AI Epidemiology: A linear regression modeling and structured machine learning protocol for the analysis of Alzheimer's Disease genomic data.

Reem Hamdan

I. Personal Section

Jarod Kintz once said, "Alzheimer's is the cleverest thief, because she not only steals from you, but she steals the very thing you need to remember what's been stolen". Alzheimer's Disease doesn't only affect the person who has been diagnosed, it affects family and friends. It affects lives as a whole. Alzheimer's Disease does not only result in a loss of memory, but a loss of all MEMORIES: memories with our loved ones that can somehow just disappear.

My grandmother was my best friend. She laughed when I laughed; she cried when I cried. I wondered why she was suddenly forgetting things she had no problem remembering before. How could my grandma suddenly forget what country she was in? How could my grandma not even remember my name? She had spent many summers with my brother and me, taking care of us. Suddenly, I was caring for her. Mornings spent making her breakfast and giving her ten different kinds of medicine prescribed by her doctor could not distract me from the question that was ravaging my mind: why? What was leading her brain to forget these memories that she had carried for more than 70 years?

I had always been interested in genetics. Our genes carry information that affects everything about us. They are what make us unique. The first thing that came to mind when pondering the causes of Alzheimer's Disease were genes. I was absolutely positive that if genetics could determine our looks and our health, then they were the culprit behind this disease. When the opportunity arose to join my school's research group, I knew that genomics research would be my focus. During the Coronavirus pandemic lockdown, I spent my days on my computer reading neuroscience journals. I created and delivered weekly presentations about different hypotheses regarding Alzheimer's Disease. Compiling the research for these weekly presentations gave me purpose during a stressful time. I was able to find a way to use my passion for good.

My research also included exploring public Alzheimer's disease databases, specifically a database from the Allen Institute which possessed all of the information that I needed to successfully figure out what genes and what uncontrollable factors were correlated with

Alzheimer's Disease. I intended to conduct multiple linear regression analyses and to eventually create a supervised machine learning decision tree model to add an additional angle. My high school research mentor Mr. William Bertolotti was my main motivator. He provided my initial introduction to the Statistical Packaging for the Social Sciences (SPSS) and Python and suggested that I incorporate supervised machine learning into my project. I spent days re-organizing the data and practicing the protocols for data analysis: SPSS for regression modeling and Python packages and machine learning protocols on Jupyter Notebook for the decision tree analyses. I created decision tree models for every one of my variables. What was important about the decision trees was that I was able to visualize connections and clearly see which genes correlated with Alzheimer's Disease.

I am content and appreciative of how my research turned out. I was able to gain much from this journey and was able to grow as a researcher, and most importantly, as a person. I learned to understand the value of hard work, determination, and passion. I want to thank my research mentor Mr. Bertolotti for all of his help. He urged me to pursue science research and he stayed by my side throughout 4 years answering any questions that I had. He has invested so much into helping me and my peers succeed and I will always be extremely grateful. I would also like to thank my grandmother who inspired this entire project. She passed away in February of 2022 and she will never be forgotten. She was my backbone, always listening to my concerns without judgment and urging me to always persevere no matter what. Her patience through multiple adversities has allowed me to acquire the strength to achieve my dreams and continue to strive for greatness. She has inspired so many and I aspire to be as brilliant and inspiring as her one day.

To the researchers of the future, I wish you all the best in your future endeavors. If you are passionate about something, follow it. If you want to research something, do it. Never be discouraged if you can't find what you are looking for. You have to keep looking no matter what. If you do not have a mentor, it is perfectly fine. There are many resources that are available to you one click away. Never give up no matter how difficult it might seem. With every failure, you are a step closer to success. Research is not easy, but the skills and lessons you take away from your journey will enable you to achieve many successes in the future. I also want to encourage aspiring researchers to ask for help. You never know how close you are to connecting with someone that could assist you. Never be embarrassed to ask for advice, there will always be someone willing to lend their time, expertise, or an open ear.

Abstract

Alzheimer's Disease is one of the most common neurodegenerative diseases in the world and is the leading cause of death among those ages 65 and older. An emerging field of science is genomic neurology, which explores the genetic basis for neurodegenerative diseases. The purpose of this project was to identify genes associated with Alzheimer's Disease. Genomic data of patients with dementia and healthy controls were collected from the Allen Institute Genome Browser. The following genes were analyzed: Apolipoprotein (APP), Complement Receptor 1 (CR1), Apolipoprotein E4 (APOE4), Clusterin (CLU). Linear regressions were run and showed the degree to which these four genes predicted AD diagnosis. Structured machine learning in the form of decision tree analyses identified which risk factors, including genes and at what levels, best predicted AD diagnosis. Significant results indicated that CR1, APOE4, and CLU were associated with the diagnosis of AD, but APP showed no significant association with the disease. The results of the study identify other potential pathways for new treatments and can help establish a more informed research process.

1. Introduction

Alzheimer's Disease (AD) is a progressive form of Dementia that causes brain cells to die and degenerate, destroying the memory of the person who has been affected. This disease causes a decline in cognitive function and interferes with daily functioning. It is most prevalent in those over 65 years of age and there is no cure. 11.3 percent of Americans age 65 and older are diagnosed with AD. AD kills more seniors than breast and prostate cancer combined. Among people aged 70, 61 percent of those with AD dementia are expected to die before the age of 80 compared with 30 percent of people without AD — a rate twice as high (Alzheimer's Disease Facts and Figures). Research into the gene or gene combinations which correlate with Alzheimer's Disease may identify a more promising avenue for finding a cure.

The Amyloid β and Tau Protein Hypotheses Explain Mechanisms of Alzheimer's Disease

Neuroscientists have narrowed down the mechanisms of AD to two hypotheses: the Amyloid Beta hypothesis and the Tau protein hypothesis. The amyloid beta cascade hypothesis identifies the causative agent of AD as the deposition of amyloid β protein, which is the main component of the plaques and that the neurofibrillary tangles, cell loss, vascular damage, and dementia that follow are a direct result of this deposition (Ricciarelli and Fedele, 2017). Evidence has accumulated over the last two decades showing that different forms of Aβ can cause synaptotoxic effects and neuronal death in a variety of in vitro and in vivo models. It has been the mainstream hypothesis until now (Kametani, 2018). However, the introduction of new technology has created doubts regarding the validity of this hypothesis. Since Alzheimer's Disease affects people differently, there is not enough evidence to implicate neurofibrillary tangles. It has been consistently shown that Aβ accumulation and deposition do not correlate with neuronal loss and cognitive decline, and that many individuals have significant amounts of amyloid plaque without showing symptoms of memory impairment (Ricciarelli and Fedele, 2017). It has been shown that increased levels of soluble amyloid-β oligomers might lead to synaptic damage and neurodegeneration (Serrano-Pozo, 2011).

An alternative hypothesis that has emerged over the past few years is the Tau hypothesis which states that excessive or abnormal phosphorylation of tau results in the

transformation of normal adult tau into PHF-tau (paired helical filament) and neurofibrillary tangles (NFTs). Mutations that alter function and isoform expression of tau lead to hyperphosphorylation. The process of tau aggregation in the absence of mutations is not known but might result from increased phosphorylation. This highly phosphorylated tau causes behavioral deficits resulting from synaptic dysfunction, axonal transport disruption, and cytoskeletal destabilization in vivo in the absence of neuronal death (Mietelska-Porowska, 2014). Hyperphosphorylated tau disassembles microtubules and creates NFTs. These NFTs, in turn, damage cytoplasmic functions and interfere with axonal transport, which can lead to neuronal death.

NFTs are a major component of both hypotheses, which suggests that it plays a big role in AD pathology. The number of neurofibrillary tangles is tightly linked to the degree of dementia, suggesting that the formation of neurofibrillary tangles more directly correlates with neuronal dysfunction. The accumulation of neurofibrillary tangles and phosphorylated tau species is associated with disturbances of the microtubule network (Brion, 1998).

Research identifies genetic factors associated with Alzheimer's Disease

Genome-wide association studies have confirmed the hypothesis that the APOE ε4 gene has a role in the mechanism of AD and may in fact be the strongest genetic risk factor for AD (Liu, 2013). The presence of this allele is associated with increased risk for early-onset AD. Clinical and autopsy-based study results demonstrate that, compared with individuals with an ε3/ε3 genotype (the third variant), risk of AD is increased in individuals with one copy of the ε4 allele (or two copies among Caucasian subjects). The presence of APOE ε4 is also associated with poorer outcomes following traumatic brain injuries (TBI), regardless of the severity of initial injury. Since TBIs are hypothesized to be associated with AD, it makes sense that the APOE gene and the fourth variant could contribute to a greater chance of developing AD, especially after a TBI has occurred. However, some research suggests that there is no significant relationship between the presence of APOE ε4 and detrimental effects on cognitive performance following TBI (Padgett, 2016).

Apolipoprotein (APP) is another gene that is linked to AD. Most missense variants in APP are associated with autosomal-dominant inheritance of AD. The discovery of mutations in APP that increase or decrease amyloid beta production and the risk for AD provides strong support for the amyloid cascade hypothesis, which hypothesizes that accumulation of amyloid beta is the primary cause of AD pathogenesis (Selkoe and Hardy, 2002). The APP gene is

encoded by 18 exons that are alternatively spliced to produce proteins ranging in size from 695 to 770 amino acids. The Aβ (Amyloid beta) peptide is encoded by parts of exons 16 and 17 (Yoshikai *et. al.,* 1990). To date, 26 pathogenic missense mutations have been reported within the APP gene (TCW and Goate, 2017).

A third gene that is hypothesized to be linked to Alzheimer's Disease is Clusterin (CLU), specifically the C allele at the rs11136000 locus. The CLU gene is the third-strongest known genetic risk factor for late-onset Alzheimer's disease. A recent genome-wide association study has also found the strongest evidence of association with CLU at rs1532278. RS is Reference SNP cluster ID indicating a location in the genome that is known to vary between individuals. The C allele is linked with faster cognitive decline in presymptomatic stages of AD progression (Thambisetty *et. al.*, 2013), and lower memory scores in both AD patients and healthy elderly individuals (Pedraza *et. al.*, 2014).

Complement receptor 1 (CR1) is a gene which promotes phagocytosis of immune complexes and cellular debris, as well as amyloid beta. Meta-analysis has confirmed several disease-associated single nucleotide polymorphisms (SNPs). CR1 currently ranks among the top 10 Alzheimer's Disease risk genes. While the role of CR1 in AD pathogenesis is not yet fully clear, accumulating evidence suggests a dysregulation of complement with effects on inflammation and amyloid accumulation. Clinically, a CR1 risk allele has been associated with faster cognitive decline and greater neuropathology burden in longitudinal aging cohorts (Alzforum: Alzrisk AD epidemiology database).

Machine Learning as an Analysis Method for Alzheimer's Disease Neuromedicine

Machine learning and Artificial Intelligence are beginning to play a large role in many fields. Some fields that utilize Machine learning and AI are of course healthcare, manufacturing, transportation, marketing and sales, and government. These are just a few of the many fields that offer the chance to use this. Machine learning is extremely important because these tools enable organizations to more quickly identify profitable opportunities and potential risks. In the healthcare field, great advancements have been made using machine learning and AI.

Machine learning also has various impacts on the study of Alzheimer's Disease. Usually, AD is diagnosed based on the presence of detecting amyloid-beta abnormality. This is usually done by extracting cerebrospinal fluid, however, this process introduces risks to the patients. The retina of the eye presents an easily accessible window for extracting biomarkers of AD (Tian, 2021). Recent studies have shown that retinal images can display features associated

with early stages of AD and other types of neurodegenerative diseases. Machine learning techniques are capable of recognizing retinal vascular features, which are extremely hard to determine for even human experts. Machine learning accurately reflects the means, standard deviations, and relations of each variable over time to the extent that synthetic data cannot be determined by a logistic regression.

My research question was simply what genes correlate with an Alzheimer's Diseased brain. My goal was to find out if there was a specific gene that had an effect on whether or not a patient had AD. By using the Allen Institute genome browser, I was able to see a wide array of patients and the genes that were present in each lobe of the brain. I have researched and read many journals by neuroscientists studying AD and have developed a list of genes that are believed to be most associated with this disease.

I hypothesized that the presence of the 4th variant of the APOE gene, APP, CLU, and CR1 genes would increase the likelihood of being diagnosed with Alzheimer's Disease as compared to healthy controls.

2. Methodology

2.1. Dependent Variables used in this study

CERAD Score. The classification of the stages of Alzheimer's Disease were independent factors. They do not in fact cause AD, but indicate what degree of damage the patient's brain has sustained. The CERAD score correlates with the neuritic plaques present. The first scale that is used to determine the stages is the CERAD score. CERAD (Consortium to Establish a Registry for Alzheimer's Disease) is a measure of the likelihood of specifically developing AD. It is measured using a 4- point scale. A score of 1 means definite AD, a score of 2 means probable AD diagnosis, a score of 3 means possible AD diagnosis, and a score of 4 reveals that there is no chance of developing AD (Mirra, 1991).

Braak Stages. Another classification of the stages of AD is called Braak stages. Braak staging in AD has six stages — I through VI. The staging focuses on the location of NFTs. Stages I and II are when the NFTs are limited to the transentorhinal region of the brain. Stages III and IV are when the NFTs are in the limbic regions, which includes the hippocampus. And stages V and VI are when the NFTs are extensive in the neocortical regions of the brain. As the stage increases, the severity of the disease also increases (Braak, 2006).

NIA Reagan Score. The last classification that I have included in my research is the NIA Reagan score. The modified NIA-Reagan diagnosis of Alzheimer's disease is based on

consensus recommendations for postmortem diagnosis of Alzheimer's disease. The criteria rely on both neurofibrillary tangles (Braak) and neuritic plaques (CERAD). A score of 1 means the likelihood of AD is high, a score of 2 means the likelihood of AD is intermediate, a score of 3 means the likelihood of AD is low, and a score of 4 means there is no chance of the patient developing Alzhiemer's (RADC, 1997).

2.2. Independent Variables

The dataset that I used was separated in half, with one side being patients with dementia and the other being patients with no dementia. The side of the dataset with no dementia patients acted as a control for the experiment, to see if the expression of the different types of genes were actually different. As you dive deeper into the patients with dementia, there are actually multiple types of dementia present. Aside from Alzheimer's Disease type, there was vascular dementia, multiple etiologies, or no dementia at all.

Genes associated with AD. The Allen Institute genome browser allows users to view the expression of genes in four different lobes of the brain: the neocortex from the posterior superior temporal gyrus, the inferior parietal lobule, white matter underlying the parietal neocortex, and the hippocampus. To narrow down my search, I researched what region of the brain was more vulnerable and susceptible to Alzheimer's Disease. Research from past neuroscientists suggested that the hippocampus, a brain area critical for learning and memory, is especially vulnerable to damage during early stages of AD. Emerging evidence has indicated that altered neurogenesis in the adult hippocampus represents an early critical event in the course of AD (Mu and Gage). Using the database, and guided by previous research, I narrowed down the selection of genes from a few hundred to four genes: APOE, APP, CR1, and CLU.

In this dataset, gene expression was measured by two scales: z-score and log2 intensity. A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean. It is also known as a standard score, because it allows the comparison of scores of different kinds of variables by standardizing the distribution (Mcleod, 2019). The logarithmic scale is used to model proportional changes rather than additive ones, which are typically more relevant (Wilhelm, 2015). The log2 scale is more fine grained than using higher bases. Log 2 measured data is more accurate to biologically detectable changes.

Sex. I chose to analyze the sex of the patients. Sex was actually believed to be one of the factors that may influence the development of Alzheimer's Disease with close to two-thirds of Americans diagnosed with AD are female (Podcasy, 2016). When we first look at gender and the effect it has on our brains, we find that mental illness and sleep patterns are involved. Women have twice the risk of depression compared to men. Depression has implications for cognition across the lifespan because mood and memory map to some of the same brain regions. Studies report that depression is a risk factor for AD dementia in both women and men with estimates as high as a 70% increased risk of AD dementia for those with depression in midlife. Therefore, because the life-long prevalence of depression is greater in women, a diagnosis of depression may have a greater overall impact for AD dementia risk among women (Mielke, 2018). In contrast to depression, men have a greater overall prevalence of sleep apnea, although the prevalence of sleep apnea in women significantly increases after menopause. Sleep apnea and poor sleep quality have been associated with cognitive decline and an increased risk of AD dementia.

Traumatic Brain Injury (TBI). TBI-induced neurovascular injuries accelerate amyloid β production and perivascular accumulation, arterial stiffness, tau hyperphosphorylation and tau/Aβ-induced blood brain barrier damage, giving rise to AD. It is hypothesized that TBI can initiate cerebrovascular pathology, which is causally involved in the development of multiple forms of neurodegeneration including AD-like dementias (Cejudo, 2018). While it is still unclear which mechanisms lead to Aβ accumulation in TBI, autopsies of relatively young TBI patients who died during the acute phase after injury show diffuse Aβ plaques similar to those found in AD patients located in the areas surrounding the lesion sites in both gray and white matter regions (reviewed in Perry *et. al.*, 2016, Johnson *et. al.*, 2010).

Age. Another independent factor that I analyzed is the age of the patients. Of course, Alzheimer's Disease is famous for emerging in those mainly over the age of 65. The greatest known risk factor for Alzheimer's and other dementias is increasing age. While age increases risk, it is not a direct cause of Alzheimer's. After age 65, the risk of Alzheimer's doubles every five years. After age 85, the risk reaches nearly one-third (Causes and Risk Factors of Alzheimer's Disease). I wanted to see when analyzing the data if the trend of old age linked with Alzheimer's Disease was present.

2.3. SPSS Data Analysis Procedures

When analyzing my data using SPSS, I only focused on the connections between the independent factors and specifically Alzheimer's Disease type dementia.

Procedures in SPSS:

1. Entered the data in a new spreadsheet following conversion from google sheets.
2. Recoded all string values to numeric values and gave each diagnosis a specific number.
3. Ran linear and multiple linear regressions.
4. Dependent variables that were run were Alzheimer's probability and presence of Dementia in patients.
5. The independent variables ran were the four genes and their respective z-scores and log2intensities, age, sex, and number of traumatic brain injuries.
6. After running the regression, the data was analyzed to see which independent variables had a significance of $p < 0.05$.

2.4. Procedures in Machine Learning:

1. Data was imported from SPSS as a .csv into Anaconda Navigator, Jupyter notebook for structured machine learning. - why- to see the connection between the variables.
2. I imported numpy and pandas.
3. Import decision tree protocol from sklearn.tree and set the tree depth to 3 branches.
4. I initialized the decision tree, organized the data frame, and declared our predictor and outcome variable.
5. We fit the model to the data frame and designated the program to generate a decision tree visual.

3. Results

3.1. Regression Results

The first dependent variable that was analyzed was the presence of dementia in patients. The independent variable in this study was sex. Sex was found to have a significant effect on likelihood of developing dementia. Males were more likely than females to be diagnosed with dementia by 0.097 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.415	.021		20.193	.000
	Sex	.097	.031	.097	3.120	.002

a. Dependent Variable: Dementia Y/N

Age, another independent variable, was also analyzed in comparison with the presence of Dementia. Age was found to be a statistically significant variable, with a p value of less than 0.05. With every increase of one standard deviation in age, a patient's chance of being diagnosed with dementia increases by 0.107 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.280	.213		-1.314	.189
	Age	.008	.002	.107	3.470	.001

a. Dependent Variable: Dementia Y/N

The 4th variant of the APOE gene was the next independent variable to be analyzed. It was found to be statistically significant in the diagnosis of dementia. For every patient with the 4th variant, their chance of being diagnosed with dementia increased by 0.257 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.390	.017		23.016	.000
	ApoE4 YN	.211	.025	.257	8.535	.000

a. Dependent Variable: Dementia Y/N

APOE4 z score and log2intensity were analyzed next. Both independent variables were found to be statistically significant, and even had the same level of significance. While they did have the same p values, their standardized coefficient betas differed. An increase in the APOE4 z score led to a decrease of possibility of dementia by 0.549 standard deviations. An increase in

the APOE4 Log2intensity led to an increase of the possibility of dementia by 0.553 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-3.529	1.676		-2.106	.035
	APOE Z score	-.336	.142	-.549	-2.364	.018
	APOE4 Log 2 intensity	.627	.264	.553	2.380	.018

a. Dependent Variable: Dementia Y/N

A gene that was found to be significant was CR1. The z-score and log2intensity were both statistically significant. An increase in the CR1 z score led to an increase of possibility of dementia by 0.49 standard deviations. An increase in the CR1 Log2intensity led to an decrease of the possibility of dementia by 0.395 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.010	.154		6.555	.000
	CR1 z-score	.334	.077	.490	4.339	.000
	CR1 log2intensity	-.833	.238	-.395	-3.502	.000

a. Dependent Variable: Dementia Y/N

A gene that was found to be significant was Clusterin (CLU). The z-score and log2intensity were both statistically significant. Each unit increase in the CLU z score was predicted to increase the diagnosis of dementia by 0.195 standard deviations. An increase in the CLU Log2intensity was associated with an increase of the possibility of dementia by 0.119 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.352	.022		16.277	.000
	CLU z score	.108	.017	.195	6.413	.000
	CLU log2intensity	5.920E-5	.000	.119	3.935	.000

a. Dependent Variable: Dementia Y/N

The next dependent variable that was studied was Alzheimer's Disease Probability. There were three options: probable, possible, and dementia: type unknown. The z-score and log2intensity of the gene CR1 was found to be statistically significant. An increase in the z-score of CR1 led to an increase in probability of AD by 0.289 standard deviations. However, an increase in the CR1 log2intensity was associated with a decrease in the probability of AD by 0.441 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.108	.223		9.438	.000
	CR1 z-score	.314	.115	.289	2.726	.007
	CR1 log2intensity	-1.447	.348	-.441	-4.162	.000

a. Dependent Variable: Alzheimer's Probability

Compared to Alzheimer's probability, only the z-score of CLU was found to be significant. As the CLU z-score increased, the probability of AD diagnosis increased by 0.158 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.075	.051		21.252	.000
	CLU z score	.124	.035	.158	3.511	.000
	CLU log2intensity	-1.546E-5	.000	-.031	-.690	.491

a. Dependent Variable: Alzheimer's Probability

The DSM provides a classification system for the diagnosis of mental health disorders for both children and adults. Alzheimer's Disease Type was set as 1, Multiple Etiologies was set as 2, Other or Unknown Cause was set as 3, Vascular Dementia was set as 4, Other Medical was set as 5, and No Dementia was set as 0. The presence of the fourth variant of the APOE gene was found to be statistically significant in the diagnosis of Alzheimer's Disease. For every patient with the 4th variant, their chance of being diagnosed with AD increased by 0.150 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.772	.044		17.537	.000
	ApoE4 YN	.315	.064	.150	4.890	.000

a. Dependent Variable: DSM4

Clusterin's z-score and log2intensity was also found to be statistically significant in the diagnosis of Alzheimer's Disease. As the CLU z-score increased, the diagnosis of AD increased by 0.068 standard deviations. As the CLU log2intensity increased, the diagnosis of AD increased by 0.95 standard deviations.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.771	.056		13.756	.000
	CLU z score	.097	.044	.068	2.212	.027
	CLU log2intensity	.000	.000	.095	3.060	.002

a. Dependent Variable: DSM4

3.2. Machine Learning Results

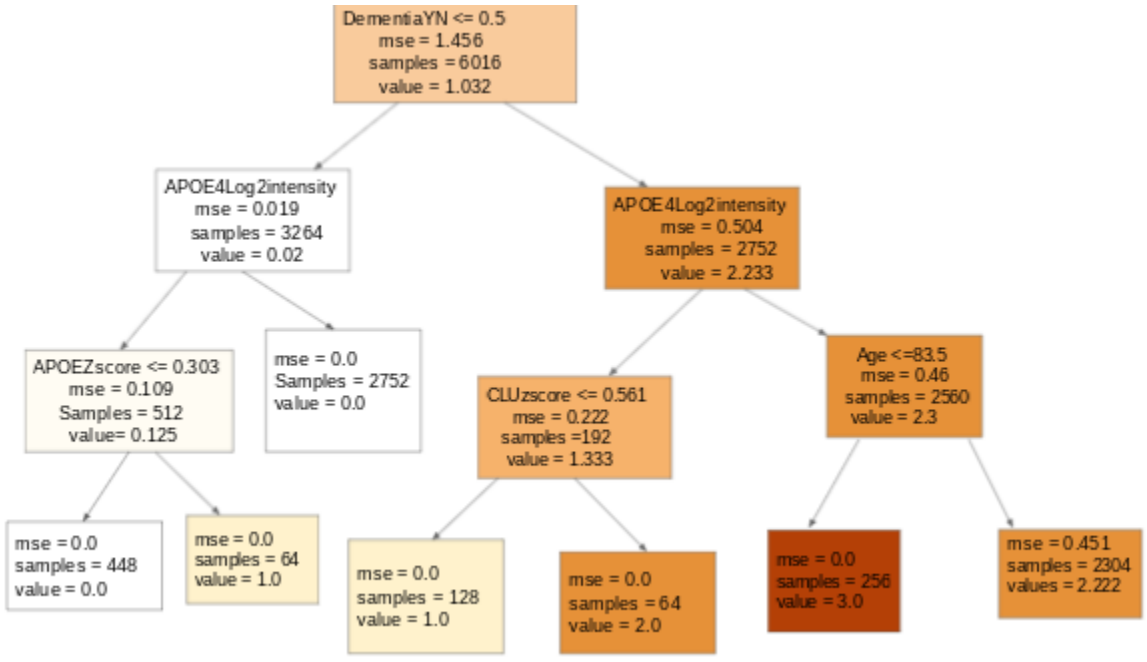

Figure 1. Decision Tree for Alzheimer's Probability as Diagnosed through Dementia

The first question this model asks is if the patient does or does not have dementia. If the patient does in fact have dementia, the next question that is asked is if the APOE4 Log2intensity is greater than 6.277 units. If the patient's Log2intensity is greater than 6.277 units, and if they

are less than 83.5 years of age, they will be most likely to be diagnosed with Alzheimer's disease.

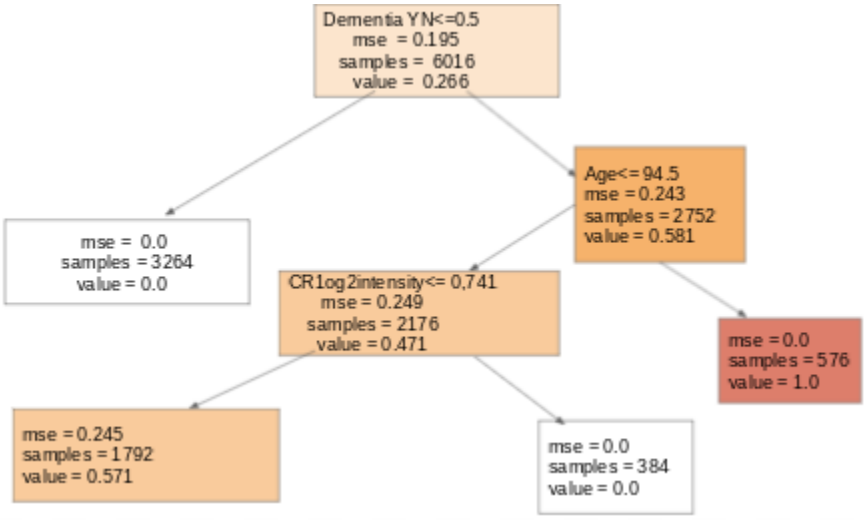

Figure 2. Decision Tree for Patients Diagnosed with AD Using the DSM4 Diagnosis

The first question that was asked was if the patient is diagnosed with dementia. If the patient does have dementia, the next question was if the patient's age was greater than 94.5 years. If the patient's age was above 94.5 years, they will most likely be diagnosed with Alzheimer's disease. If the patient's age was less than 94.5 years, then the next question is if the CR1 log2intensity is more or less than 0.741 units. If the CR1 log2intensity is greater than 0.741 units, they will most likely not be diagnosed with Alzheimer's Disease. If it was less than 0.741 units, then they will be likely to have AD, but less than if they were older than 94.5 years of age.

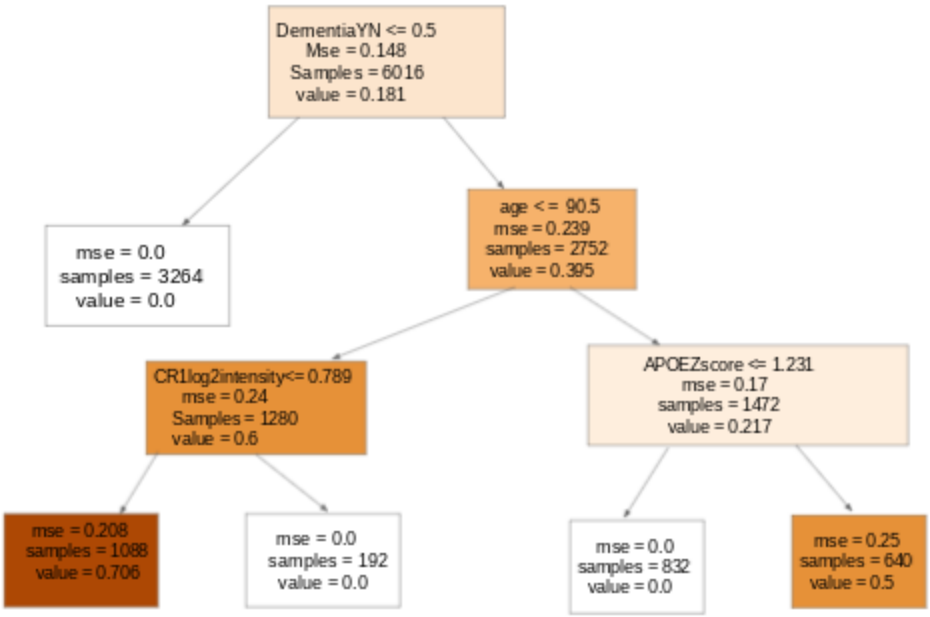

Figure 3. CR1log2intensity was the strongest predictor for Dementia

The first question that was asked was whether the patient had dementia. If the patient did have dementia, the next factor considered was age. If the patient's age was greater than 90.5 years, and their APOEzscore was greater than 1.231, then they were more likely to have a probable diagnosis for Alzheimer's Disease. If the patient's age was less than 90.5 and their CR1 log2intensity was less than 0.789, then they would also have a probable diagnosis for Alzheimer's Disease.

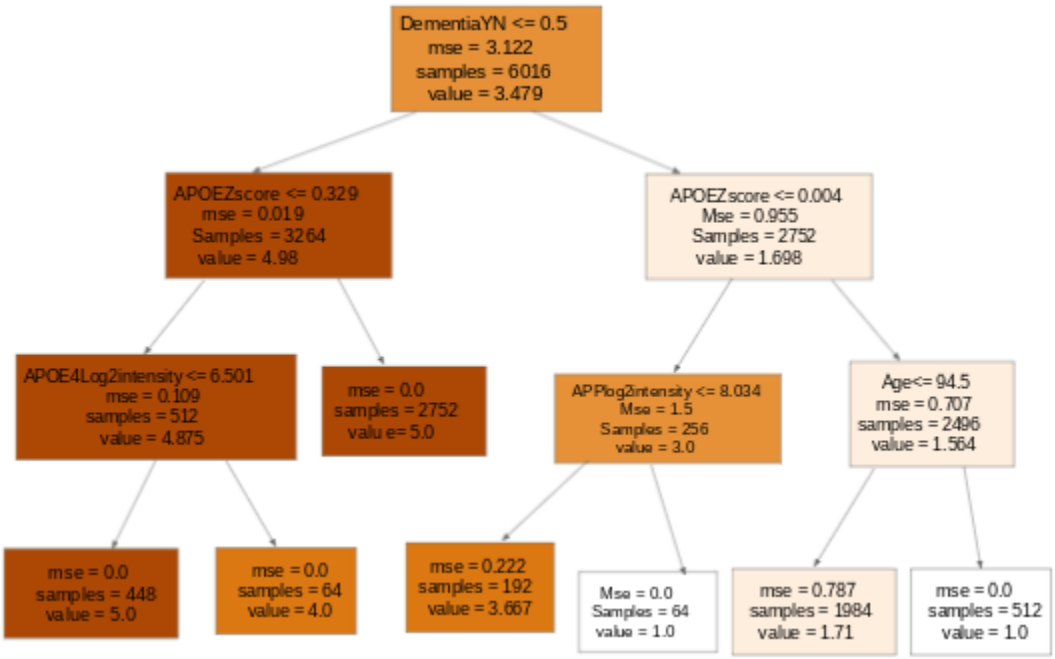

Figure 4. APOE4 was the strongest predictor for Dementia

A probable AD was more likely in individuals that were 90.5 years of age and younger. If the individual was 90.5 years of age and younger and their CR1 log2intensity score was less than 0.789, they had the greatest chance of being diagnosed with Alzheimer's disease.

4. Discussion

I hypothesized that the 4th variant of the APOE gene would lead to an increase in the probability of Alzheimer's Disease. Through genomic analysis, a statistically significant association was found between the diagnosis of AD and the presence of the fourth variant of the APOE gene. An increase in the APOE4 z-score and log2intensity both increased the likelihood of patients developing AD. This was consistent with research and hypotheses in the neurogenetics community (Padgett, 2016).

One of the genes that appeared most likely in the research conducted by Zhao (2020) to be associated with AD was Apolipoprotein (APP). I hypothesized that this gene would be

statistically significant in relation to AD, but no significant relationship was found. Not only was there no correlation between APP and Alzheimer's Disease probability, but there was also no correlation between the diagnosis of dementia and the expression of the APP gene. APP served only as a predictor for AD probability.

Another gene that was hypothesized to have an association with AD was Clusterin (CLU). As the expression of the z-score and log2intensity of CLU increased, the probability of AD diagnosis also increased. This has been supported by research conducted by Thambisetty *et. al.* (2013) and Pedraza *et. al.* (2014) that shows that specifically the C allele at the rs11136000 locus is linked with faster cognitive decline in presymptomatic stages of AD progression and lower memory scores.

The final gene that was analyzed was Complement Receptor 1 (CR1). This gene was hypothesized to be statistically significant and the genomic analysis showed that there was a significant association between AD probability and the expression of CR1. However, only an increase in the z-score was found to increase the likelihood of being diagnosed with AD. An increase in the log2intensity was associated with a decrease in the likelihood of developing AD.

Through research by Mielke (2018), I expected that Alzheimer's Disease would be more present in females. Results indicated no statistically significant relationship between sex and AD probablity; however, there was a significant relationship between Dementia and sex. Males were more likely to be diagnosed with Dementia. Research has shown that depression and mental health issues predict dementia. Furthermore, men who are widowed and have suffered detrimental effects to their mental health are more likely to be diagnosed with dementia than those who did not suffer such experiences (Mielke, 2018).

Dementia and AD have long been associated with an aging population. It was hypothesized and expected that the older a person gets past the age of 90, the more likely they would be to receive a Dementia or Alzheimer's Disease diagnosis. Linear regression results confirmed an association between age and AD predicting that for every year a person aged, their chance of being diagnosed with AD increased. Furthermore, multiple structured machine learning decision tree models supported this hypothesis. However, several decision tree models for the presence of an AD diagnosis found that patients with greatest risk of being diagnosed with AD were younger than 90.5 years of age.

These findings are relevant to today's society because it opens a new discussion. Instead of focusing on the Amyloid-beta and Tau hypotheses, neuroscientists should focus more on the genes themselves instead of mainly the proteins. There have been many inconsistencies and inaccuracies within past research. Millions of dollars were spent on Amyloid beta research,

and then researchers moved on to the Tau hypothesis which although has generated a bit of success, is widely controversial. Even the new FDA approved drug Aducanumab targets Amyloid Beta in the cells. Although some evidence suggests that this drug has been successful, it is still widely criticized and controversial (McGinley, 2021). Why haven't we focused on drugs that minimize the expression of certain genes? Why are we only focusing on Tau and Amyloid Beta? We need to look through a wider lens. Gene expression is the process by which the information encoded in a gene is used to direct the assembly of a protein molecule. I hypothesize that there would be greater success in this research if we started from the root of the process: the gene. Genes make proteins, and if we focus on the genes that make these proteins, then we could have greater success. Any new pathways that can be covered will be beneficial to finding a way to slow this neurodegenerative disease before it kills the brain any further.

Works Cited:

Allen Institute. (n.d.). Brain Map - Brain-map.org. Allen Brain Map. Retrieved August 2021, from https://portal.brain-map.org/

Alzforum: Alzrisk AD epidemiology database. Alzforum: AlzRisk AD Epidemiology Database. (n.d.). Retrieved October 2021, from http://alzgene.org/

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., & Del Tredici, K. (2006, October). *Staging of alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry*. Acta neuropathologica. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3906709/.

Brion, JP. (1998). Neurofibrillary tangles and Alzheimer's disease. European neurology. Retrieved October 2021, from https://pubmed.ncbi.nlm.nih.gov/9748670/.

Causes and risk factors for alzheimer's disease. Alzheimer's Disease and Dementia. (n.d.). Retrieved September 2021, from https://www.alz.org/alzheimers-dementia/what-is-alzheimers/causes-and-risk-factors

Facts and figures. Alzheimer's Disease and Dementia. (n.d.). Retrieved November 9, 2021, from https://www.alz.org/alzheimers-dementia/facts-figures#:~:text=More%20than%206%20million%20Americans%20of%20all%20ages%20have%20Alzheimer's,11.3%25)%20has%20Alzheimer's%20dementia.

Foster, E. M., Dangla-Valls, A., Lovestone, S., Ribe, E. M., & Buckley, N. J. (2019, February 28). Clusterin in alzheimer's disease: Mechanisms, genetics, and lessons from other pathologies. Frontiers in neuroscience. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6403191/.

Johnson, V. E., Stewart, W., & Smith, D. H. (n.d.). Traumatic brain injury and amyloid-β pathology: A link to alzheimer's disease? Nature reviews. Neuroscience. Retrieved November 10, 2021, from https://pubmed.ncbi.nlm.nih.gov/20216546/.

Kametani, F., & Hasegawa, M. (2018, January 30). Reconsideration of amyloid hypothesis and tau hypothesis in alzheimer's disease. Frontiers in neuroscience. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5797629/.

Liu, C.-C., Liu, C.-C., Kanekiyo, T., Xu, H., & Bu, G. (2013, February). *Apolipoprotein E and alzheimer disease: Risk, mechanisms and therapy*. Nature reviews. Neurology. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3726719/.

McGinley, L. (2021, July 16). The controversial approval of an alzheimer's drug reignites the battle over the underlying cause of the disease. The Washington Post. Retrieved November 9, 2021, from https://www.washingtonpost.com/health/2021/07/05/aduhelm-new-alzheimers-drug-amyloid/.

Mcleod, S. (2019, May 17). Z-score: Definition, Calculation and interpretation. Study Guides for Psychology Students - Simply Psychology. Retrieved November 9, 2021, from https://www.simplypsychology.org/z-score.html.

Mielke, M. M. (2018, November). *Sex and gender differences in alzheimer's disease dementia*. The Psychiatric times. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6390276/.

Mietelska-Porowska, A., Wasik, U., Goras, M., Filipek, A., & Niewiadomska, G. (2014, March 18). Tau protein modifications and interactions: Their role in function and dysfunction. International journal of molecular sciences. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3975420/.

Mirra SS;Heyman A;McKeel D;Sumi SM;Crain BJ;Brownlee LM;Vogel FS;Hughes JP;van Belle G;Berg L; (n.d.). The consortium to establish a registry for alzheimer's disease (CERAD). part II. standardization of the neuropathologic assessment of alzheimer's disease. Neurology. Retrieved November 10, 2021, from https://pubmed.ncbi.nlm.nih.gov/2011243/.

Mu, Y., & Gage, F. H. (2011, December 22). Adult hippocampal neurogenesis and its role in alzheimer's disease. Molecular Neurodegeneration. Retrieved November 10, 2021, from https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/1750-1326-6-8.

NIA-Reagan diagnosis of AD. RADC. (n.d.). Retrieved November 9, 2021, from https://www.radc.rush.edu/docs/var/detail.htm;jsessionid=7BF04AAD7D18ADE361BF3DFCBF79466E?category=Pathology&subcategory=Alzheimer%27s%2Bdisease&variable=niareagansc.

Padgett, C. R., Summers, M. J., & Skilbeck, C. E. (2016, October). Is APOE ε4 associated with poorer cognitive outcome following traumatic brain injury? A meta-analysis. Neuropsychology. Retrieved November 10, 2021, from https://pubmed.ncbi.nlm.nih.gov/26986748/.

Pedraza, O., Allen, M., Jennette, K., Carrasquillo, M., Crook, J., Serie, D., Pankratz, V. S., Palusak, R., Nguyen, T., Malphrus, K., Ma, L., Bisceglio, G., Roberts, R. O., Lucas, J. A., Ivnik, R. J., Smith, G. E., Graff-Radford, N. R., Petersen, R. C., Younkin, S. G., & Ertekin-Taner, N. (2014, March). Evaluation of memory endophenotypes for association with CLU, CR1, and Picalm variants in black and white subjects. Alzheimer's & dementia : the journal of the Alzheimer's Association. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3815516/.

Perry, D. C., Sturm, V. E., Peterson, M. J., Pieper, C. F., Bullock, T., Boeve, B. F., Miller, B. L., Guskiewicz, K. M., Berger, M. S., Kramer, J. H., & Welsh-Bohmer, K. A. (2016, February). Association of traumatic brain injury with subsequent neurological and psychiatric disease: A meta-analysis. Journal of neurosurgery. Retrieved November 10, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751029/.

Podcasy, J. L., & Epperson, C. N. (2016, December). *Considering sex and gender in alzheimer disease and other dementias*. Dialogues in clinical neuroscience. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5286729/.

Ramos-Cejudo J;Wisniewski T;Marmar C;Zetterberg H;Blennow K;de Leon MJ;Fossati S; (n.d.). *Traumatic brain injury and alzheimer's disease: The cerebrovascular link*. EBioMedicine. Retrieved November 9, 2021, from https://pubmed.ncbi.nlm.nih.gov/29396300/.

Ricciarelli, R., & Fedele, E. (2017). The amyloid cascade hypothesis in alzheimer's disease: It's time to change our mind. Current neuropharmacology. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5652035/.

Roussotte, F. F., Gutman, B. A., Madsen, S. K., Colby, J. B., Thompson, P. M., & Alzheimer's Disease Neuroimaging Initiative. (2014, May 7). Combined effects of alzheimer risk variants in the CLU and ApoE genes on ventricular expansion patterns in the elderly. The Journal of neuroscience : the official journal of the Society for Neuroscience. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012312/.

Selkoe, D. J., & Hardy, J. (2016, June 1). The amyloid hypothesis of alzheimer's disease at 25 Years. EMBO molecular medicine. Retrieved November 10, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888851/.

Serrano-Pozo, A., Frosch, M. P., Masliah, E., & Hyman, B. T. (2011, September). Neuropathological alterations in Alzheimer’s disease. Cold Spring Harbor perspectives in medicine. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234452/#A006189C98.

TCW, J., & Goate, A. M. (2017). Genetics of β-amyloid Precursor Protein in Alzheimer's Disease. Cold Spring Harbor perspectives in medicine. Retrieved September 2021, from https://pubmed.ncbi.nlm.nih.gov/28003277/

Tian, J., Smith, G., Guo, H., Liu, B., Pan, Z., Wang, Z., Xiong, S., & Fang, R. (2021, January 8). Modular Machine Learning for Alzheimer's disease classification from retinal vasculature. Scientific reports. Retrieved August 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7794289/

Thambisetty, M., Beason-Held, L. L., An, Y., Kraut, M., Nalls, M., Hernandez, D. G., Singleton, A. B., Zonderman, A. B., Ferrucci, L., Lovestone, S., & Resnick, S. M. (2013, March 1). Alzheimer risk variant clu and brain function during aging. Biological psychiatry. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488132/.

Wilhelm, J. (2015). Why do we usually use log2 when normalizing the expression of genes? Research Gate. Retrieved September 2021, from https://www.researchgate.net/post/Why_do_we_usually_use_Log2_when_normalizing_the_expression_of_genes

Yoshikai S;Sasaki H;Doh-ura K;Furuya H;Sakaki Y; (n.d.). Genomic organization of the human amyloid beta-protein precursor gene. Gene. Retrieved November 10, 2021, from https://pubmed.ncbi.nlm.nih.gov/2110105/.

Zhao, J., Liu, X., Xia, W., Zhang, Y., & Wang, C. (2020, August 4). Targeting amyloidogenic processing of APP in alzheimer's disease. Frontiers in molecular neuroscience. Retrieved November 9, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418514/.