# Diagnosing Alzheimer’s Disease with Machine Learning

### By Eugene Fong

### Spring 2023

### Prof Chaojie Wang’s Capstone in Data Science class

### DATA 606 @UMBC

# OVERVIEW

### PHASE 1 – Proposal & Planning

###### Literature search: Alzheimer’s Disease

###### Diagnosing Alzheimer’s is hard!

###### OASIS dataset

### PHASE 2 – Data prep, EDA, & Data Viz

###### Stats

###### Data viz

###### Filtering

### PHASE 3 – Model training & Deployment

###### Data cleansing & transformation

###### Stratified train/test split

###### Model training

###### Prediction evaluation

# PHASE 1 – Proposal & Planning

# Literature search: Alzheimer’s disease (AD)

Alzheimer’s disease (AD) is a progressive neurodegenerative disease which causes a decline in cognitive functions until death. It is the main cause of dementia and rapidly increasing worldwide, roughly doubling every 5 years. The direct cause is still an issue of hot debate, but the literature has many risk factors listed such as: “increasing age, genetic factors, head injuries, vascular diseases, infections, and environmental factors” (Breijyeh & Karaman, 2020). From a clinical standpoint, AD patients exhibit “memory loss… change of personality… progressive loss of cognitive functions” (Breijyeh & Karaman, 2020). From a biomolecular examination, autopsies have revealed neuritic plaques and neurofibrillary tangles, which are predicted to be caused by amyloid-beta (AB) plaques interfering with acetylcholine (ACh), its receptors, and/or production in the nucleus basalis of Meynart (NBM) in the basal forebrain (Breijyeh & Karaman, 2020). However, whether AB plaques are truly the root cause of Alzheimer’s is now under scrutiny after neuroscientist and physician, Matthew Schrag, reexamined the initial research identifying AB plaques as the root cause by Sylvain Lesne. Some of the results from Lesne’s 2006 papers may have been manipulated (*Potential Fabrication in Research Images Threatens Key Theory of Alzheimer’s Disease*, n.d.). Prior to 2021, the AD drugs available could only address some of the individual symptoms of AD, with no effect on the final outcome of the disease (Breijyeh & Karaman, 2020). Beginning in 2021, the FDA approved 3 new AD drugs: Aducanumab in 2021, Lecanemab and Donanemeb in 2023 – all with their own fair share of controversy:

* “Aducanumab does not cure or reverse [AD]… reduced amyloid plaque levels, but that did not translate to any clinical effect… Potentially serious harms are common” (Woloshin & Kesselheim, 2022).
* “Lecanemab slowed clinical decline by 27%” (*Lecanemab, the New Alzheimer’s Treatment*, n.d.).
* “Donanemab slowed mental decline by 35%” (Reardon, 2023).

These 3 new drugs target AB plaques, but none of them can reverse or stop Alzheimer’s completely, which may be an indication that AB plaques may not be the root cause of Alzheimer’s and that as a result of fabricated research results, researchers have been pursuing the wrong target.

The process of diagnosing AD has changed quite a bit over the decades. The Mini-Mental State Exam (MMSE) was one of the early attempts in 1973 to diagnose AD using a simple questionnaire (*Mini-Mental State Exam (MMSE) Test for Alzheimer’s / Dementia*, n.d.). However, the sensitivity ranged from a low 23% up to 76%, while specificity ranged from 40% - 94% (Arevalo-Rodriguez et al., 2021), but was limited to patients who already had mild cognitive impairment (MCI) and works better with repeated follow-ups to mark the pace of mental degradation converting to AD (Arevalo-Rodriguez et al., 2021). The next major iteration was in 1984 with the National Institute on Aging and the Reagan Institute Working Group on Diagnostic Criteria for the Neuropathological Assessment of Alzheimer’s Disease (NIA-Reagan) criteria, which could diagnose patients after death via an autopsy looking for AB plaques and tangle formation (Beach et al., 2012). This method could only identify 3 combinations of dementia with low, medium and high probability ranges, and had a subset of patients that could not be classified at all, leaving a large patient population underserved (Beach et al., 2012). Further refinement led to the creation of the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) diagnostic in 1991 which has grown to include more methods, such as: autopsy, biopsy, brain scans which can look for more biomarkers of 25 different combinations of AD and be applied in living patients (Beach et al., 2012). Sensitivity improved to 71% - 87%, specificity range of 44% - 70% (Beach et al., 2012). These ranges are certainly an improvement upon the MMSE and NIA-Reagan methods, however, it still leaves a large range of uncertainty.

Recently, there have been attempts to diagnose AD with the help of automated methods and Machine Learning (ML) algorithms. In 2018, a paper by Ammar and Ayed used ML speech analysis to identify AD patients with a 56% - 79% accuracy rate (Ben Ammar & Ben Ayed, 2018). Researchers can also utilize a group of “atlas normalization” techniques to compare brain scans across different patients, while accounting for individual differences in head shape, developmental and internal brain matter. Atlas Scaling Factor (ASF) matches individual head size to the standardized reference atlas of the human brain (Buckner et al., 2004). ASF is used to compute the estimated Total Intracranial Volume (eTIV), which includes all of the internal brain structures (Buckner et al., 2004). Normalized whole-brain volume (nWBV) uses digital segmentation tools to label gray/white matter from cerebral spinal fluid (CSF) and sum up the gray/white matter areas (Buckner et al., 2004). AD patients develop lesions which fill with CSF as parts of the brain waste away, so this is essentially a measurement of brain matter or empty space. In 2022, Goulikar Laxmi Narasimha Deva’s research paper used these measurements provided in one of the Longitudinal OASIS datasets as the basis for ML Alzheimer’s diagnosis and achieved 76% - 86% accuracy, 65% - 80% recall, and 76% - 87% on Area Under the Curve (AUC) scores (Deva, n.d.)

The Open Access Series of Imaging Studies (OASIS) research group “is … aimed at making neuroimaging data sets of the brain freely available to the scientific community” (Marcus et al., 2007). Since 2007, OASIS has been publicly releasing their collection of Alzheimer’s brain scans along with patient biomarkers to encourage open development of improved AD diagnosis and research (Marcus et al., 2007). There are currently 4 OASIS datasets. published from 2007 – 2020, utilizing several different brain scanning techniques, such as: CT, MRI, PET, and looking at different patient cohorts like healthy non-demented controls VS dementia in various stages, as well as longitudinal studies where patients return for follow-up VS cross-sectional studies with just one observation of each patient (Marcus et al., 2007).

I located my dataset on Kaggle, which is a merged subset of the OASIS-1 Cross-sectional dataset sans MRI brain scan images (Boysen, n.d.). This is textual data in a CSV format looking at a cohort of 416 dementia VS non-demented patients (Marcus et al., 2007). They have MRI scans at one point in time and while some do appear again for follow-up, I will not be using the follow-up or Longitudinal data, nor will I be analyzing any of the MRI images (Marcus et al., 2007). The variables measured and their labels in the dataset were:

* ID = Patient ID
* M/F = Gender
* Hand = Handedness
* Age = Age
* Educ = Education level
* SES = Socioeconomic Status
* MMSE = Mini-Mental State Exam
* CDR = Clinical Dementia Rating
* eTIV = Estimated Total Intracrananial Volume
* nWBV = Normalized Whole Brain Volume
* ASF = Atlas Scaling Factor
* Delay = When patients returned within a 90-day window for follow-up MRI scans (for longitudinal study)

TODO – replace with MD chart?

TODO

* Convert into MD format