Ensemble Methods for Prediction of Longitudinal Risk and Development of Alzheimer's Disease

Daniel Wu (G5)

CSCI5525: Machine Learning
Paul Schrater
wuxx1495 (5214001)

Abstract

Effective and accurate diagnosis of Alzheimer's disease (AD) as well as the prodromal stage (mild cognitive impairment (MCI)). So far multiple papers present a brain T1-weighted structural magnetic resonance imaging (MRI) biomarker that combines several MRI biomarkers to attempt predictive diagnosis. however, most existing research focuses on only a single modality of biomarkers for diagnosis of AD and MCI. Through the Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge, we attempt to combine several imaging clinical modalities along with genetic and epidemiological data to create a predictive model. We will use biomarkers from MRI scans, functional imaging (FDG-PET) for hypometabolism, CSF biomarkers. Two predictive models were used, a ensemble method for a neural network using bagging, and an ensemble method using gradient boosted decision trees (XGBoost) that combines several weak classifiers. These methods was developed, trained, and evaluated using a standardized dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Using the multi-class area under the receiver operator curve (mAUC) and the balanced classification accuracy (BCA) as performance metrics, the stacked XGBoost method had an mAUC of .593 . This proved to be worse than the base classifiers used to train the stacked ensemble method, the best being the Random Forest classifier, ending up with a mAUC of .692. In the end, the bagged neural network produced a better result, with an mAUC of .645. Overall, we were unable to develop a satisfactory method with the features selected from the data provided. Further analysis of the data and feature selection is necessary.

1 Introduction

Alzheimer's Disease is a disease that remains a key challenge for the 21st century healthcare. The statistics show that by age 80 20% of the population will suffer from dementia, which AD is the most common cause. Dementia has a higher health and social cost tan cancer, stroke, and chronic heart disease combined. The costs are projected to be 1 trillion in 2018 and doubled by 2030.

No current treatments provably cure or even slow AD. Of the several clinical trials that have been approved or at the approval stage, and none of them have managed to show or prove any disease-modifying effect. A proposed reason for the lack of effectiveness of these treatments is that it is difficult to identify patients at early stages of the disease where treatments are most likely to be effective.

The goal of this project was to develop a predictive model that helps with computer-aided diagnosis methods based on clinical data gathered in. There have been a variety of methods that have shown

promising results in literature based on MRI-based diagnostic classification that have been presented in literature, but frequently the methods are optimized for specific data sets. It is unclear how the algorithms would perform on unseen data, which would be a better predictor of its value in a clinical setting. This is the fundamental goal of the predictive models based on current clinical data. Models that can predict the diagnosis and potential risk one has in progressing to dementia is vital to developing better techniques and treatment that can affect the disease beyond management of symptoms. Once dementia can be definitely clinically diagnosed, brain degeneration has already occurred and irreplaceable brain tissue limits the ability of treatments to be able to make a positive impact in potentially halting or reversing the dementia.

However currently, the limited understanding of AD makes prediction of symptoms onset hard. However, several approaches have been effective in scientific literature, and some methods will be described below

1.1 Manual Prediction

An informed clinician experienced in interpreting multi-modal data can judge prognosis and predict conversion to a different diagnostic category by drawing on their own knowledge of clinical history of specific patients and their general knowledge and experience with AD.

1.2 Statistical Learning

Regression has been used as a statistical technique to model the relationship between age and disease risk using several biomarkers, such as anatomical volumes from MRI, cognitive test scores, cognitive decline, and other biomarkers. Machine learning methods, such as support vector machines, random forests, and artificial neural networks have been used to attempt to learn the relationship between values of a set of predictors and their labels. They can prove very effective in high dimensional classification and regression problems. In this project, we will attempt to improve on these methods for a more generalized population with several more patients.

1.3 TADPOLE Challenge

The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) Challenge used data collected from primarily the ADNI collection of open source neuroimaging data, which contains a variety of patients with AD, healthy controls, from around the nation. The validation set is a subset of the data available from the datasets that have longitudinal data that can be used for prediction.

In current literature, we can categorize patients into three classes with Alzheimer's disease (AD), patients with mild cognitive impairment (MCI), and cognitively normal individuals (CN). These are diagnosis criteria that have been developed which has been common practice in the studies of computer-aided diagnosis methods [1]. Confirmation of diagnosis of Alzheimer's disease is challenging as the only way to develop a ground truth diagnosis of dementia is with an autopsy. Therefore the importance of accurately predicting these clinical diagnoses may help in being able to target clinical research and treatment to patients based on clinical data prior to them developing AD.

This project aims to use data provided by the challenge to predict the AD classification using a wide variety of imaging modalities and features. We will compare the results with several ensemble classifiers, support vector machines, and neural networks. Two other meta-algorithms will also be implemented: a stacking ensemble method of 5 machine learning base classifiers using XGBoost, and a bagged neural network model.

2 Data

The dataset includes a comprehensive longitudinal data set for training and a list of some longitudinal data on rollover subjects for forecasting. Because of the nature our models, we will discard longitudinal data and instead treat this as a cross-sectional study using current data to predict future outcomes.

Several quantitative biomarkers, which are medical measurements that can indicate a disease, are available from the ADNI dataset. The biomarkers can be roughly divided into two categories:

measures of the amyloid beta protein and measures of damage to nerve cells. Amyloid beta protein can be measured either using cerebrospinal fluid (CSF) or amyloid position emission tomography (PET). For the second category, a wide variety of methods are available, including CSF, tau-PET, quantifying brain metabolism using fluoro-deoxyglucose (FDG) PET or atrophy using MRI.

In the ADNI Data set, there are over 2000 features. However, because of the nature of the data acquisitions being done in several clinical sites around the country with no set standard in how the data is organized, some features are relatively sparse. For the purpose of this project in investigating the predictive power of the dataset, most features were eliminated due to the sparsity of the data, and only 22 key features remained to be incorporated in the model. Justification for the specific features selected is discussed here and preprocessing techniques will be discussed in the methods section.

In addition to biomarkers, the models also incorporate risk factors, such as age, sex, genetic risk factors as well. APOE is the gene that is primarily investigated as a risk factor for developing AD.

1. Main cognitive tests - neuropsychological tests administered by a clinical expert
    (a) CDR Sum of Boxes
    (b) ADAS13
    (c) MMSE
    (d) RAVLT
    (e) MOCA
    (f) ECOG
2. MRI Regions of Interests (ROIs) (Freesurfer) - measures of brain structural volume and integrity
3. FDG PET ROI averages - measures cell hypometabolism where cells affected by AD show reduced signal.
4. AV45 PET ROI averages - measures amyloid-beta load in the brain, where amyloid-beta is a protein that misfold, increase the risk of developing AD or related dementia
5. CSF Biomarkers - amyloid and tau levels in the cerebrospinal fluid (CSF), as opposed to the cerebral cortex.
6. Others
    (a) APOE status - a gene that is a risk factor for developing AD
    (b) Demographic information: age, gender

2.1 MRI measures

MRI is a technique used to image the anatomy and the physiological processes of the brain. In the ADNI dataset, quantification of ROIs were done using a standardized protocol. For this particular model, we choice to use whole volume measurements with these ROIs rather than cortical thickness, with the hope that individual features in the model be as uncorrelated as possible. In the model, the volume measurements of the Ventricles, Hippocampus, whole brain, entorhinal region, Mid temporal region, and fusiform region.

2.2 Cognitive Tests

Cognitive tests are neuropsychological tests administered by a clinical expert which assess several skills like general cognition, memory, language, vision, etc. These are cognitive tests that give an overall sense of whether a person is aware of their symptoms, and general motor and cognitive functions. These are important in the context of AD because they help quantify the cognitive decline as disease progresses. However, these tests suffer from bias introduced when patients end up memorizing the answers to the test after taking it multiple, and also exhibit floor effects, as many people reside in the extremes and not much in between.

2.3 PET measures

Positron Emission Tomography detects pairs of gamma rays emitted by a radioactive tracer, which is introduced into the body of a biologically active molecule. Three-dimensional images of tracer

concentrations within the body then produce an image through computational analysis. The patient is injected with contrast agent which helps visualization of areas with low glucose intake (i.e. hypometabolism) which may indicate areas of atrophy and degeneration in the brain. FDG-PET give information of the neuronal cell metabolism that refers to activity going on inside neuronal cells.

2.4 CSF measures

The cerebrospinal fluid is a clear, body fluid found in the brain and spinal cord that acts as a cushion or buffer for the brain. Measures of CSF are very important for dementia research, and the concentration of these fluids are a strong indicator of AD. However, the lumbar puncture procedure is quite invasive and often not done in a routine clinical exam.

3 Methods

3.1 Data Preprocessing

The main difficulty of this problem was the sparsity of the data. Because of the nature of the open source project of the current ADNI training set, there are several features are missing because of the nature of these clinical scans, and the varying differences in protocol of how data is acquired from site to site. The challenge coordinators were able to create a centralized list of features, but of only 27 features. Out of those features, some were cut because of sparsity or too highly correlated with other features in the training data set that would not yield additional helpful information. Because many of the libraries used for the models used do not handle missing data, multiple imputation by chained equations (MICE) [3] was used as the principled method of dealing with missing data. This method as a method for addressing the missing data by creating multiple imputations that account for statistical uncertainty in the imputations. This chained approach can handle both continuous and binary as well as complexities such as bounds and survey skip patterns.

In addition to data imputation, a correction factor was introduced for the MRI volume ROIs. Because of the need to account for the difference in people's head sizes and brain sizes, a normalization factor using the entire intracranial volume (ICV) is introduced to normalize the volume ROIs. The baseline intracranial volume is used to account for potential decrease in intracranial volume as AD disease progresses.The normalization factor is, as follows:

$$Corrected\ ROI = \frac{Old\ ROIs}{ICV}$$

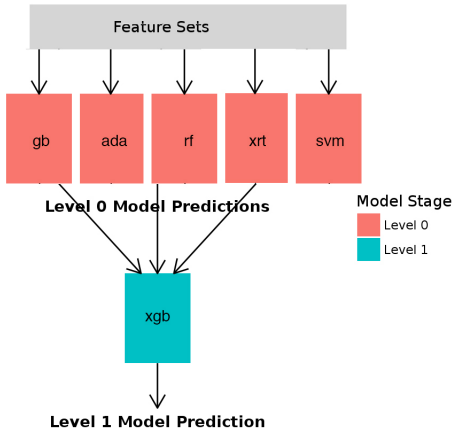

Figure 1: Model of Ensemble Method.

3.2 Stacked Ensemble Method using XGBoost

Ensemble modeling is now well-established to boost predictive accuracy by combining the predictions of multiple machine learning models. Model stacking is an efficient ensemble method in which the predictions that are generated by using different learning algorithms are used as inputs in a second-level learning algorithm. Model stacking has been successfully used on a wide variety of predictive modeling problems to boost the models prediction accuracy beyond the level obtained by any of the individual models. It has been suggested that the diversity of models in a library plays an important role in building a better ensemble model. Dietterich (2000) emphasized the importance of diversity by stating that the ensemble models gain more accuracy and robustness despite potential concern of overfitting and leakage of training.

In order to build our library of diverse models, we will use many different machine learning algorithms with hyperparameter tuning for Level 0.. We will combine five different machine learning algorithms, gradient boosting, AdaBoost, random forest, extremely randomized trees, and support vector machine algorithms. We will briefly described each algorithm below.

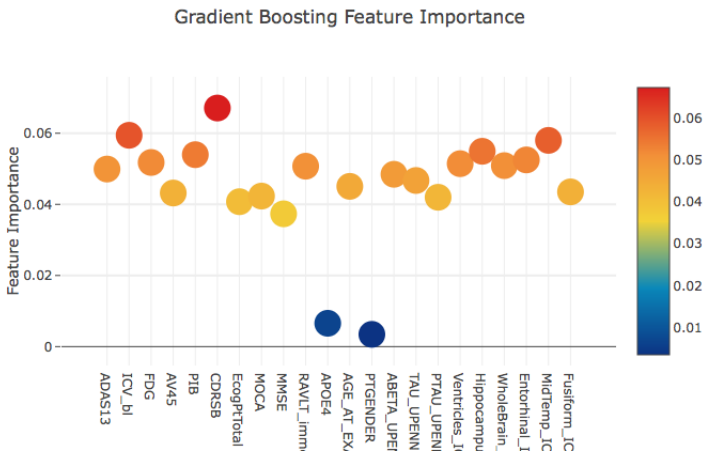

Figure 2: Feature importance graph for Gradient Boost.

Gradient Boost and AdaBoost

Boosting is a class of machine learning methods based on the idea that a combination of weak learners is capable of becoming a strong learner.[6] The idea of Gradient Boost is to use the same weak learning algorithm for repeated set of learners. Each step, weak classifiers compare their classification labels, and the portions in which they disagree becomes the boundaries in which the next classifier learns. For the purpose of classifying images we choose soft-max objective as the loss function for Gradient Boost. Adaboost utilizes a Decision stump, which are weighted versions of the same training data rather than subsamples of the data. Thus each weights update each iteration and allows the weak learner to focus on patterns that were not already classifier. [6] In order to select the best hyper-parameters, grid search cross-validation is implemented. For the grid search cross-validation, we consider 2 parameters, the number of estimators and the learning rate. After careful selection, the parameters chosen for our Adaboost model are 1000 estimators and a 0.75 learning rate. The feature importance graph of our Adaboost model can be seen in Figure 3. For the Gradient Boosting method, an additional parameter of the depth of the tree can be changed. We tuned that parameter to a max depth of 5. The feature importance of our Gradient Boost model can be seen in Figure 2.

Random Forest and Extremely Random Trees

Random forests is a classifier that contains trees in an ensemble that are built from a sample drawn with replacement of the training set. The split that is picked is the best split among a random subset of features. As a result of this randomness, the bias of the forest usually increases but because of the averaging of the algorithm, the variance also decreases, usually an acceptable compromise in the real world. Extremely randomized trees replicate this behavior, but thresholds are drawn at random

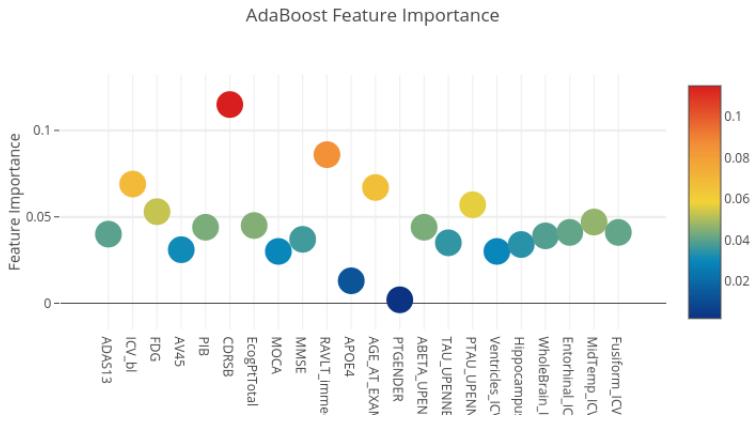

Figure 3: Feature importance graph for Adaboost.

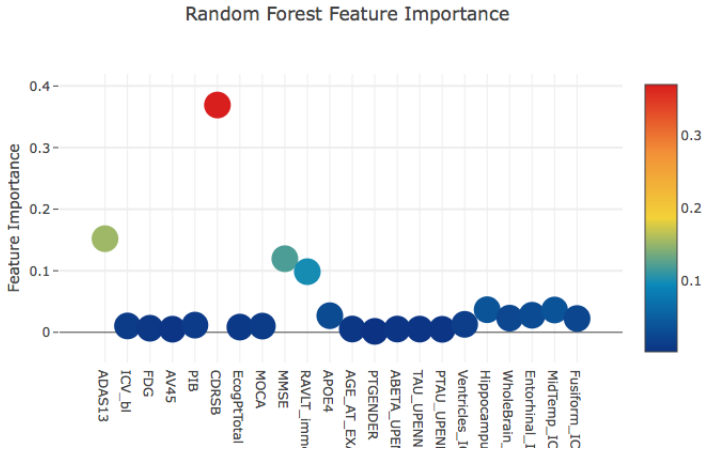

Figure 4: Feature importance graph for Random Forest.

for each candidate feature and the beast of the thresholds is picked as the splitting rule. This further reduces variance at the expense of increased bias.

Similar to what we did when choosing hyper-parameters in Gradient Boost, grid search cross-validation is used here to select parameters. For the grid search cross-validation, we consider 3 parameters, the number of trees, the max-depth of each tree, the proportion of features to be used in building each individual tree. The parameters we used for the Random Forest classifiers are: 1000 trees and a max depth of 6. The model's feature importance can be seen in Figure 4. For the Extreme Randomized Trees Classifier, we ended up using 1000 trees and a max depth of 8 per classifier. The model's feature importance can be seen in Figure 5.

Support Vector Machines

The support vector machines implemented in sciki-learn support multiclass classification using the one-vs-the-rest approach, which to summarize, means that separate models are trained for each class. We decided to go with a linear kernel, and set $C = 0.025$ *Ensemble Method - Extreme Gradient Boost*

Extreme Gradient Boosting is an implementation of the boosted tree supervised model.[5] Boosted trees are essentially the same model as random forests, except the training of the model utilizes a loss function that can be optimized with a gradient boosting framework. The uniqueness of this model is the utilization of parallel tree boosting that helps solve data science problems in a fast and accurate way and is scalable. We can see in Figure 8 the features that are overall important in each model.

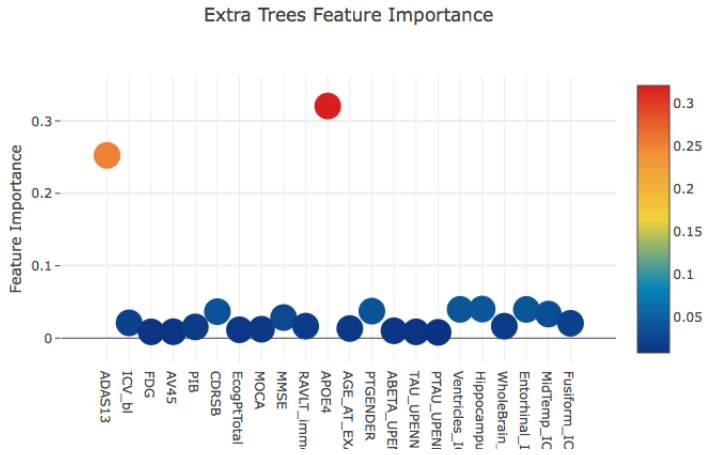

Figure 5: Feature importance graph for Extremely Randomized Trees.

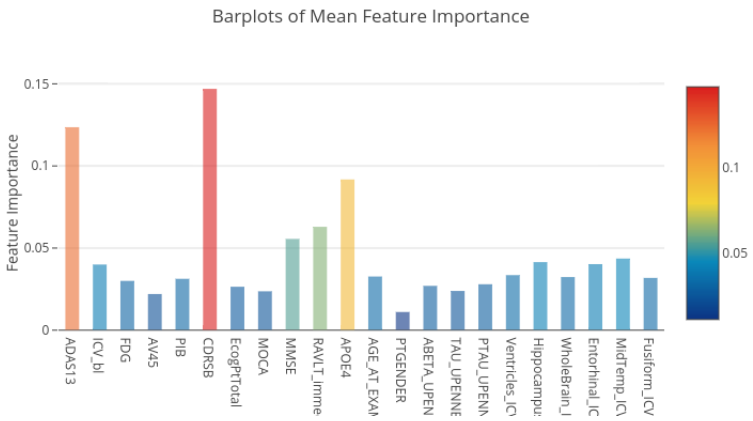

Figure 6: Feature importance graph averaged from all models.

3.3 Bagging Neural Network Model

The neural network models were trained as shown in Figure 6. The model essentially has two hidden layers, one with 500 neurons and the other with 200 neurons. Both layers use Parametric Rectified Linear Unit (PReLU) activation function. Two dropout layers are also incorporated after each hidden layer, set to a dropout rate of 0.2. Finally, the output layer is a softmax layer that classifies our result into three separate probabilities.

Bagging Bootstrapping is a general purpose sample-based statistical method in which several (non disjoint) training sets are obtained by drawing randomly, with replacement, from a single base dataset. Bagging, or bootstrap aggregation is a technique which uses bootstrap sampling to train multiple models and average each model to attempt to reduce variance or improve accuracy of the predictor. The main advantages of this method is to improve the classification result stability and accuracy while also reducing classifier variance.

4 Results

4.1 Feature Importance

Four of the 5 ensemble methods are shown in Figures 2-5 that indicate the feature importance of several models. When comparing these feature importances to the mAUC results, we can note some

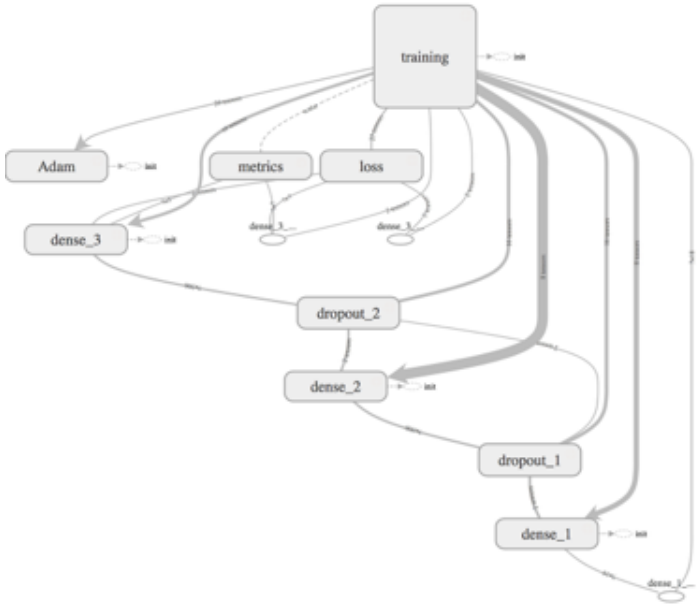

Figure 7: Neural Network Model.

Table 1: Performance Metrics from Ensemble Methods

Method	mAUC	BCA
AdaBoost	.576	.540
RF	.694	.535
GradBoost	628	.571
ET	.649	.543
XGB	.654	.535
XGB	.593	.571

observations. Firstly, the Random Forest and Extremely Randomized Tree models performed better than the GradBoost and Adaboost models, and also show less variance in feature importance. Both placed significant emphasis on non-imaging features, such as ADAS13 and CDRSB survey scores, as well as the genetic factor APOE4.

4.2 Performance Metrics

The receiver operating characteristic curve shows the trade-off between different classification outcomes. When dealing with True positives (TP), false positives (FP), True Negatives (TN), and False Negatives (FN), they can be an alternative to calculating test accuracy.

Multi-class area under the receiver operating curve (mAUC) was primarily used to evaluate the performance of our models.[2] A reliable mAUC estimate can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. The AUC is an overall measure of the ability to discriminate positive and negative cases. For multi-class problems, such as TADPOLE's clinical status prediction, the AUC $\hat{A}(c_i|c_j)$ for each class c_i can be calculated as

$$\hat{A}(c_i|c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j}$$

where S_i is the sum of the ranks of the class i test points. In the multiclass case, we take the average of each pair's AUC and calculate the overall mAUC

$$mAUC = \frac{2}{L(L-1)} \sum_{i=2}^{L} \sum_{j=1}^{i} \hat{A}(c_i|c_j)$$

Table 2: Performance Metrics from Neural Networks

Method	mAUC	BCA
RegNN	.625	.598
Bagging	.645	.606

In order to calculate BCA of each class, it is defined by this formula

$$BCA_i = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$$

The overall BCA is given by the mean of all the balanced accuracies for every class.

The result of the models are shown in Tables 1 and 2. The bagged neural network performed slightly better than the RegNN, and also yielded lower variance than the regular neural network. In terms of classification accuracy, it performed the best out of all the models available.

5 Discussion

The TADPOLE challenge proved to be a unique opportunity to be able to test training ensemble methods in a field in which there is increasing interest in developing computer aided diagnosis in diseases that have large amounts of data but relatively little fruit in terms of advancing treatment for those diseases. This dataset allowed the opportunity to attempt to predict models using a wide variety of data sources to show what particular clinical modalities may be worth focusing efforts in in future disease prediction models. The feature importance from ensemble classifiers indicate a number of things. First, APOE4 gene again is shown to be a more important features that aligns with the biological consideration of that gene's risk factors for people with AD. Second, the clinical surveys that evaluate cognitive function were weighted more importantly by our models than imaging modalities and more quantifiable measurements. The reason for this can be many, one of which is the simple explanation that the other quantifiable data still is widely variable and sparse depending on where it is gathered and who it is gathered by. The image processing and analysis was already completed for this project, but it would be interesting to be able to develop potential Convolutional Neural Networks that are trained on the raw data of the images themselves than pixel values from ROIs currently used now. Those may yield better results.

Another interpretation of these results indicate that perhaps these survey results are biasing our models, and because of the well-studied problems of survey results, they may perhaps be better excluded the model in its entirety.

The framework of the challenge required that our models yield forecasted predictions taking into consideration the time frame from the baseline scan that we would have been given data on. Time lapsed between scans would be a significant variable that is unaccounted for.

5.1 Further Work

Recent studies have also shown that AUC is quite noisy as a classification measure and has significant problems as a model comparison. In reality AUC is a tradeoff between performance of sensitivity and specificity. Investigating a better performance measurement maybe helpful in understanding the performance of our model.[4]

Because the ensemble methods of combining various machine learning algorithms yield unsatisfactory results, more work is required in being able to evaluate what went wrong. There are several points in the data processing pipeline in which we need to re-evaluate the decisions and motivations behind those decisions, or also consider revising our methods.

One of the possible potential changes to investigate is obviously the feature selection. Most of the feature selection was guided by medical knowledge of what has been helpful in guiding clinicians to a reasonable method in combination with complete data being at a premium in this open source imaging data bank by ADNI. It might be worth considering developing separate models for each clinical site in which there is a better chance that within each training set it is less likely that the

majority of the patients will be missing data. Then we can combine the models and perform an ensemble method to train our data. This method has not been explored but it might prove to be fruitful given that these models will most likely be relatively weak classifiers since they are trained only on a subset of the data. Even though MICE was utilized to help impute missing data, this method will also alleviate the need for imputation and thus be able to create hopefully a more robust model that is able to achieve higher accuracy.

5.2 Conclusion

Overall machine learning methods show promise in being able to predict categorical Clinical Diagnosis. The complexity of our stacked ensemble method proved to be a mixed bag, improving our mAUC but decreasing our classification accuracy.

Furthermore, this work was designed to aid in research and treatment of AD. The value of these results is lessened to some degree when considering the amount of effort required to be able to conduct the clinical studies, imaging, and other data that the models trained on in this project require. A more reasonable clinical setting would be to consider perhaps one or two of these modalities for. However, this project may prove to be a valuable resource for researches to help in the diagnosis of AD and developing treatments.

References

[1] L. Sorensen, C. Igel, A. Pai, I. Balas, C. Anker, M. Lillholm and M. Nielsen, Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry, NeuroImage: Clinical, vol. 13, pp. 470-482, 2017.

[2] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner and E. Dougherty, "Small-sample precision of ROC-related estimates", Bioinformatics, vol. 26, no. 6, pp. 822-830, 2010.

[3] M. Azur, E. Stuart, C. Frangakis and P. Leaf, "Multiple imputation by chained equations: what is it and how does it work?", International Journal of Methods in Psychiatric Research, vol. 20, no. 1, pp. 40-49, 2011.

[4] Ferri, C. Hernandez-Orallo, J. and M. Salido. Volume under the roc surface for multi-class problems. In ECML 2003.

[5] Chen, T.; Guestrin, C. arXiv preprint arXiv:1603.02754 2016

[6] Ferreira, A. Figueiredo, M. T. Boosting algorithms: A review of methods theory and applications in Ensemble Machine Learning New York:Springer pp. 35-85 2012.