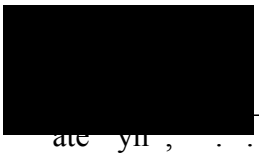
# MACHINE-LEARNING AIDED DIAGNOSIS OF ALZHEIMER'S DISEASE

Hazel Genevieve Helfman

Engineering Honors
The University of Texas at Austin

December 2023

Benjamin Keitz, Ph.D.
McKetta Department of Chemical Engineering
Supervising Professor



ate yn , . .
McKetta Department of Chemical Engineering
Second Reader

## Abstract

Alzheimer's disease is a neurodegenerative disorder characterized by the accumulation of amyloid-beta proteins in the brain, leading to loss of neuronal function and eventual death. Though the incidence of Alzheimer's has risen in recent years, in no small part due to increasing lifespans, there has been little progress in the diagnosis and prevention of the disease. Diagnosis premortem is possible, but mainly through costly imaging or invasive brain biopsies, the latter of which is not recommended due to the possibility of further brain damage in the AD patient. Furthermore, AD treatments are difficult to study due to the difficulty of identifying patients as well as the diseases' stubborn progression. Thus, there is an area of opportunity in accurately identifying these patients for both diagnostic and therapeutic purposes. There are many biomarkers correlated with the presence of AD, whether that be noticeable brain damage via scanning, the biomarkers of neuron cell death, or latent biomarkers which may cooccur in the progression of the disease. Given that these are non-linear relationships, computer-aided diagnosis may help in elucidating the diagnosis of AD. Random Forest models, given their ability to generate human-understandable trees and decision surfaces, are primed to assist medical professionals with the diagnosis of AD. This thesis analyzes several such models and evaluates their accuracies, as well as providing an overview of the state of the computer-aided medical diagnostics field.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

I would like to thank my advisor, B.K. Keitz and my second reader, Nathaniel Lynd, for assisting me in the writing of this thesis. I would also like to thank my friends and family for getting me through these past four and a half years of college; their support has meant everything to me. This paper is for my grandfather; may the anguish of Alzheimer's be someday wiped from Earth and, as with smallpox, be referred to with *was* rather than *is*.

# 1. Diagnostic Models of Alzheimer's Disease

As worldwide average life expectancies have drastically increased in the modern era, so have rates of Alzheimer's disease, a neurodegenerative disorder that results in loss of cognitive facilities, memory, and ultimately brain function (Rocca & White, 2011). Though there is no cure to the disease, there are medicines – acetylcholine (AChE) inhibitors and memantine among them – which may increase patients' quality of life and slow disease progression (National Health Service, 2021). Thus, early diagnosis is of the utmost importance.

Unfortunately, the diagnosis of Alzheimer's is nontrivial, requiring an accounting of symptoms which may be unreliable given that they come from the patient with memory loss themselves. The accounts of people close to the patient is required, but people are fallible. Memory tests may show cognitive decline, but not necessarily that it is caused by Alzheimer's. CT scans can show mental degeneration, but these are rarely monocausal, leaving us with blood tests which may provide the most accurate diagnosis (National Institutes of Health, 2022).

Of interest in regard to biomarkers are homocysteine and isoprostane. Patients with Alzheimer's tend to have much higher blood plasma levels of homocysteine than the general population, controlling for age. There is debate about whether hyperhomocysteinemia is a causative factor or marker of Alzheimer's, given that a deficiency of B vitamins (which also may cause cognitive decline) is a cause of hyperhomocysteinemia itself, and that that it causes cardiovascular issues which may be upstream of a future Alzheimer's diagnosis. Regardless of whether it lies upstream or downstream of the progression of Alzheimer's, homocysteine is a valuable biomarker in its diagnosis (Zhuo, Wang, & Pratico, 2011). Alzheimer's progression is marked by brain damage and isoprostane levels are elevated in patients with brain damage due to

lipid peroxidation. Studies consistently find that while isoprostane levels are higher in Alzheimer's patients, they are not indicative enough to diagnose Alzheimer's in itself (Irizarry, Yao, Hyman, Growdon, & Practico, 2007).

This thesis seeks to generate a model for the diagnosis of Alzheimer's, drawing upon the indicative studies showing elevated levels of isoprostane and homocysteine in patients. To do so, a random forest regression upon a large dataset of Alzheimer's patients will be generated, optimizing the model via hyperparameterization (the modification of parameters used in creating this regression such that a more accurate model will be produced). An initial quantitative check of the hypothesis was generated, ensuring that there exists a relationship between homocysteine, isoprostane, and Alzheimer's progression scores. Further, keeping with industry standard, the data was split into a 90/10 train/test such that we could train the random forest regressor and allow the hyperparameterization to perform a random search. This thesis details the creation of the model and outcomes from its predictions.

## 2. Overview of Alzheimer's Disease

Alzheimer's disease (AD) is a debilitating neurodegenerative disorder caused by agglomeration of amyloid-beta (Aβ) and tau proteins, causing inflammation, oxidation, and ultimately neuronal degeneration, in turn causing symptoms of dementia (Blennow, de Leon, & Zetterberg, 2006). This process is irreversible and hypothesized to be exponential – once initial plaques form, healthy proteins interact with them, conforming to their shape, which in turn increases the seeding population of misshapen proteins, and so on (Kawarabayashi, et al., 2001). This accumulation results in mild cognitive impairment, often akin to that of aging, making it difficult to determine the exact onset of the disorder given the common attribution of senility to

ageing. In general, this differentiation between senility and AD is difficult. However, AD causes a drastic loss of memory ability, cognitive function, and general functioning over that generally attributed to aging. Diagnosis rarely happens before the symptoms become debilitating for this reason.

The inability to diagnose during the prodromal phase of AD has brought the discussion of biomarkers in the clinical setting to the forefront, though this is still an emerging field, especially given those behaviors which increase the risk of AD – smoking, drinking, high cholesterol, diabetes – may cause both upstream and downstream effects upon the brain. Also, some biomarkers, such as homocysteine, carry their own secondary effects through the body on cardiovascular health, raising a classic chicken-and-egg problem. Some diagnostic methods, namely brain imaging, may identify neurodegeneration via visual changes in grey matter and blood flow, but this is not easily differentiated from other forms of brain damage and age-related neurodegeneration, not to mention the other forms of dementia. Aβ and tau protein presence in the brain is indicative of AD, however, these tests are invasive, and are not justifiable in a normal doctor's office. Thus, the field has turned to markers contained in bodily fluid such as blood and urine (Gunes, Aizawa, Sugashi, Sugimoto, & Rodrigues, 2022).

### 2.1. Homocysteine

B-vitamin deficiency is associated with both neurological decline, as seen in Korsakoff syndrome among alcoholics, and hyperhomocystemia itself causes both worse cardiovascular outcomes and oxidative stress in the body, identifying the disease as both a possible risk factor and cause of AD itself. This multifold relationship led to a meta-analysis stating that there is a definite link between AD and high homocysteine levels and that this high concentration comes

before the diagnosis of AD, making it a prime biomarker for diagnosis (Morris, 2003). While solving the association between B-vitamins, hyperhomocysteimia, and AD is tangled, and prospective studies for curing the former disease via B supplementation may not actually cure the outcomes of the disease, it is nonetheless a promising avenue for further research (Martí-Carvajal, Solà, Lathyris, & Dayer, 2017).

### 2.2. Isoprostane

Isoprostane, specifically F2-isoprostanes, are a less studied biomarker for the progression of AD. First discovered in the 1990s, they are a marker of oxidative stress as they are products of lipid peroxidation and cell death, which makes them ideal for the detection of neuron death in the progression of AD, as well as other cerebral and cardiovascular issues such as stroke and brain injury. Furthermore, they are stable as compared to other biomarkers and can be detected in urine, plasma, and breath, making their detection notably noninvasive (Janssen, 2001).

Specifically, in AD, isoprostane levels are elevated, and an association between these levels in blood and AD has been found. Aβ and tau accumulation in the brain leads to oxidative stress, specifically targeting neurons, causing a rise in levels of isoprostane as cell death occurs. While increased levels are associated with the progression of AD, they do not seem to predict the incidence of the disease, making it a prime biomarker for further research. Complicating the issue is that isoprostane is a marker of general cell breakdown, with studies finding higher levels in patients with cystic fibrosis, degenerative disorders, smokers, and others with negative health behaviors. Thus, any causative relationship would have to be carefully discovered and many variables need to be controlled for (Trares, Chen, & Schöttker, 2022).

## 3. Data Analysis

Data for used in this thesis was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), run by the University of Southern California, which takes longitudinal data on various biomarkers and AD progression from Alzheimer's patients, compiling them for study. Data collection began in 2004 and has extended to the present day in various phases. Aside from the biomarker and disease progression data used in this paper, further data in the form of brain scans were taken to assess disease progression, which itself presents a promising opportunity for those interested in computer vision and machine learning to create a predictive model of Alzheimer's based upon changes from baseline MRI/PET data (Alzheimer's Disease Neuroimaging Initiative, 2017).

To begin the process, ADNI data was sorted and cleaned for processing. To do so, each individual participant's unique RID had to be matched across datasets in order to sort patients and split them via their ID as an anonymous identifier. After matching, graphs of isoprostane levels, homocysteine levels, and disease assessment scores were generated as a sanity check to ensure that the data had a visual correlation. As an aside, disease assessment scores track the progression of AD by testing cognitive abilities affected by neurodegeneration such as memory. An initial relationship was found among the variables, so a model was built.

Furthermore, to test the positive correlation between elevated homocysteine and isoprostane levels and the incidence of AD, a multilinear regression was run on the data. This resulted in a linear regression with a poor $R^2$ of 0.0014 and coefficients of 0.089 and 0.16 for isoprostane and homocysteine respectively. While this verifies that homocysteine has a larger correlation with the incidence of AD, it also demonstrates the difficulty past researchers have had in generating diagnostic models based on biomarkers – the correlation between the two is

very poor and it is difficult to draw conclusions as to cause and effect or on diagnostic validity. The latter is very important; the diagnosis of AD is a lifechanging event. Receiving a false diagnosis causes unnecessary dread within patients; a false negative result in a failure to plan for when AD progresses, resulting in worse outcomes for the patient and their family.

## 4. First Model

A random forest classifier was selected due to its decision tree structure. In a random forest classifier, many decision trees are built, with "branches" being logical operations upon data which eventually select a "leaf", or class (Studer, Ritschard, Gabadinho, & Müller, 2011). As the decision trees are fitted to the data in training, a model is built. When testing data is sent through the model, each individual tree classifies the data; the average prediction of the trees is returned as the output for the classifier. This model, in particular, was chosen as the decision tree structure resembles medical diagnostics in its own right, and therefore may be understandable to a medical professional given a simplified version of the model. Thus, future feedback from doctors may be taken into account for diagnostics.

First, as a baseline, the ADNI data grouped by RID was broken into a 90%/10% train and test split. Then the training data was fed into the random forest regressor. This was set up to establish a baseline performance before improvements were made to the model. Using this model, an accuracy of 76.0% was achieved, which was below what is optimal for diagnostic practice; thus, further improvements were needed.

### 4.1. Hyperparameterization

To improve the model, hyperparameterization was implemented. Any regressor has various variables which may be tuned in order to improve the accuracy of the model. However,

there are many variables with more possible values, creating an insurmountable number of permutations for one person to execute to find a global optimum. To decrease the time spent optimizing these parameters, hyperparameterization executes a random search over the universe of possible parameters for the regression model such that this search is automated. Other methods are possible, as a local optimum can be searched for in a shorter amount of time, but for the purposes of this thesis, a random search was sufficient.

In the hyperparameterization, the number of estimators, maximum number of features, maximum depth, minimum samples split, minimum samples leaf, and bootstrap variables were changed. First, the number of estimators varied between 200 and 2000 with a distance of 10; this variable represents the number of decision trees generated for the regression. The maximum number of features refers to the maximum number before creating a new branch in a decision tree, this is decided by either a square root function or automatic built-in function. The maximum depth is the highest level a tree may reach before being terminated, set between 10 and 110 with a distance of 11. The minimum samples split sets the lowest amount of data before a new branch is created, with values of 2, 5, or 10. The leaf variable sets the same for a leaf at 1, 2, or 3. Bootstrap determines whether points are sampled with replacement. These options allow the random search to iterate through 64,800 permutations to find a global optimum for the model's parameters (Koehrsen, 2018). A graph of the search is provided below.

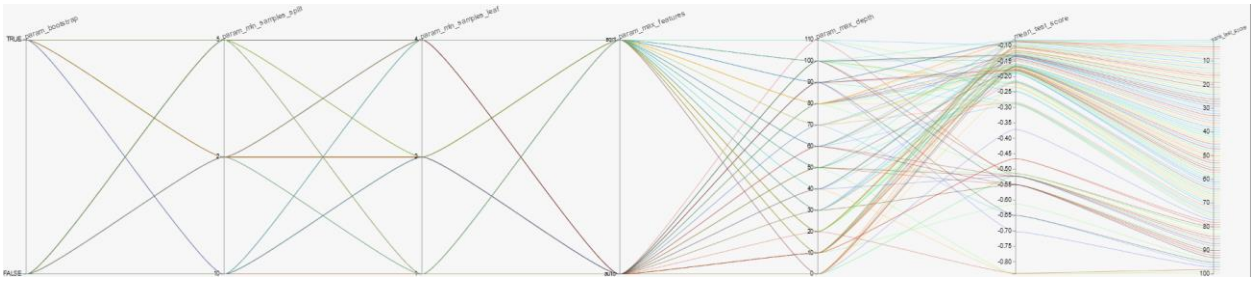


Figure 1. Graph of Parameter Permutations and Test Scores

This graph may be interpreted via each line, for which each represents a permutation of the random forest's parameters. As the line travels from left to right, it selects a unique combination of these parameters as described above; each model is tested against the validation data in the model such that a test score representing the difference from real world data is generated, and each instance is ranked to produce the final model and parameters that will be analyzed.

To analyze the effect of each parameter on the accuracy of the model, a multilinear regression was performed on the testing data, with each parameter coded to numerical values. Iterating over the data, the bootstrap parameter had the most effect, with the maximum features after it, minimum samples leaf following, and minimum samples split. The other parameters - # of estimators and maximum depth – were negligible, with p-values of 0.83 and 0.73 respectively. The regression output is below.

Table 1. Multilinear Regression on Parameters

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | -0.619 | 0.044188 | -14.0085 | 1.22E-24 |
| param_n_estimators | 3.8E-06 | 1.8E-05 | 0.211426 | 0.833018 |
| param_min_samples_split | 0.007059 | 0.003 | 2.352634 | 0.020749 |
| param_min_samples_leaf | 0.029897 | 0.007883 | 3.792709 | 0.000265 |
| param_max_features | 0.228485 | 0.019953 | 11.45102 | 1.85E-19 |

| | | | | |
|---|---|---|---|---|
| param_max_depth | -9.6E-05 | 0.00028 | -0.34266 | 0.732625 |
| param_bootstrap | 0.242507 | 0.019579 | 12.38632 | 2.18E-21 |

We can visualize these parameters' effects by graphing them against the test score, where a test score closer to zero represents a higher degree of accuracy. Taking a closer look at the graphs, there are several conclusions. First, a higher number of estimators – on average – results in a higher mean test score. However, the low variance of its outcomes kept it from outperforming the final model. In another model, it seems that there are two regions of optimality ranging around 600-800 estimators and 1600. Further, we can see that the mean test score for each option of minimum samples split is about equal; all that differs is the variance in outcome. Minimum leaf samples has a similar result; the mean for each option is about equal; if a few percentage points of performance need to be added to a model, it seems that a higher value results in slightly better performance.

The maximum features parameters result in some interesting outcomes. For the multilinear regression, 'auto' was coded to 0 and 'sqrt' to 1. For the automatic setting, where the maximum number of features is equal to the number of features, there was a much lower mean test score than that where the square root of the number of features was equal to the maximum number of features. However, the variance of the former was much higher; resulting in some outcomes more optimal than would be expected given its low average.

The maximum depth parameter demonstrated great variation; with two islands of high-performance centering around values of 20 and 80; however, the lowest values exhibited the highest variability and could conceivably result in metastable outcomes. Finally, the bootstrap parameter, with 'false' coded to 0 for the regression and 'true' coded to 1, had latter far outperform the former, though again a high region of variability for false implies that there are

some models which may outperform the current model, though these may have been inaccessible in the first random search.

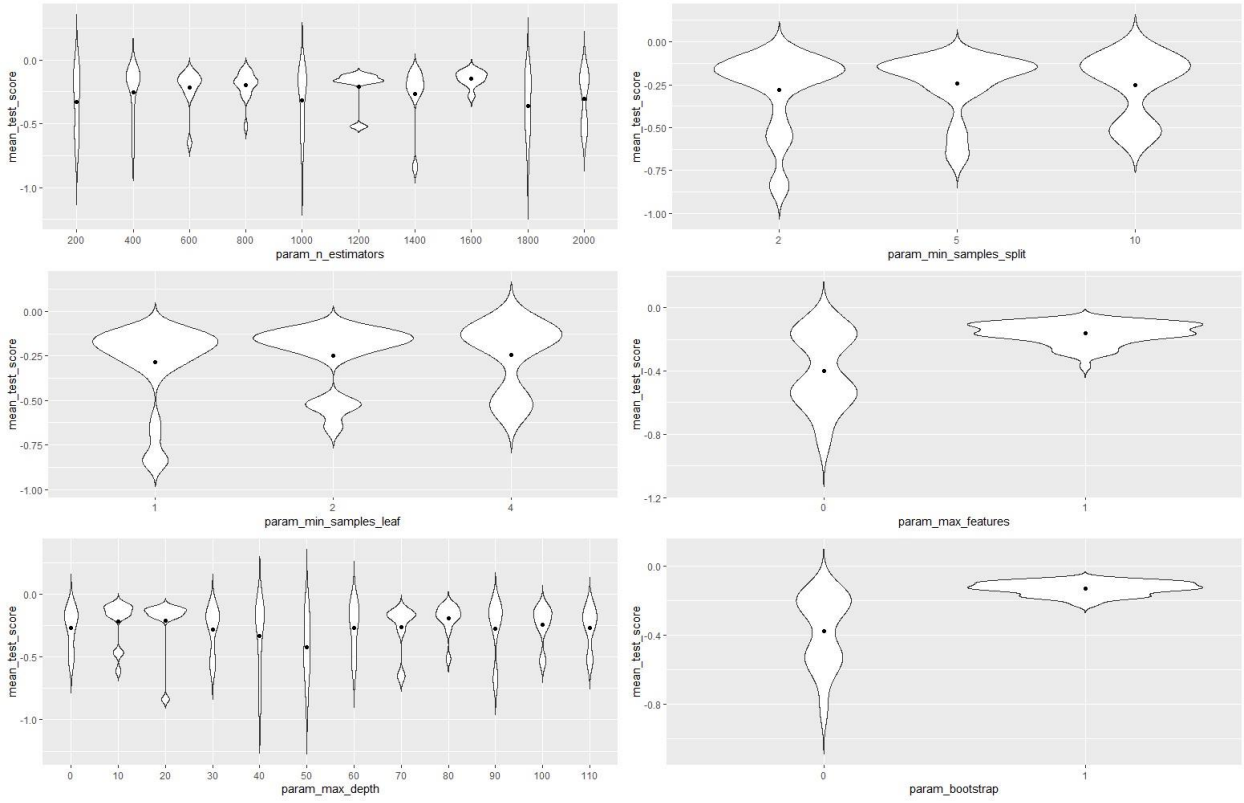



Figure 2. Graphical Representation of Parameter Scores

To illustrate the importance of hyperparameterization, a base model was built with one specified parameter – the number of estimators being 10 – resulting in an accuracy of 29.07% when compared to the testing dataset. Thus, the care taken in the hyperparameterization step resulted in an accuracy improvement of 50% over baseline. Any training success without this essential step is up to chance, and a future paper could explore the use of a grid search for parameters rather than the random search used to gain a higher degree of accuracy in training the model.

Another reason for not continuing with hyperparameterization is that this can cause the model to overfit on data. In other words, it performs perfectly on the training data, but only that, and when applied to data in the testing set or in real-world applications, it does not accurately predict outcomes. Avoiding this outcome is contingent on testing with real-world data to ensure that the model makes logical predictions and watching accuracy predictions; if one obtains predictions that are near-perfect, the model is likely overfitted.

### 4.2. Evaluation

Using the 10% of data set aside when the model was built, model accuracy was tested, finding that the hyperparameterized model has an accuracy of 88.34% with the following parameter values:

Table 2. Selected Random Forest Regression Parameters

| N Estimators | Min Samples Split | Min Samples Leaf | Max Features | Max Depth | Bootstrap |
|---|---|---|---|---|---|
| 400 | 2 | 4 | Sqrt | 10 | true |

To properly visualize this data, a graph of the results versus the actual results was generated, with actual Alzheimer's scores as circles, and predicted as Xs.

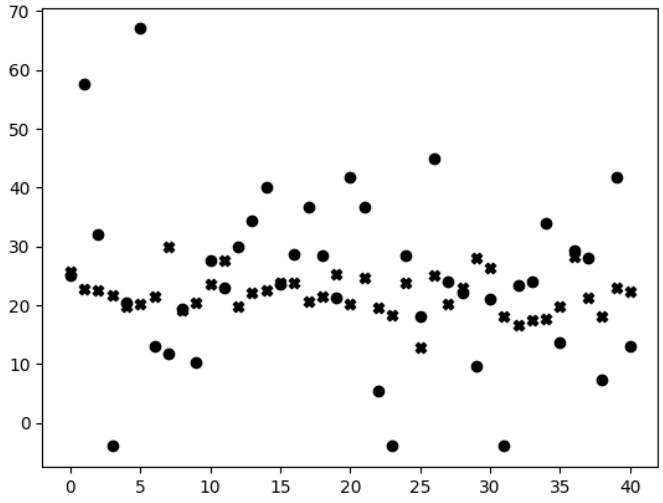



Figure 3. Random Forest Classifier Residuals

From this graph, it is evident that the model is able to predict Alzheimer's progression with reasonable accuracy, though there are several differences. The model has a much lower variance in outputs as compared to real-world data; it seems reluctant to classify past a score of 30 though real-world scores may extend to 70s. Despite this, it does vary with real-world data – i.e., as the ADNI testing score increases, so does the model, and it decreases the same as well, resulting in the prior accuracy statistic.

Another statistic of interest in evaluating machine learning models is recall, otherwise a confusion matrix, which compares predicted and actual labels from a model such that one can view true positives and negatives against false predictions. In a diagnostic model, this is especially important, as a false diagnosis of Alzheimer's can be life-altering; a false positive causing undue stress in patients and a false negative resulting in a failure to plan for the disease's progression. Comparing predicted and actual labels, we obtain the below graph, where a lighter color represents a higher-frequency outcome.

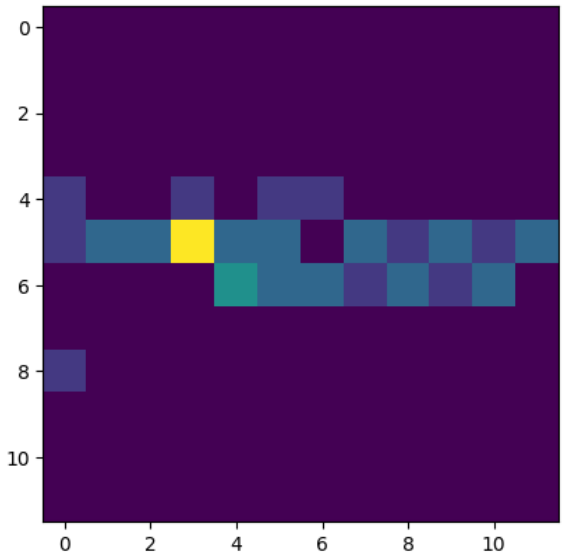



Figure 4. Random Forest Classifier Confusion Matrix

This graph identifies several areas for improvement. Visually, the ideal shape for a confusion matrix is a diagonal line from the top left to bottom right, as this indicates a higher ability for precise predictions. However, this graph does indicate diagnostic ability, as there is a group which demonstrates AD diagnoses and that without, though this is very noisy. To view this, the resolution of this graph was reduced. The general cutoff for AD diagnosis on the ADAS is 17 (Monllau, et al., 2007).

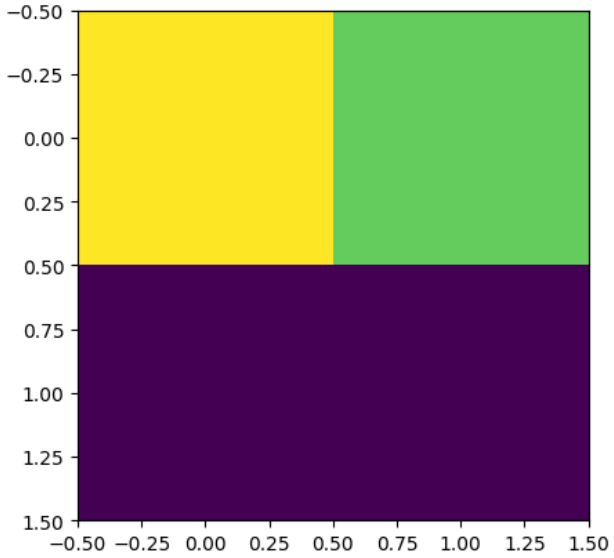



Figure 5. Simplified Confusion Matrix

Lowering the resolution of the graph results in the above, where the top left corner is the number of successful AD diagnoses (22), the top right corner the number of false positives (!7), the bottom left false negatives (1), bottom right negative diagnoses (1). From this, we can conclude that the false negative rate must be decreased to have a successful diagnostic model.

## 5. Retraining the Model

In this thesis, an initial assumption was that the ADAS score should predict the diagnosis of AD in a continuous distribution. This proves inaccurate with the above data. Instead, a superior way to build this model is via discrete variables. After all, a patient either has AD or does not. To visualize these differences, the above graphs plotting the relationship between homocysteine and isoprostane levels versus ADAS scores were regenerated, with red values indicating lack of AD and blue values indicating AD. Here, there are two separate groups clearly

visible. This indicates a higher probability of success for generating a diagnostic model than attempting to predict AD progression via ADAS scores, as it reduces model variability.

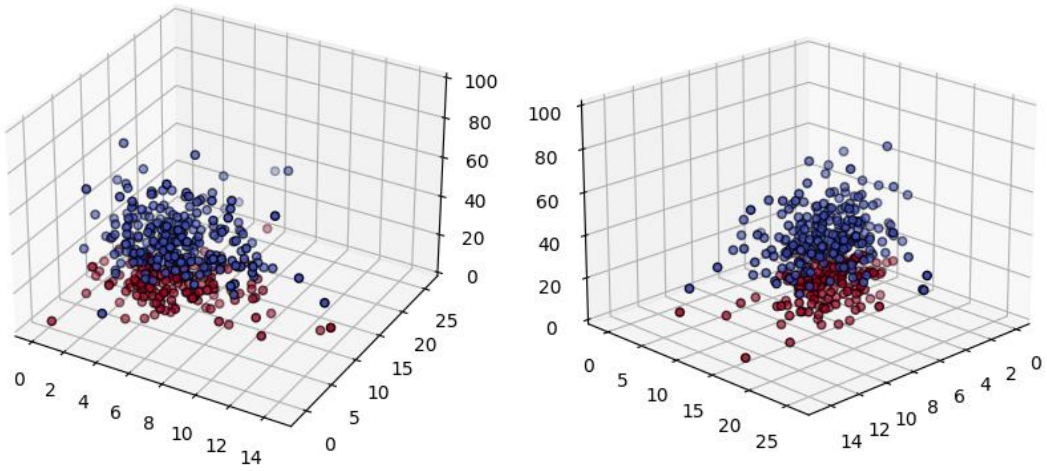



Figure 6. Homocysteine and Isoprostane versus ADAS Scores

To rebuild this model, an ADAS cutoff of 17 was used as in the literature (Monllau, et al., 2007). Given that a discrete variable is now used, a random forest classifier will be used in the place of a regressor. Regressors are typically used for continuous data such as ADAS scores. Given that the model type changed, hyperparameterization was rerun as well, with the same parameter ranges as before. As well, the method for splitting the data into training and testing was changed to sklearn's train_test_split as this reduces variability during training. Furthermore, given that there are less patients without AD than with AD, to ensure model accuracy, the training data was stratified such that it receives more control patients while training to ensure that the model does not overfit on patients with AD, which may have contributed to the high false positive rate in the prior model.

Hyperparameterization was again used with the same parameters, producing the below plot of permutations. To analyze each instance's effect on the model, a multilinear regression was run, producing the below table. In this, the number of estimators used in the model, the minimum number of samples required to split a branch, the minimum number of samples to create a leaf, and the bootstrap parameter were statistically significant. This differs from the prior model, which was not dependent on the number of estimators, but was dependent on the maximum number of features per leaf. In other words, this model requires a larger forest of decision trees to average the results of; the prior model requires greater tuning of the algorithm used to create a branch split.

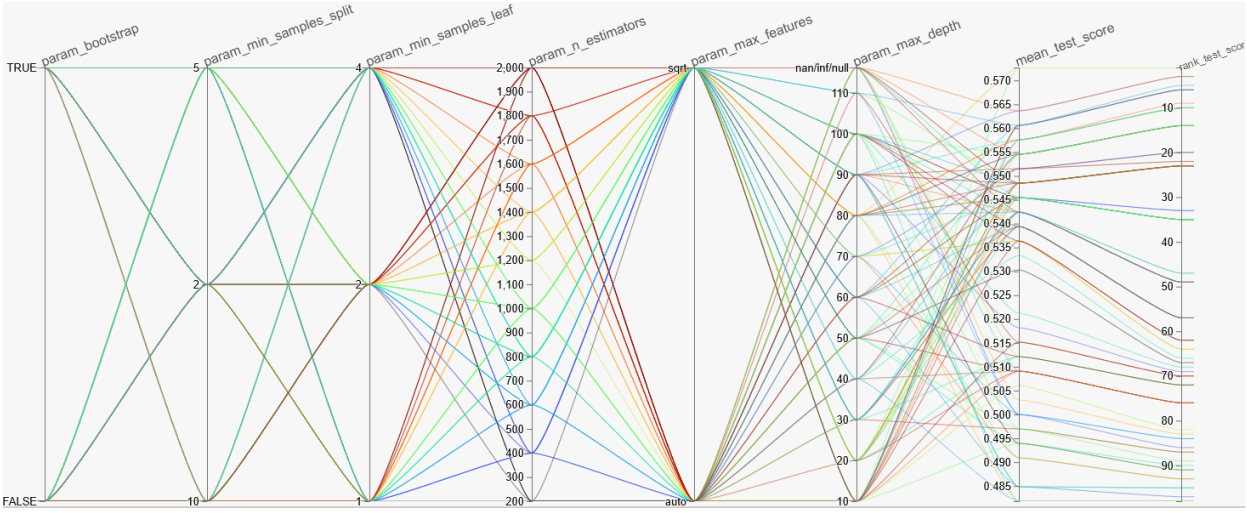



Figure 7. Random Forest Classifier Parameter Scores

Table 3. Multilinear Regression of Random Forest Classifier Parameters

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 0.481538 | 0.005076 | 94.85896 | 2.38E-94 |
| param_n_estimators | 5.06E-06 | 2.06E-06 | 2.452233 | 0.016063 |
| param_min_samples_split | 0.000771 | 0.000345 | 2.238143 | 0.027598 |
| param_min_samples_leaf | 0.011355 | 0.000906 | 12.53932 | 1.06E-21 |
| param_max_features | 0.000355 | 0.002292 | 0.155045 | 0.877122 |

| | | | | |
|---|---|---|---|---|
| param_max_depth | -6.1E-05 | 3.22E-05 | -1.88176 | 0.062996 |
| param_bootstrap | 0.03596 | 0.002249 | 15.98767 | 1.99E-28 |

To better visualize this data, violin plots of each variable and their effects on the test result were generated. A higher test score – the y-axis – is a better result. The best model for hyperparameterization has 1400 estimators, so it is surprising that the graph demonstrates that, on average, this selection has the most suboptimal results. However, this was likely chosen due to the fact that this selection has the highest variance in test scores, and its low average can be explained by the fact that there is much more room to have a lower score than a higher one. For the minimum number of samples to split a branch, there is little variation in the average result. Again however, the selected value – 10 – is due to this value's higher variance, producing much higher and lower test scores than the other options. The minimum samples for a leaf in a decision tree differ greatly from the first model. Despite the highest value having a much higher average test score than the other options, the middle value is selected due to its longer tail. In the first model, all three options had a much higher variance than this model. Furthermore, as its p-value suggests, the maximum number of features makes little difference, with both distributions appearing nearly identical.

Though not statistically significant, the maximum depth parameter merits some discussion. Each value is radically different from each other, more so than in the first model. On average, a value of 20 seems to perform better, but a value of 90 has higher variance and may lead to better outcomes if enough fits are performed. Finally, bootstrap values have the greatest difference in mean test scores of any parameter, with TRUE outperforming FALSE by five-hundredths. As expected, the former value was chosen during the fit.

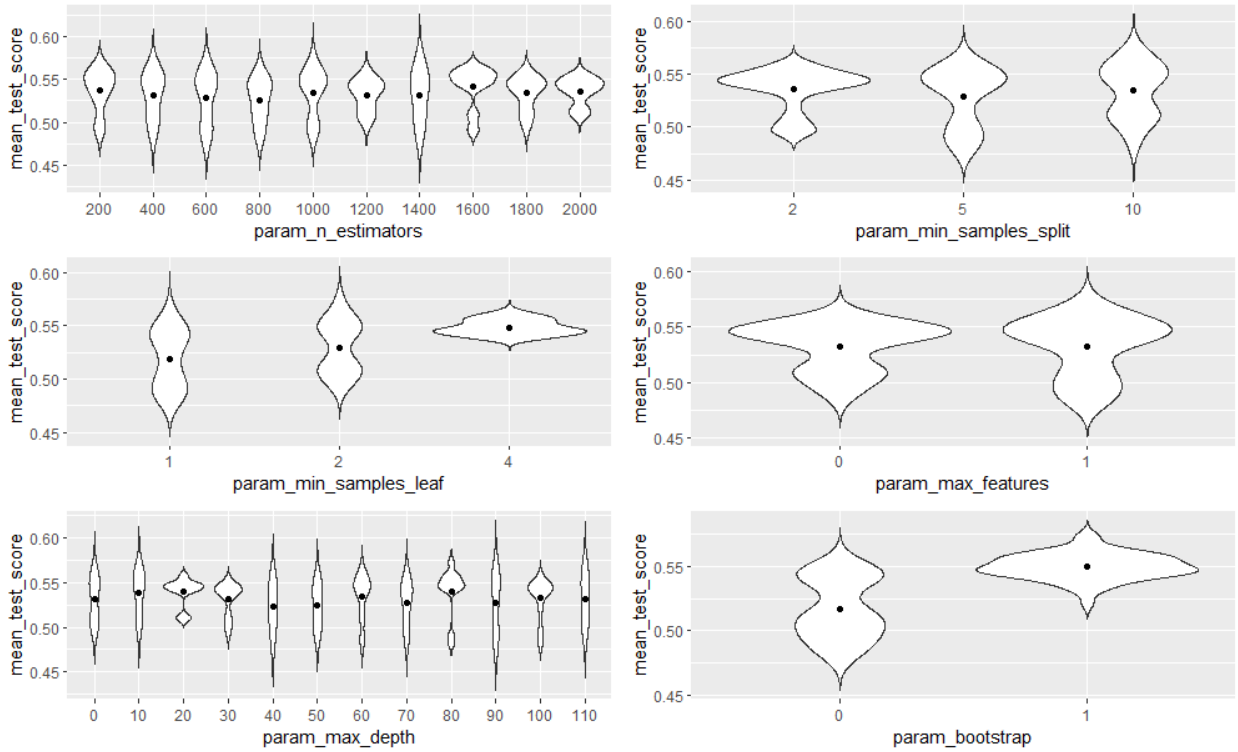



Figure 8. Visualization of Parameter Variation Data for Random Forest Classifier

## 5.1. Evaluation

Evaluating the model with a different 80/20 train/test split so that a higher number of testing results may be analyzed, a model with an accuracy of 53.0% was obtained, lower than the prior model. The following parameters were obtained, representing the model with the best performance from hyperparameterization.

Table 4. Random Forest Classifier Parameters

| # Estimators | Min Samples Split | Min Samples Leaf | Max Features | Max Depth | Bootstrap |
|---|---|---|---|---|---|
| **1400** | 10 | 2 | sqrt | 80 | TRUE |

Though the accuracy statistic is lower than in the prior model, the false positive rate was successfully reduced, correctly diagnosing 15 AD patients and identifying 29 patients without AD. However, the number of false positives and negatives were 19 and 20 respectively. This is still higher than optimal. As a comparison, the false positive rate of mammograms over *ten* years is approximately 50% (Pace, 2022). The false negative rate of mammograms is about 20% (National Cancer Institute, 2023).

For this model, an F1 score was generated. This score is a function of accuracy and recall, recall being one minus the false negative rate. This value was 59.3%, indicating that the model does perform better than chance (with 100% being perfect diagnostic quality). While these results are an improvement on the previous model, improvements can be made.

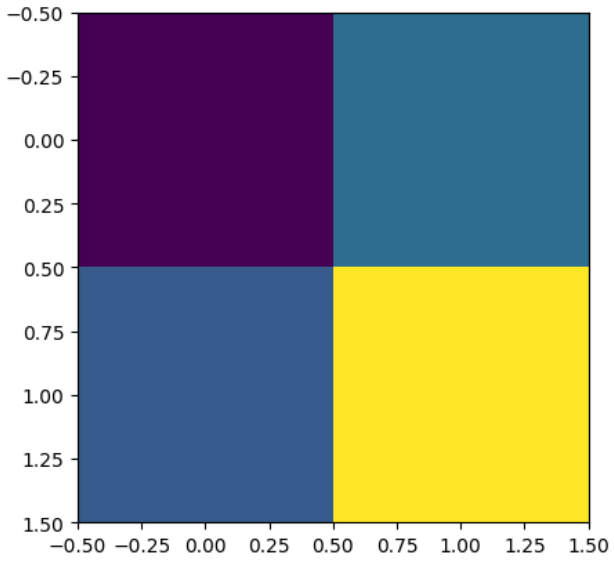



Figure 9. Random Forest Classifier Confusion Matrix

## 6. Boosting the Model

XGBoost starts as a typical random forest model, building an ensemble of decision trees and creating the forest. The difference between XGBoost and a random forest is that the former's trees depend on the results of the trees before it, and the random forest is the average of many trees. This results in a more accurate model (xgboost developers, 2022). Hyperparameterization was again performed, this time with a smaller number of permutations due to less available parameters. The number of estimators varied between 100 and 1400; the maximum depth between 10 and 110, and maximum leaves between 2, 5, and 7 with an unlimited option. Given the limited amount of data available, cross-validation was used to train and test the model. This splits the data into a number of data subsets, all but one of which train individual models, the last one tested against. The average performance of the models generated is the accuracy metric. This does tend to generate worse accuracy results for smaller datasets such as the ADNI data than expected. However, this does result in better models. The hyperparameterized model's accuracy statistic is 52.6%. The parameters of this model are 200 estimators, a maximum depth of 30, and unlimited leaves.

To better visualize the results of this data, a confusion matrix (using cross-validation) was generated. It properly diagnosed 116 AD patients, with 58 false positives and 31 false negatives. It identified 208 patients correctly as not having AD. This improves greatly over the prior two models.

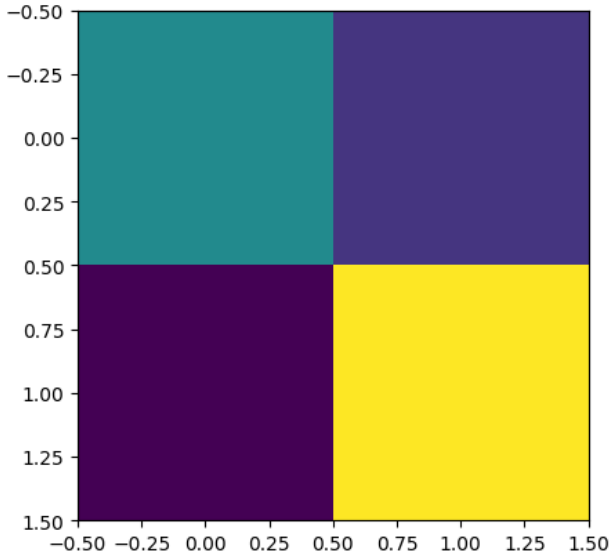



Figure 10. XGBoost Cross-validated Confusion Matrix

## 7. Model Comparison

The random forest classifier method was chosen because of its ability to demonstrate its decision structure in a human-readable format for diagnosis of AD patients. First, the decision surface of the random forest classifier was generated, revealing a noncontinuous relationship between homocysteine and isoprostane levels and the diagnosis of AD. Generally, given the relationship in the literature, one would expect that as homocysteine and isoprostane levels increase, there is a likelier chance of AD. Indeed, at very low levels of both of these biomarkers, this holds true. However, discontinuous islands of AD appear at higher levels, but the highest levels of the biomarkers do not reveal a connection between those and AD. There are several narrow bands with negative diagnoses combined with larger islands. This model is suboptimal due to the presence of these islands – it suggests overfitting of the data. This means that the model is too attuned to the training data and may not generalize to the real world; it is able to do so by recognizing patterns in the noise of the patient data (Carremans, 2018). Though there is not

another population that may be tested on at the present moment, we can assume that it may not work as well in the real world. Given the second model's performance, this is not good.

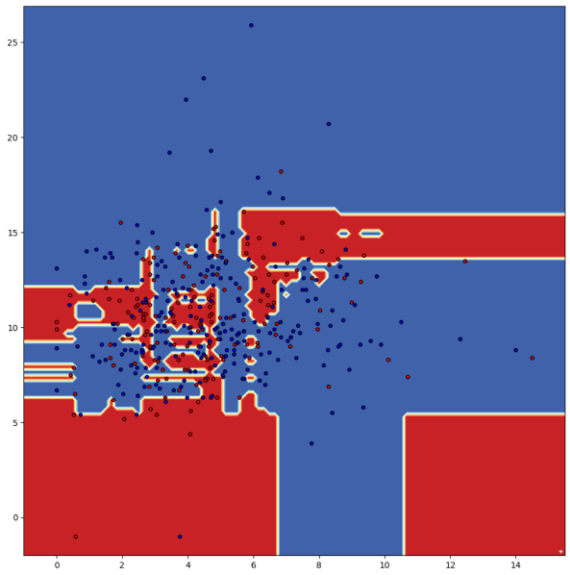



Figure 11. Decision Surface, Random Forest Classifier

The general pattern of island diagnoses holds for the XGBoost model. However, the same islands do not appear. The main difference is that the model more freely diagnoses AD at a lower incidence of isoprostane than the other. Also, there are less sparse diagnostic islands than in the prior model, which makes it more generalizable for aiding medical diagnoses. Though there is more than what may be optimal, the lower number reveals that there may be less overfitting in the XGBoost Model which allows it to be used for diagnostic purposes. However, it would benefit from more training data. Ideally, there would be no islands of diagnoses and while we do not expect to see a linear relationship, there ought to be less gaps in the data.

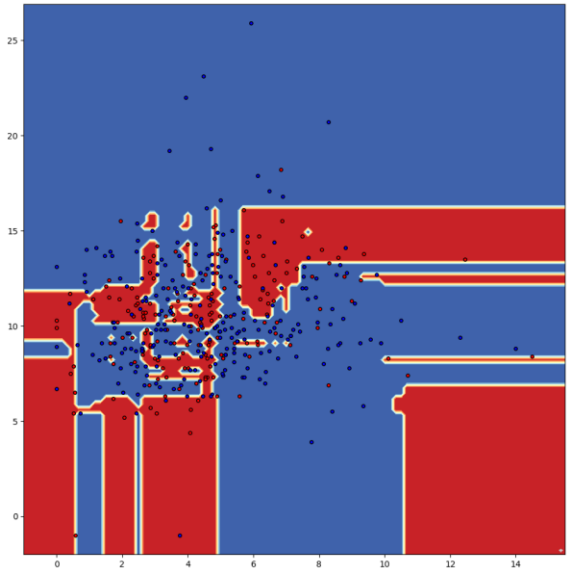



Figure 12. Decision Surface, XGBoost

XGBoost works by chaining various decision trees together to produce a forest. Though not the best tool for diagnosis, this allows the generation of the tree which outputs the final decision and has been generated below where f1 represents isoprostane and f0 homocysteine. This iterated series of trees gives a few conclusions. First, while there is a positive association between the incidence of AD and elevated isoprostane and homocysteine levels, it is not necessarily linear. Therefore, as shown in past studies, linear models are insufficient for this task. Also, homocysteine is the largest contributor to the diagnosis of AD out of the two biomarkers. This is not a new finding (Morris, 2003). However, that this tree independently came to this conclusion demonstrates some diagnostic validity. And, this finding points to some treatment areas if this is a causative factor, namely B vitamin supplementation and methylated B vitamins if the patient has an MTFHR mutation that makes them unable to process cyanocobalamin and the like (Maron & Loscalzo, 2009). Given more data, this tree could likely be collapsed into a doctor-readable decision tree to be used in the diagnosis of AD without a computer.

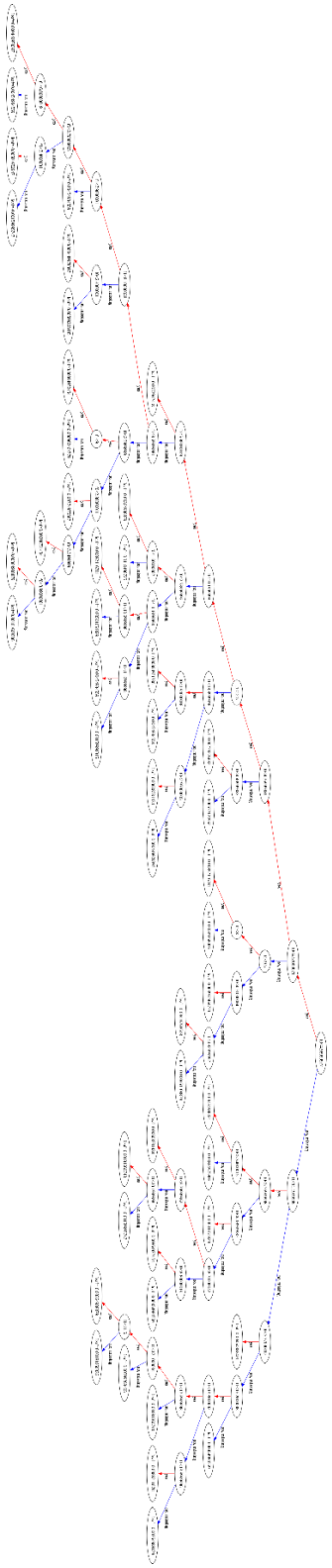


Figure 13. Decision Tree, XGBoost

## 8. Future Progress

The data collected for this thesis represents a very small subset of AD patients. Most machine learning models benefit from much more data than that used in this paper. Given that homocysteine and isoprostane levels are easily collected from blood tests, a study could be created to take these levels and generate a more generalizable model for the diagnosis of AD. Furthermore, the ADNI data did not necessarily have patients which represent the general population; all ADNI patients are eligible by virtue of being between the ages of 55 and 90. The levels of homocysteine and isoprostane for younger people are likely different and therefore this model may not work for those outside these demographics. Furthermore, these patients are already likely to have elevated levels of these two biomarkers, making it more difficult to diagnose them when only looking at patients with these prior risk factors. A researcher with better access to patient data across the United States may be able to consider age in their model and look at demographics not covered in this study. A promising avenue of research would be measuring these levels as the patient grows older, looking for spikes in these biomarkers which precede an increase in ADAS scores to determine a cause-and-effect relationship, whether elevated isoprostane and homocysteine cause AD or these increased levels are a biproduct of the disease. A personal conjecture is that hyperhomocystemia and AD are both diseases caused by an underlying factor and that hyperhomocystemia happens to be more visible first since it is more easily tested for. However, higher isoprostane levels are likely evidence of already-existing neurodegeneration. However, these are simply conjectures and are very interesting avenues for future research.

As well, machine learning as we know it is still in its infancy. The first random forest classifier used in this model was developed in 1999; XGBoost in 2014 (United States Patent No. US6009199A, 1999) (Chen & Guestrin, 2016). And, these classification techniques have only just had their first forays into medicine given their newness and time it takes for new technologies to diffuse into other fields (Alam, Rahman, & Rahman, 2019). Given the recent rate of advancement in machine learning (e.g., ChatGPT and other LLMs), the gap between technology and that used in the medical field may only widen.

This gap is not monocausal. While the rate of advancement in the machine learning field versus medicine is large, there are also less incentives for advancement in medicine specifically. Novelty is incentivized over working off older research; academic code does not often follow the same standards as code written in industry. As well, the job market for AI engineers and the like is much higher than the pay afforded to medical professionals (Leming, et al., 2023).

There are additional challenges aside from the positive incentives, namely the likelihood of future regulation of machine learning. Increased scrutiny has been applied to AI not only due to its exponential rate of increase but also due to its new applications which augment human judgement, and nowhere is the quality for human judgement more valued than that of the medical profession (Candelon, di Carlo, De Bondt, & Evgeniou, 2021). Furthermore, medical associations are worried about racial bias in diagnostics given the recent scandal in artificial upward adjustment of Black patient's kidney filtration scores, denying some patients necessary transplants (Robeznieks, 2021). This was noticed in a relatively uncomplicated algorithm which directly mentioned race (Inker, et al., 2021). But in a more complicated algorithm in which race is not present, race may be correlated with both biomarker levels and outcomes may hold racial

bias in an unexpected way. Thus, great care must be taken to ensure the absence of bias. There are very recently developed methods to explain behavior in LLMs – this will go a long way in reducing bias in larger models such as IBM's Watson diagnostic tool (Bills, et al., 2023). However, for models used in this paper such as random forest classifiers and XGBoost, racial bias must be elucidated through statistical methods. To solve this problem, careful processing of the data may help – preprocessing seems to significantly reduce racial bias in one iteration of XGBoost – as well as better model selection (Allen, et al., 2020).

## Conclusion

Alzheimer's disease has an extremely complex etiology which has led to its increased diagnosis as lifespans increase. Given its eventuality in the ageing process, much more research into its causes and prevention must be incentivized. Of the biomarkers involved in the progression of the disease, homocysteine and isoprostane are of interest, seeming to act as a marker of the underlying causes of AD for the former and a sign of neuronal death and further degeneration for the latter. Several models analyzing these biomarker levels and their effect on ADAS scores were analyzed, with random forest regression on raw ADAS scores falling short, but higher accuracy was achieved by converting this to a diagnosis model with a ADAS score greater than 17 as a cutoff. Furthermore, hyperparameterization was proven to help with diagnostic outcomes. To further improve this model, XGBoost with chained decision trees was implemented, reducing the proportion of false negatives and positives in the ADNI population. Despite this improvement, far more progress can be made in the diagnosis of AD if larger datasets of AD patients are available, as a sample size of <1000 is not optimal for the creation of a general model of diagnosis for the disease. To these ends, it would be helpful for more patient data to be open-sourced. This would allow for better models with more researchers working on them. Also, LLMs present a promising compliment to these models. Given the recent success of LLMs such as ChatGPT 3 and 4, one could envision an LLM trained on patient charts and histories covering several years such that a general diagnostic intelligence may be trained from this set. The proliferation of diagnostic models will help doctors make more informed decisions for patients, help properly diagnose patients, and assist in research on diseases such that false positives are excluded from study populations, allowing for more targeted treatments.

# References

Alam, Z., Rahman, S., & Rahman, S. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*.

Allen, A., Mataraso, S., Siefkas, A., Burduck, H., Braden, G., Dellinger, R. P., . . . Das, R. (2020). A Racially Unbiased, Machine Learning Approach to Prediction of Mortality: Algorithm Development Study. *JMIR Public Health and Surveillance*.

Alzheimer's Disease Neuroimaging Initiative. (2017). *About*. Retrieved from ADNI: https://adni.loni.usc.edu/about/

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., . . . Sunders, W. (2023). *Language models can explain neurons in language models.* OpenAI.

Blennow, K., de Leon, M., & Zetterberg, H. (2006). Alzheimer's disease. *Lancet*, 387-403.

Candelon, F., di Carlo, R. C., De Bondt, M., & Evgeniou, T. (2021, September). AI Regulation Is Coming. *Harvard Business Review*.

Carremans, B. (2018, August 23). *Handling overfitting in deep learning models*. Retrieved from Towards Data Science: https://towardsdatascience.com/handling-overfitting-in-deep-learning-models-c760ee047c6e

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv*.

Gunes, S., Aizawa, Y., Sugashi, T., Sugimoto, M., & Rodrigues, P. (2022). Biomarkers for Alzheimer's Disease in the Current State: A Narrative Review . *International Journal of Molecular Sciences*, 4962.

Ho, T. K. (1999). *United States Patent No. US6009199A.*

Inker, L. A., Eneanya, N. D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., . . . Guti\'{e}rrez, O. M. (2021). New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. *The New England Journal of Medicine*, 1737-1749.

Irizarry, M., Yao, Y., Hyman, J., Growdon, J., & Practico, D. (2007). Plasma F2A Isoprostane Levels in Alzheimer's and Parkinson's Disease. *Neurodegenerative Diseases*, 403-405.

Janssen, L. (2001). Isoprostanes: an overview and putative roles in pulmonary pathophysiology. *American Journal of Physiology-Lung Cellular and Molecular Physiolog*, 1067-1082.

Kawarabayashi, T., Younkin, L. H., Saido, T. C., Shoji, M., Ashe, K. H., & Younkin, S. G. (2001). Age-Dependent Changes in Brain, CSF, and Plasma Amyloid β Protein in the Tg2576 Transgenic Mouse Model of Alzheimer's Disease. *Journal of Neuroscience*, 372-381.

Koehrsen, W. (2018, January 9). *Hyperparameter Tuning the Random Forest in Python*. Retrieved from Toward Datascience: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Leming, M. J., Bron, E. E., Bruffaerts, R., Ou, Y., Iglesias, J. E., Gollub, R. L., & Im, H. (2023). Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. *npj Digital Medicine*.

Maron, B. A., & Loscalzo, J. (2009). The Treatment of Hyperhomocysteinemia. *Annual Review of Medicine*, 39-54.

Martí-Carvajal, A., Solà, I., Lathyris, D., & Dayer, M. (2017). Homocysteine-lowering interventions for preventing cardiovascular events. *Cochrane Database of Systematic Reviews*.

Monllau, A., Pena-Casanova, J., Blesa, R., Aguilar, M., Bohm, P., Sol, J., & Hernandez, G. (2007). Diagnostic value and functional correlations of the ADAS-Cog scale in Alzheimer's disease: data on NORMACODEM project. *Neurologia*, 493-501.

Morris, M. S. (2003). Homocysteine and Alzheimer's disease. *The Lancet Neurology*, 425-428.

National Cancer Institute. (2023, February 21). *Mammograms*. Retrieved from cancer.gov: https://www.cancer.gov/types/breast/mammograms-fact-sheet

National Health Service. (2021, July 5). *Alzheimer's disease - Treatment*. Retrieved from NHS.uk: https://www.nhs.uk/conditions/alzheimers-disease/treatment/

National Institutes of Health. (2022, December 8). *How Is Alzheimer's Disease Diagnosed?* . Retrieved from National Institute on Aging: https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed

Pace, L. E. (2022). False-Positive Results of Mammography Screening in the Era of Digital Breast Tomosynthesis. *JAMA Network Open*.

Robeznieks, A. (2021, June 23). *Feds warned that algorithms can introduce bias to clinical decisions*. Retrieved from American Medical Association: https://www.ama-assn.org/delivering-care/health-equity/feds-warned-algorithms-can-introduce-bias-clinical-decisions

Rocca, W. P., & White, L. (2011). Trends in the incidence and prevalence of Alzheimer's disease, dementia, and cognitive impairment in the United States. *Alzheimer's & Dementia*, 80-93.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy Analysis of State Sequences. *Sociological Methods & Research*, 471–510.

Trares, K., Chen, L.-J., & Schöttker, B. (2022). Association of F2-isoprostane levels with Alzheimer's disease in observational studies: A systematic review and meta-analysis. *Ageing Research Reviews*.

xgboost developers. (2022). *Introduction to Boosted Trees*. Retrieved from xgboost: https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Zhuo, J.-M., Wang, H., & Pratico, D. (2011). Is hyperhomocysteinemia an Alzheimer's disease (AD) risk factor, an AD marker, or neither? *Trends in Pharmacological Sciences*, 562-571.

# Biography

Hazel G. Helfman was born in Dallas, Texas, electing to attend the University of Texas at Austin for college, and majoring in both Chemical Engineering and Plan II. During her time there, she worked in undergraduate research on computer-aided analysis of amyloid-beta proteins, was the lead graphic designer for the Texas Triple Helix, worked in theater tech, tutored children, was the scribe, then regent, of Theta Tau, worked for the Texas Commission on Environmental Quality, took a semester off school to work at Bayer's Iowa plant, then again worked there the next summer, and somewhere in this time came into existence.