# A comparison and characterization of machine learning models using the OASIS dataset

Blake Andreou
Email: blake.a.andreou@vanderbilt.edu

***Abstract* – Since its discovery, Alzheimer's disease has increasingly grown to become a global health crisis. Scientists are desperately searching for ways to identify the disease early to prevent symptom development from occurring. The current study focuses on this goal by characterizing a set of machine learning techniques' abilities to detect individuals with Alzheimer's given demographic and cognitive data using the Open Access Series of Imaging Studies (OASIS) dataset. We conclude that the support vector machine provided the best discriminatory power out of the selected models.**

***Keywords – Alzheimer's Disease, OASIS***

## I. Introduction

### *A. Alzheimer's Disease*

Alzheimer's Disease is a gradually progressive neurological disease with devastating effects on those afflicted [1]. Symptoms include those in cognitive, neurological, and social domains. In the cognitive domain, individuals afflicted with Alzheimer's disease report a loss of memory, difficulty in completing tasks once familiar to them, and difficulty understanding and producing language, among other symptoms. In the neurological domain, we see a plethora of dementia-specific features pointing to mass cell death in the cortex, including beta-amyloid plaques as well as neurofibrillary tangles [2]. Socially, individuals afflicted with this disorder see the downstream impact of the previous two symptom domains, and often have trouble understanding and empathizing with those around them. All symptoms of the disorder will worsen over time, and the disease is terminal.

### *B. Prevalence & Severity of Alzheimer's Disease*

Alzheimer's Disease has devastating consequences not only for the individuals affected, but also the country at large. Alzheimer's Disease is currently far and away the largest cause of dementia in elderly individuals in the United States, accounting for nearly 70% of cases. As of 2023, 6.7 million individuals are currently living with Alzheimer's dementia in the United States. This accounts for 10.8% of the population of those 75 and older in America, or nearly 1 in 9 [3]. Furthermore, evidence suggests that the proportion of individuals with the disease is rising as our society becomes collectively older. It is estimated by Tarawneh and Holtzman that this 6.7 million prevalence of Alzheimer's disease will increase to approximately 7.7 million people by 2030, and nearly 12 million by 2050 [4]. There is therefore a substantial and increasing need for the development of mechanisms to treat Alzheimer's Disease.

### *C. Treatment of Alzheimer's Disease*

The treatment of Alzheimer's disease has, especially in recent years, primarily focused on the identification of relevant biomarkers for early diagnosis of the disorder, so that preventative measures can be put in place to prevent the progression of the disease [4]. This is for two reasons. The first is that there have not been any treatments that have successfully returned cognitive abilities to individuals with advanced stages of the disease. The second is that the symptoms themselves are deleterious to the day-to-day life of the individuals afflicted with the disorder, and it is logical to prevent the credence of the symptoms rather than handle existing ones. Recent research has suggested that the pathology

of Alzheimer's disease begins substantially earlier than the emergence of any Alzheimer's-related symptoms. Therefore, for the well-being of individuals at-risk for developing the symptoms of the disorder, it makes the most sense to aim to develop a way to identify the disorder prior to the emergence of symptoms. Extending from the need for early identification of this disorder, it would be extremely helpful for us to automate this process to allow for efficient and accurate processing of patients at-risk for the disorder [5].

*D. Aim of Current Study*

The current study aims to provide insight into the automation of Alzheimer's detection.

We will be utilizing the OASIS (Open Access Series of Imaging Studies) dataset [6]. This dataset provides a series of demographic information and neural statistics for individuals within three classes: individuals with Alzheimer's disease, individual without Alzheimer's disease, and individuals who were initially identified as not meeting the criteria for Alzheimer's but were later identified as having the disease.

After running an EDA, we plan to use the following machine learning models in an attempt to effectively separate the data classes using the demographic/neurological data: multimodal logistic regression, decision tree model, support vector machine. These were chosen from two reference papers detailing similar machine learning techniques with this dataset [5,7].

We will make the following considerations for optimizations of each model. For the multimodal logistic regression, we will adjust the regularization techniques to attempt to maximize our efficacy. For the decision tree model, we will test various depths aiming for a balance between efficiency and memory storage that the resulting tree requires. For the support vector machines, we will test out various kernels to ensure that we are using the most strongly associated kernel for each analysis. For each model, we will analyze using 5-fold cross validation.

For the ablative portion of the study, we will focus on removing various features of the data itself. First, we will attempt a run of the data without standardization. Then we will run all future data runs after first standardizing the data. We will then attempt to use the models using only the three neurological datapoints for each visit. Then, we will slowly add relevant factors such as socio-economic status and gender, until we arrive at the most general model, including all features not identified as problematic in the EDA. Details on the ablation techniques specifically will be discussed in the following section.

## II. Methods

*A. OASIS Dataset*

The OASIS Dataset consists of a series of demographic datapoints (ID, Visit #, Time since first visit, Gender, Handedness, Age, Education Level, Socio-Economic Status, Mini Mental State Examination (MMSE), Clinical Dementia Rating (CDR)), and neurological datapoints (Estimated Total Intracranial Volume (eTIV), Normalized Whole Brain Volume (nWBV), Atlas Scaling Factor (ASF)) for individuals within three categories: non-Alzheimer's, converted to Alzheimer's, Alzheimer's. Converted refers in this case to individuals who were deemed non-Alzheimer's on the first visit but were labeled as Alzheimer's on a subsequent visit. Each of the 150 individuals in this study was right-handed, and they ranged in age from 65 to 96.

*B. Exploratory Data Analysis*

We first aim to characterize all features within the dataset to ensure that we are not including any features that would provide directed information towards an Alzheimer's

diagnosis, as well as to ensure that our features are not substantially correlated within themselves.

Using a variety of techniques from the Python pandas library [8], we visualized each datapoint, and compared it to our dependent variable of diagnosis, as well as to other independent variables that were thought could potentially have an impact on that variable. From this, we arrived at our target ablation groups, discussed in the ablation section of the methods. Results from the EDA will be elaborated on in the results area of the manuscript.

*C. Machine Learning Techniques*

We will be using scikit-learn for the creation, training, validation, and testing of each of our models [9]: a multinomial logistic regression model, a support vector machine, and a decision tree model. We elected to choose these three models as they represented an appropriate spread of possible ML methods.

To optimize the logistic regression model, we will attempt to use both L1 and L2 regularization to strengthen the power of the model. To optimize the decision tree model, we will attempt to limit the number of layers that can be created in order to produce a model of strongest accuracy with the smallest amount of computational power. Finally, to optimize the support vector machine, we will be adjusting the kernel between a linear, polynomial, sigmoid, and radial basis function.

In all cases, we will be evaluating the model using 5-fold cross validation, with a pre-separated external test set after the cross validation. For the evaluation metrics themselves, we will be assessing the models on the basis of accuracy, recall, and precision. To choose the best optimizer, we will choose the model with the highest accuracy. In the event of an approximate tie, we will then move to recall, precision, and runtime, in that order, to choose the most effective optimization criteria for each model.

*D. Ablation*

Based on our EDA, we arrived at the following 10 datasets to classify our data:

1. Full Data – RAW
2. Full Data
3. Full Data – Visit 1
4. Neuro Data – RAW
5. Neuro Data
6. Neuro Data – Visit 1
7. Neuro Data + MMSE
8. Neuro Data + MMSE – Visit 1
9. Neuro Data + MMSE + Age
10. Neuro Data + MMSE + Age – Visit 1

Full data refers to all data except for the categories of subject ID, Time since visit, handedness, and clinical dementia rating. The first two categories were removed as they are arbitrary and not associated with the disease. Handedness was removed as it was controlled for in the study by requiring right-handedness for all subjects. The final category was removed during the EDA process, as this category is a physician rating used specifically to diagnose the disease.

All data is standardized unless indicated otherwise with the ‘RAW’ suffix. Additionally, all data is collapsed across all visits unless indicated otherwise.

For the strongest performer, selected incorrect responses will be analyzed directly.

## III. Results

*A. Exploratory Data Analysis*

Our exploratory data analysis followed these general steps: visualization of each variable on their own, visualization of variables in association with other variables with suspected association.

Notable findings from the EDA include the following:

1. Subject ID, MRI ID, and Time since First Scan variables could be discounted as they provided no meaningful information about either the disorder or the subject.

2. Visit number at first glance felt inappropriate to add to the analysis, as it would be more reflective of the passing of time than the progression of the disease. However, we can see that there remains a similar distribution of each group across visits, as well as a similar brain volume across all visits, so we will leave all visits for analysis (Fig 1,2). We will, however, use the ablative step of considering only visit 1 for each subsequent set we create.

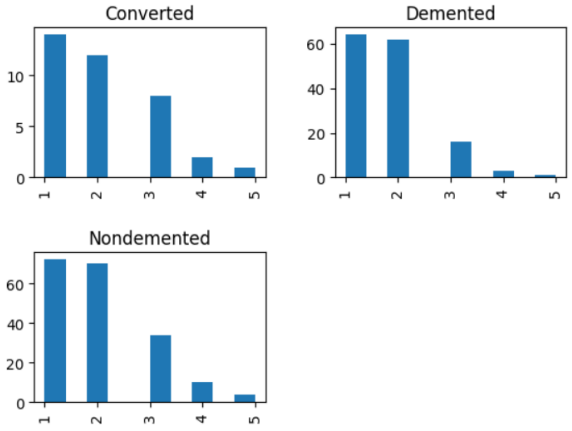


*Figure 1. Count of each Category per Visit*

3. Age was somewhat correlated with brain measurements. This is in line with previous research regarding age and brain health, but we consider the ablative step of removing age.

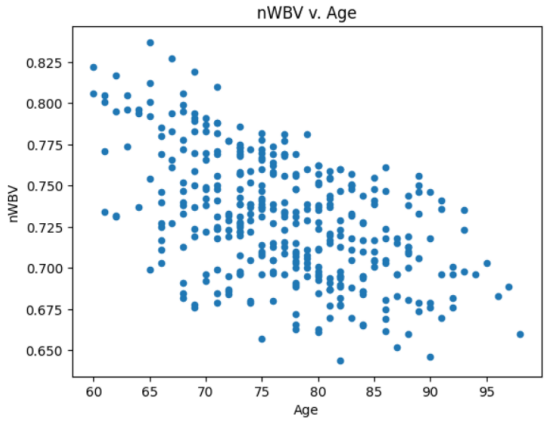


*Figure 2. normalized brain volume compared to age*

4. CDR could not be considered for the analyses as it functions as a flag for dementia, as described in the methods section (Fig 3).

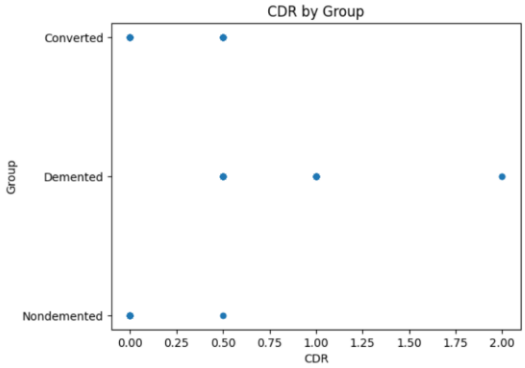


*Figure 3. Clinical Dementia Rating by Group.*

5. Estimated total cranial volume and atlas scaling factor are strongly negatively correlated. This makes sense, as the scaling factor is directly related to the cranial volume(Fig 4). We will, however, keep both variables in the analysis to preserve the full set of neurological data in the dataset.

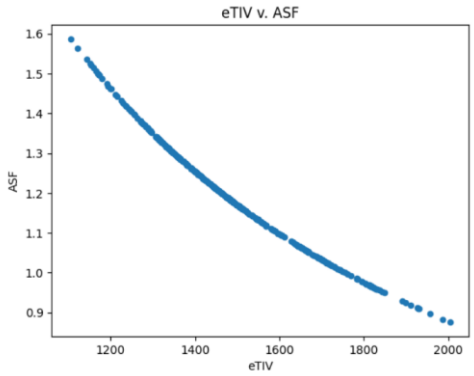


*Figure 4. cranial volume and scaling factor.*

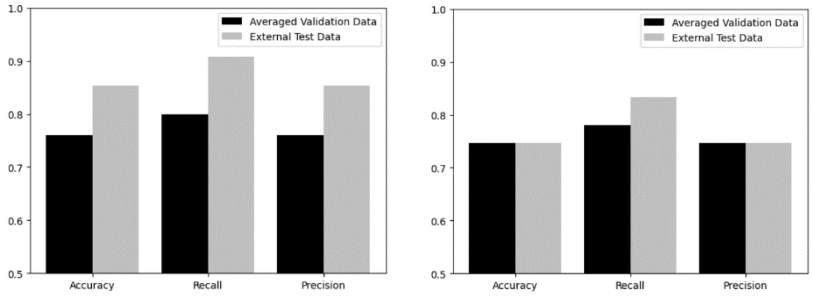


*Figure 6. Comparison of un-regularized and regularized performances, full RAW data.*

Since we are aiming to see neurological results as they occur prior to cognitive symptoms emerging, we will first look at the neurological data alone [5]. We will then add the MMSE, which is a cognitive score, demonstrating the potential increase in performance when cognition information is provided. We then add age, another known factor associated with dementia [4]. Finally, we will add all other potentially relevant variables, such as gender and socio-economic status. In all cases, we will also consider a visit 1 subset of the data. Additionally, for the smallest and largest datasets we consider the raw data as well as the standardized data.

*B. Modeling Results and Comparison*

We will first discuss the optimization used in each modeling technique in each case, We will then only consider this model as the 'strongest' and enumerate the differences between the cases (i.e. results of the ablative portion of the study) using only this technique.

Beginning with the logistic regression modeling, it was noted in this modeling that regularization did not substantially improve model performance for any of the datasets trained (Fig 6).

Next, in the decision tree model, we noticed that the max depth of the tree did not have a substantial impact on the eventual accuracy of the model when compared to a test set. Predictably, as more variables were added the models, the performance of shallower trees weakened. Interestingly, performance of shallower trees was stronger than deeper trees if the variable size was small, specifically in the cases of the Neuro-Only datasets (Fig 7).

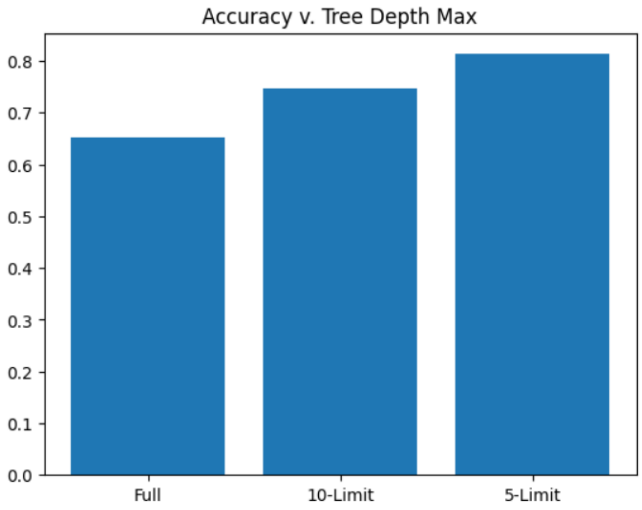


*Figure 7. Comparison of Accuracy at various max decision-tree depths -- Neuro Data – Visit 1.*

Finally, in the support vector machine optimization, the linear kernel was across the board the strongest kernel used (Fig 8).

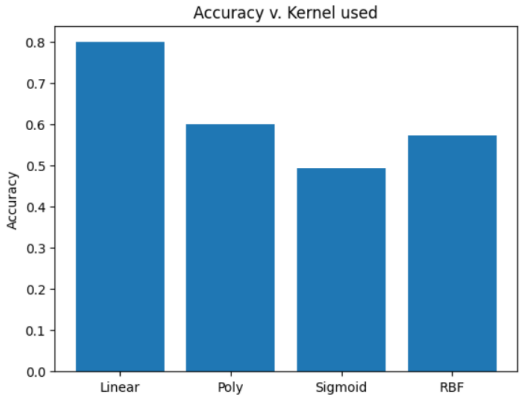


*Figure 8. Accuracy Comparison across kernels – Neuro+MMSE data*

Comparing all modeling techniques to each other, it becomes apparent that the linear kernel support vector machines provide the most consistent accuracy, precision, and recall for this dataset. When considering logistic regression, which is occasionally stronger in performance metrics in a few of our cases, we opted to use consistency as our next metric since our logistic regression model occasionally created models with severe overfitting (Fig 9).

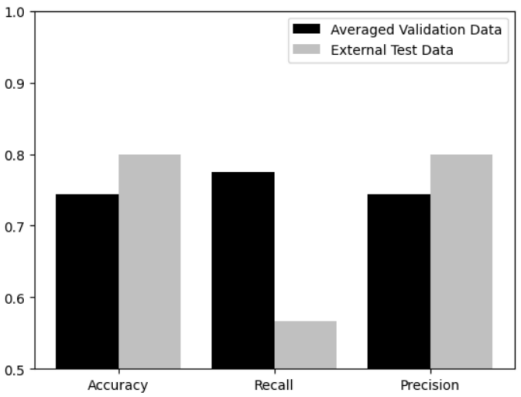


*Figure 9. Results of Neuro+MMSE run with Logistic Regression. Note overfitting in Recall.*

### *C. Ablation Results*

All the following results are taken from direct comparison from the combined metrics of accuracy, recall, and precision (Fig 10).

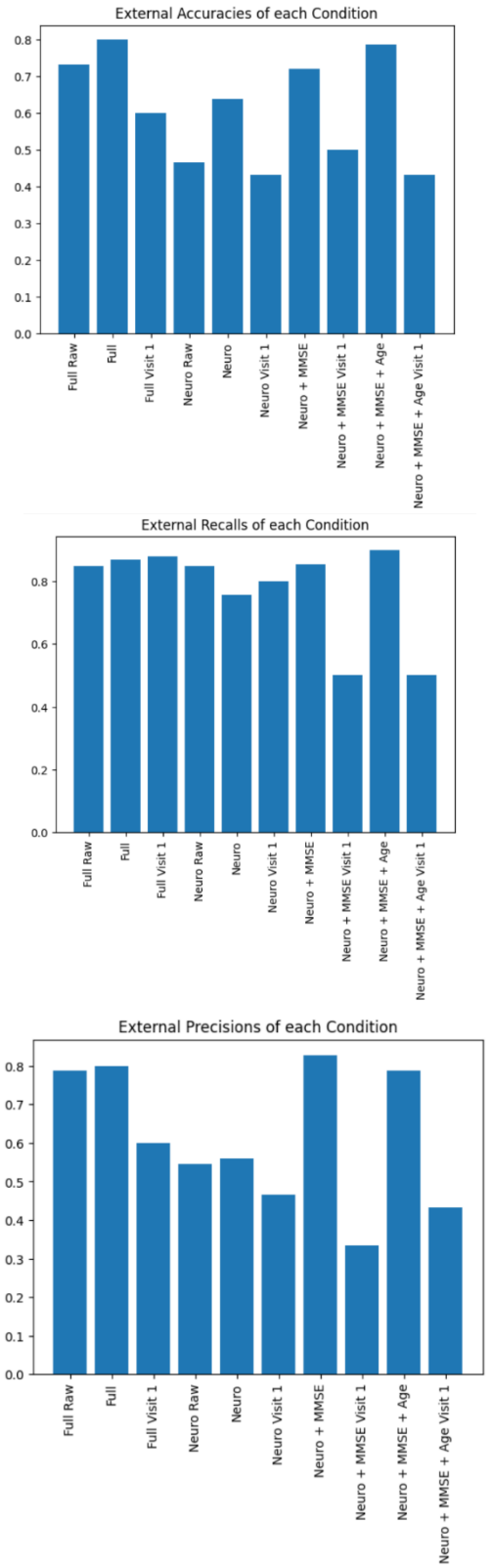


*Figure 10. Metric Comparison for each condition.*

First, we compare the raw model results to their standardized counterparts. We see no consistent difference in the performance of the linear support vector machines in any of the metrics listed (Fig 10).

Next, we compare the visit 1 model results to the full visit model counterparts. We notice, particularly in accuracy and precision, that the visit 1 models underperform their full visit counterparts.

Finally, we move on to the comparison of the models with only the neural features, to the inclusion of MMSE, age, and other demographic information, respectively. The inclusion of MMSE score massively improved the model performances from the neurological data alone. Additionally, when all demographic information was added to the model, we noted a similar performance across all metrics to the model trained on neurological data and MMSE alone.

## IV. Discussion

Recent developments in both computer science as well as neuroimaging have allowed us to automate the detection of neurological diseases for the first time. Our goal in the current study was to provide insight into the potential of various typical machine learning techniques with the OASIS dataset.

We were able to successfully train a logistic regression model, a decision tree model, as well as a support vector machine to classify individuals in the OASIS dataset as non-Alzheimer's, Alzheimer's, or converted at an accuracy of approximately 75%. We noticed that a support vector machine with a linear kernel provided the strongest classification power out of all options with all listed optimizations attempted. Within this support vector machine data, between classes we noted that training just using neurological data provided just almost as strong a model as a model with full demographic data added. This is a critical insight, as it demonstrates the promise of neurological data as a metric to reliably train diagnostic algorithms in the future. Furthermore, we noted that the simple inclusion of cognitive data, the MMSE, to the neurological data allowed for the training of models with similar power as the fully informed models. This suggests that cognitive tests may be helpful to administer regularly to best inform these types of diagnoses.

Continuing on this, we would like to briefly mention interesting failure cases of the model. Specifically looking at individuals who were incorrectly characterized in the smaller datasets (neurological data only), we were able to track these individuals to being consistently mis-classified as other features were added. This suggests that the model failure may be consistent and could be due to a non-linear separation between the classes.

Future studies would benefit by adding additional machine learning techniques to further strengthen the classification power demonstrated here, such as boosting. Additionally, the usage of datasets that contain more specific neurological information may provide even greater insight into the potential for diagnostic automation.

## V. References

[1] "10 Early Signs and Symptoms of Alzheimer's." *Alzheimer's Disease and Dementia*, www.alz.org/alzheimers_disease_10_signs_of_alzheimers.asp. Accessed 8 Aug. 2023.

[2] Armstrong, Richard A. "The molecular biology of senile plaques and neurofibrillary tangles in Alzheimer's disease." *Folia neuropathologica* vol. 47,4 (2009): 289-99.

[3] *2023 Alzheimer's Disease Facts and Figures*, www.alz.org/media/Documents/alzheimers-facts-and-figures.pdf. Accessed 8 Aug. 2023.

[4] Tarawneh, Rawan, and David M Holtzman. "The clinical problem of symptomatic Alzheimer disease and mild cognitive impairment." *Cold Spring Harbor perspectives in medicine* vol. 2,5 (2012): a006148. doi:10.1101/cshperspect.a006148

[5] Baglat, Preety, et al. "Multiple Machine Learning Models for Detection of Alzheimer's Disease Using Oasis Dataset." *Re-Imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation*, 2020, pp. 614–622, https://doi.org/10.1007/978-3-030-64849-7_54.

[6] Marcus, Daniel S., et al. "Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults." Journal of Cognitive Neuroscience, vol. 22, no. 12, 2010, pp. 2677–2684, https://doi.org/10.1162/jocn.2009.21407.

[7] Kavitha, C et al. "Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models." *Frontiers in public health* vol. 10 853294. 3 Mar. 2022, doi:10.3389/fpubh.2022.853294

[8] McKinney, Wes. "Data Structures for Statistical Computing in Python." *SciPy* (2010).

[9] Pedregosa et al. "Scikit-learn: Machine Learning in Python." *JMLR 12*, pp. 2825-2830, (2011).