csv and oasis_longitudinal.csv. OASIS stands for Open Access Series of Imaging Studies, which publishes "neuroimaging datasets freely available to the scientific community" (Oasis Brains). We will first explore the data from the oasis_cross-sectional.csv file.

In the oasis_cross-sectional.csv file, there are 436 rows and 12 columns. The column headings are shown in Figure 1. The first one is ID which is the identification of the MRI image; M/F is the gender of the subject; Hand shows the dominant hand of the subject; Age is the subject's age in years; Education is the education level. The rest of the column headings can be seen in Figure 2.



	ID	M/F	Hand	Age	Educ	SES	MMSE	CDR	eTIV	nWBV	ASF	Delay
0	OAS1_0001_MR1	F	R	74	2.0	3.0	29.0	0.0	1344	0.743	1.306	NaN
1	OAS1_0002_MR1	F	R	55	4.0	1.0	29.0	0.0	1147	0.810	1.531	NaN
2	OAS1_0003_MR1	F	R	73	4.0	3.0	27.0	0.5	1454	0.708	1.207	NaN
3	OAS1_0004_MR1	M	R	28	NaN	NaN	NaN	NaN	1588	0.803	1.105	NaN
4	OAS1_0005_MR1	M	R	18	NaN	NaN	NaN	NaN	1737	0.848	1.010	NaN

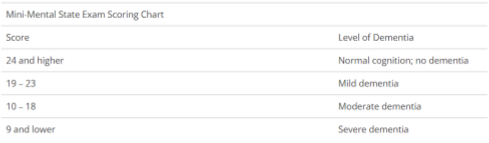
Figure 1. Oasis Cross-Sectional Column Headings



COL	FULL-FORMS
EDUC	Years of education
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

Figure 2. Data Column Descriptors (Choi, Song, & Parikh).

Mini mental state exam is a 30-point test aimed at accessing the exam taker's cognitive ability. There are 4 levels of dementia severity which can be seen from the chart below. Mild dementia is from 19-23 and scores lower that indicate more severe dementia.



Mini-Mental State Exam Scoring Chart

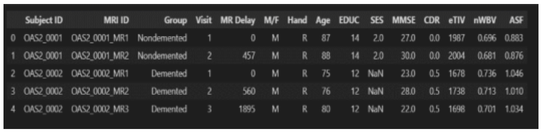
Score	Level of Dementia
24 and higher	Normal cognition; no dementia
19 – 23	Mild dementia
10 – 18	Moderate dementia
9 and lower	Severe dementia

Figure 3. MMSE Scoring Chart

Clinical Dementia Rating shows the stages of severity of dementia symptoms. Since current study only concerns itself with early AD, only patients with CDR ratings of 0.0, or 0.5 or 1.0 were included.

Estimated total intracranial volume is a measure of head size in brain studies. Normalize whole brain volume is used to normalize the brain volume according to head size. Atlas scaling factor is "a one-parameter scaling factor that allows for comparison of the estimated total intracranial volume (eTIV) based on differences in human anatomy" (Fulton, et al. 2019). The last column, Delay, is not explicitly described and we are not clear on what that exactly means.

For the oasis_longitudinal.csv file, there are 373 rows and 15 columns. The column headings are as follows: Subject ID, MRI ID, Group, Visit, MR Delay, M/F, Hand Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, and ASF. Figure 4 below shows the column names and the first 5 rows of the data.



	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1698	0.701	1.034

Figure 4. Oasis Longitudinal Column Headings

Subject ID is the ID of the subject while MRI ID is the ID of the MRI image. Group is the group that the subject is categorized as nondemented or demented. Visit is the number of visit the MRI image was taken. MR Delay is delayed enhancement after a contrast agent is injected. The other headings are the same as the oasis_cross_sectional dataset. In our report, we will focus on oasis_longitudinal dataset to better understand detection of Alzheimer's disease. The longitudinal dataset will help us analyze the subjects and observe if there are any that went from nondemented to demented throughout the study period.

3. MACHINE LEARNING METHODS

In this section, we will briefly explain what each machine learning algorithm is and how is works. Then, we will describe our process with implementing each algorithm on the oasis_longitudinal.csv file.

3.1 K-Nearest Neighbors Algorithm

K-Nearest Neighbors (KNN) algorithm is a supervised learning algorithm that can be used for classification or regression problems. More commonly, it is used for classification. KNN algorithm identifies the nearest neighbors of a given data point and then assigns the appropriate label to that data point (IBM, n.d.). The assigning of the label is based on what label that is often associated with the nearest neighbors around that given data point.

3.2 Logistic Regression Algorithm

Logistic Regression algorithm is one of the supervised machine learning techniques which is used for classification and regression problems. It is typically used when dealing with binary classification. There are several types of logistic regression: binary or binomial, multinomial and ordinal. In this report, we will be dealing with binary type since we only have two classes to predict.

The logistic regression model is implemented in two steps: first we'll train the model without fine tuning and lastly build the model using optimal hyperparameters. For each trained model, we'll evaluate its performance based on these metrics – accuracy, recall, precision, F1 score and area under the curve (AUC) score. These metrics will be calculated for both train and test datasets. In order to fine tune the logistic regression model, we'll apply both Cross Validation and GridSearchCV methods to find the best parameters (C, penalty, etc.) to improve our model's accuracy.

3.3 Support Vector Machine (SVM) Algorithm

Support Vector Machine algorithm is one of the machine learning techniques which is mainly used for classification problems but can be used to solve regression problems as well. It is fast and dependable technique which performs well for binary classification especially if you have limited data. The different SVM kernels that will be considered for this report are linear, polynomial, and radial basis function kernels.

The support vector machine model is implemented in two steps: first we'll train the model without fine tuning and lastly build the model using optimal hyperparameters. For each trained model, we'll evaluate its performance based on these metrics – accuracy, recall, precision, F1 score and area under the curve (AUC) score. These metrics will be calculated for both train and test datasets. In order to fine tune the support vector model, we'll apply both Cross Validation and GridSearchCV methods to find the best parameters (kernel, C, gamma, etc.) to improve our model's accuracy.

3.4 Decision Tree Algorithm

The Decision Tree Algorithm is supervised machine learning algorithm that sorts data based on a series of binary decision sets. Decision Trees can be split into two categories: Classification Trees and Regression Trees. Classification Trees seek to answer a yes/no question by dividing data into categories. Regression trees work with continuous data based on the related principles of entropy and information gain (Xoriant Corporation, n.d.).

3.5 Random Forest Algorithm

Decision Trees, while effective for many classification problems, have a tendency to succumb to overfitting. A Random Forest is a collection of low correlation individual Decision Trees that work together to make a prediction. When the feature selection order is randomized, some trees will be right, and others will have errors that would be misleading if they were on their own. However, the group of trees can better point to which direction is the right one (Yiu, 2019).

4. RESULTS AND DISCUSSION

In this paper, we evaluate various performance metrics like accuracy, recall, precision and F1 score. We perform cross validation to determine the best parameter for each model: Logistic

Regression, SVM. Furthermore, the results of different classification algorithms with parameter confinement are compared and each classifier is evaluated in terms of ROC-AUC curve. Moreover, we evaluated the performance of each machine learning model using confusion metrics. Recall is the proportion of people correctly classified as having Alzheimer's disease and precision represents the rate if people accurately classified as not having disease. On the other hand, F1 is the weighted average of recall and precision, while accuracy is the proportion of people correctly classified. Based on the results, the patient gets a report that tells them what stage of Alzheimer disease they are currently in. Additionally, knowing the stage helps doctors to better understand how the disease is affecting them.

The model's function, model prediction, and results will be discussed in this section for final submission.

5. CONCLUSION

The main goal of this project is to predict Alzheimer's disease. For predicting Dementia or Alzheimer's disease in adult patients, we used the MRI and Alzheimer's dataset provided by the open Access Series of Imaging Studies (OASIS). First, we visualized the dataset and checked for missing values and replaced it with median values. Furthermore, we preprocessed the dataset by dropping out unnecessary features. Moreover, we standardized the values so that it can fit in the machine learning model. Next, the dataset has been used to train Logistic Regression Model, SVM Model. Furthermore, we calculated the evaluation metrics, accuracy, precision, recall, F1 score, ROC/AUC and confusion matrix. To improve the overall result, the grid search method is used to fine-tune all the machine learning model.

We will discuss comparison of all the models: Logistic regression model, SVM, KNN, Decision tree, Random Forest in the final submission.

6. REFERENCES

Choi, Hyunseok, Song, Kyuri, Parikh, Saurin, "Detecting Early Alzheimer's Using MRI Data and Machine Learning", retrieved Nov. 02, 2022 from https://www.kaggle.com/code/hyunseokc/detecting-early-alzheimer-s/

Fulton, L. V., Dolezel, D., Harrop, J., Yan, Y., &amp; Fulton, C. P. (2019, August 22). Classification of alzheimer's disease with and without imagery using gradient boosted machines and resnet-50. MDPI. Retrieved November 5, 2022, from https://www.mdpi.com/2076-3425/9/9/212/htm#:~:text=Atlas%20scaling%20factor%20%28ASF%29%3A%20%280.88%E2%80%931.56%29%20%28observed%29.%20The%20ASF,100%25%20complete%20%28416%20of%20416%29%20%5B%2010%20%5D.

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. O'Reilly. (ISBN: 978-1-492-03264-9)

Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Frontiers in Public Health, 10. https://doi.org/10.3389/fpubh.2022.853294

IBM Cloud Education. (n.d.). What is machine learning? IBM. Retrieved October 25, 2022, from https://www.ibm.com/cloud/learn/machine-learning

IBM.(n.d.). What is the K-nearest neighbors algorithm? Retrieved November 12, 2022, from https://www.ibm.com/topics/knn

Jo T, Nho K, Saykin AJ. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. Front Aging Neurosci. 2019 Aug 20;11:220. doi: 10.3389/fnagi.2019.00220. PMID: 31481890; PMCID: PMC6710444.

Mayo Foundation for Medical Education and Research. (2022, February 19). Alzheimer's disease.

Xoriant Corporation. (n.d.). Decision Trees for Classification: A Machine Learning Algorithm. Xoriant. https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm

Yiu, T. (2019, June 12). Understanding Random Forest. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2