Depression Patient SPECT Brain Scan Analysis

Maime Guan, Lingge Li, Duy Ngo, Fulya Ozcan, Dustin Pluta, Yuxiao Wang

December 13, 2016

Abstract

We analyzed data from 954 depression patients with logistic regression, random forest and other methods to predict treatment outcome and achieved 56% accuracy. Several brain regions are found to be statistically significant along with compliance indicator. The same methods are able to separate depression patients from ADHD patients with AUC 0.72.

1 Introduction

Single photon emission computed technology (SPECT) is a recent neuroimaging technique that has been applied to identify and understand various brain related health issues. For this project, our main objective is to build a predictive model for patient outcome based on SPECT scans before treatment. We dichotomized patient outcome into two categories: responder vs non-responder by pre and post treatment Beck Depression Inventory (BDI) scores. Overall Among patients who have the same pre-BDI, the top 50% in BDI change are considered responders. We applied multiple common statistical and machine learning models for the purpose of classifying responders vs non-responders. A Bayesian regression analysis is also performed to identify brain regions of interest and other important variables. In addition, we repeated a similar study on the combined depression and ADHD dataset as a confirmatory step. The remainder of the paper is organized as follows. Section 2 gives

data description and exploratory analysis. Our analysis methods are discussed as follows: analysis of different machine learning methods in Section 3; regression analysis in Section 4; ADHD/depression analysis in Section 5. Finally, Section 6 gives the conclusions.

2 Data Description and Feature Reduction

The data set contains 954 observations (patients) with 1654 possible features. There are 11 main feature sets: Demographics, Comorbidities, T_Baseline, T_Concentration, Baseline, Concentration, Max cluster size, Min cluster size, Max cluster T_Baseline, Min cluster T_Baseline. There are two challenges with this data set. The first challenge is that most columns have missing values. Some columns have more than 90% of their values missing, that is because some patients did not have those measurements.

In order to be able to continue our analysis and reduce the noise, we eliminated the missing values in the following way. We first removed the columns that had more than 90% missing values, since they provide information for only a few patients in the data set. Then, we imputed the remaining missing values using Gelman's `mi` package in R, which imputes the missing values in an approximate Bayesian framework. This strong imputation technique allowed us to continue our analysis without losing any information.

The second challenge with this data set was that there were 1654 possible features that we could use, some of which were the different versions of the SPECT readings. We focused our analysis on T_Baseline feature set from the SPECT data. To decide which columns (brain regions) in the T_Baseline feature set are important in terms of their effect on co-diagnoses, we did recursive feature elimination. What this algorithm does is that it finds the important features by repeatedly constructing a model and removing the features that have low importance. We determined 10 important co-diagnoses which for depression, and found which brain regions were affecting them. These co-diagnoses were Anxiety Disorder, Childhood Disorder, Attention Deficit Disruptive Behavior, ADHD, Substance Abuse Disorder, Temporal Dysfunction, Frontal Lobe Dysfunction, Diagnosed

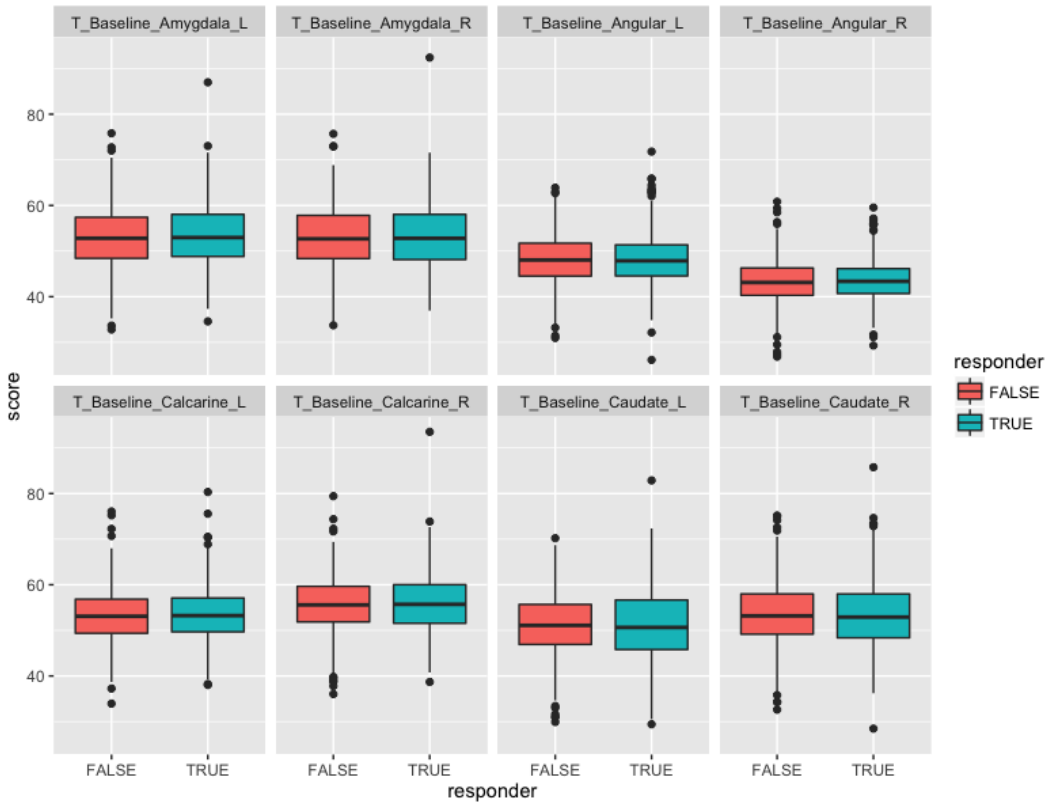

Figure 1: Boxplots of T scores in first eight brain regions colored by responder vs non-responder show minimal difference

Brain Trauma, GAD, PTSD.

Anxiety Disorder
T_Baseline_Cerebellum_Crus2_R,T_Baseline_Parietal_Sup_R, T_Baseline_Temporal_Inf_Post_L, T_Baseline_Vermis_8, T_Baseline_Occipital_Inf_R, T_Baseline_Occipital_Inf_L, T_Baseline_Precentral_R,T_Baseline_Postcentral_R, T_Baseline_Cerebellum_Crus1_L, T_Baseline_Occipital_Sup_L
Childhood Disorder
T_Baseline_Frontal_Mid_Orb_R_10, T_Baseline_Cerebellum_7b_R, T_Baseline_Cerebellum_8_L, T_Baseline_Paracentral_Lobule_L, T_Baseline_Hippocampus_R, T_Baseline_Pallidum_L, T_Baseline_Putamen_L, T_Baseline_Frontal_Inf_Orb_R, T_Baseline_Temporal_Inf_Ant_R, T_Baseline_Amygdala_L
Attention Deficit Disruptive Behavior
T_Baseline_Cerebellum_7b_R, T_Baseline_Cerebellum_8_L, T_Baseline_Frontal_Mid_Orb_R_10, T_Baseline_Amygdala_L, T_Baseline_Paracentral_Lobule_L, T_Baseline_Frontal_Mid_Orb_L_9 T_Baseline_Pallidum_R, T_Baseline_Temporal_Inf_Post_R, T_Baseline_Caudate_L, T_Baseline_Temporal_Inf_Ant_R
ADHD
[T_Baseline_Frontal_Inf_Orb_R, T_Baseline_Frontal_Mid_Orb_R_10, T_Baseline_Cerebellum_7b_R, T_Baseline_Hippocampus_R, T_Baseline_Cerebellum_8_L, T_Baseline_Temporal_Mid_Ant_R, T_Baseline_Angular_R, T_Baseline_Hippocampus_L, T_Baseline_Paracentral_Lobule_L, T_Baseline_Cerebellum_Crus1_L
Substance Abuse Disorder
T_Baseline_Heschl_L, T_Baseline_Temporal_Sup_Post_R, T_Baseline_Paracentral_Lobule_L, T_Baseline_Hippocampus_L, T_Baseline_Calcarine_L, T_Baseline_SupraMarginal_R, T_Baseline_Temporal_Mid_Post_R, T_Baseline_Cerebellum_3_R, T_Baseline_Temporal_Inf_Ant_L, T_Baseline_Vermis_6
Temporal Dysfunction
T_Baseline_Amygdala_R, T_Baseline_Vermis_6, T_Baseline_Cingulum_Mid_R, T_Baseline_Lingual_L, T_Baseline_Insula_L, T_Baseline_Caudate_L
Frontal Lobe Dysfunction
T_Baseline_Cingulum_Ant_R, T_Baseline_Cingulum_Mid_R , T_Baseline_Cerebellum_Crus2_R, T_Baseline_Rectus_R, T_Baseline_Putamen_R, T_Baseline_Temporal_Inf_Ant_R, T_Baseline_Cingulum_Ant_L, T_Baseline_Cerebellum_Crus2_L, T_Baseline_Frontal_Mid_Orb_R
Diagnosed Brain Trauma
T_Baseline_Cerebellum_8_R, T_Baseline_Temporal_Mid_Ant_L, T_Baseline_Paracentral_Lobule_R, T_Baseline_Cingulum_Ant_L, T_Baseline_Amygdala_R , T_Baseline_Frontal_Sup_Medial_L, T_Baseline_Frontal_Sup_Orb_R, T_Baseline_Temporal_Sup_Ant_R, T_Baseline_Amygdala_L
GAD
T_Baseline_Frontal_Mid_R, T_Baseline_Temporal_Inf_Mid_L , T_Baseline_Occipital_Inf_L, T_Baseline_Cerebellum_Crus2_L, T_Baseline_Caudate_R, T_Baseline_Cerebellum_3_R, T_Baseline_Frontal_Sup_Medial_R, T_Baseline_Frontal_Sup_R, T_Baseline_Frontal_Mid_Orb_R_10, T_Baseline_Occipital_Sup_L
PTSD
T_Baseline_Cingulum_Ant_R, T_Baseline_Frontal_Inf_Oper_L, T_Baseline_Supp_Motor_Area_L, T_Baseline_Cerebellum_Crus1_R, T_Baseline_Temporal_Inf_Mid_R, T_Baseline_Temporal_Pole_Sup_R, T_Baseline_Temporal_Mid_Ant_R , T_Baseline_Parietal_Inf_R , T_Baseline_Supp_Motor_Area_R, T_Baseline_Temporal_Inf_Ant_L

Table 1: Important features for co-diagnoses

As a result of the recursive feature elimination, we identified 10 regions for each co-diagnoses. After identifying important regions for co-diagnoses, we also identified significant brain regions that affect the outcome using logistic regression. The 9 significant regions at 5% level from logistic regression were: Cerebellum_9_L, Cingulum_Post_L, Fusiform_L, Fusiform_R, Pallidum_L, SupraMarginal_L, Temporal_Mid_Post_L, Temporal_Mid_Post_R, Temporal_Pole_Sup_L,Vermis_3.

In this section, we consider the problem of predicting the response (yes or no) based on the characteristics of the patients. In this work, we mainly focus on the t_baseline features and the t_concentration features, both of which are measures of the brain image in different regions.

3 Classification models

There is no single model that can work well for all datasets. In this work, we evaluated different types of classification algorithms in order to find the ones that are most suitable for this problem. Specifically, we consider the following classification models.

- k nearest neighbors (kNN): the prediction is based on the majority vote by labels of the neighboring training data.
- support vector machine (SVM): a discriminative classifier defined by a separating hyperplane, where the margin of the training data is maximized. Multiple kernels can be used to incorporate features of more complex structures. Popular kernels include linear kernel, polynomial kernel and RBF(Gaussian) kernel.
- Gaussian process: a classifier based on Gaussian process.
- Decision tree: an algorithm that recursively breaks down the training set into smaller pieces according to one feature at each time, and thus constructing a tree-structured model, where the prediction is computed via the leaf nodes.
- Random forest: ensemble of decision trees that are constructed using a subset of the features and training data. It usually has better performance comparing to decision tree as it reduces the possibility of model overfitting.
- Multilayer perceptron: a feedforward artificial neural network model that maps the input data to output through multiple hidden layers.
- AdaBoost: uses a sequence of weak learners that are tweaked in favor of the error made by the previous learners.
- Naive Bayes: a simple Bayesian classifiers where the features are assumed to be independent conditional on the classification labels.
- Quadratic discriminant analysis: the algorithm is similar to linear discriminant analysis (LDA) but there are no assumptions on the covariance matrices.

The dataset is randomly split into training and testing data, where the testing data contains about 20% of the data while the training data consists of the rest 80%. The model training is purely based on the training data and the evaluation is performed on the testing data. The performance of the model is evaluated by the accuracy, which is the proportion of the testing data that the model gives the correct prediction.

Table 2 displays the classification accuracy for the classification algorithms. All of the classification accuracy is around 50%, which means all the models are no better than random guessing. Considering that the labels are not balanced, the models actually perform worse than just predicting 1 for all test samples. Figure 2 shows the decision boundary for the models, plotted on a 2-D space spanned by the first two principal components. The projection onto the space spanned by the first two principal components accounts for 70% of the total variation. It is observable that the two classes are overlapping and are not linearly separable.

The results show that non of the single model has a reasonably well performance. Though disappointing, it does not necessarily mean that the models are useless for this problem, as the results are obtained for single model that are trained on raw features (no transformations at all). Better feature engineering, model stack and ensembling could boost the performance, when there are enough training samples. Moreover, the output of the models can be informative, for example, random forest model can also rank the importance of the input features.

model_name	only t_baseline	only t_concentration	both
Nearest Neighbors	0.492147	0.518325	0.492147
Linear SVM	0.481675	0.528796	0.481675
RBF SVM	0.523560	0.523560	0.523560
Gaussian Process	0.481675	0.518325	0.518325
Decision Tree	0.554974	0.460733	0.513089
Random Forest	0.513089	0.544503	0.539267
Multilayer Perceptron	0.486911	0.502618	0.481675
AdaBoost	0.465969	0.476440	0.513089
Naive Bayes	0.492147	0.465969	0.455497
Quadratic Discriminant Analysis	0.539267	0.560209	0.507853

Table 2: Evaluation of the classification accuracy of different models.

4 Bayesian Logistic Regression

To investigate the relationship between the SPECT image scans and some other covariates of interest, we fit three different logistic regression models to the data in a Bayesian framework using 'MCMCpack', software that facilitates MCMC-based computational Bayesian inference (Martin, Quinn, & Park, 2016). We use uninformative priors for the coefficients of

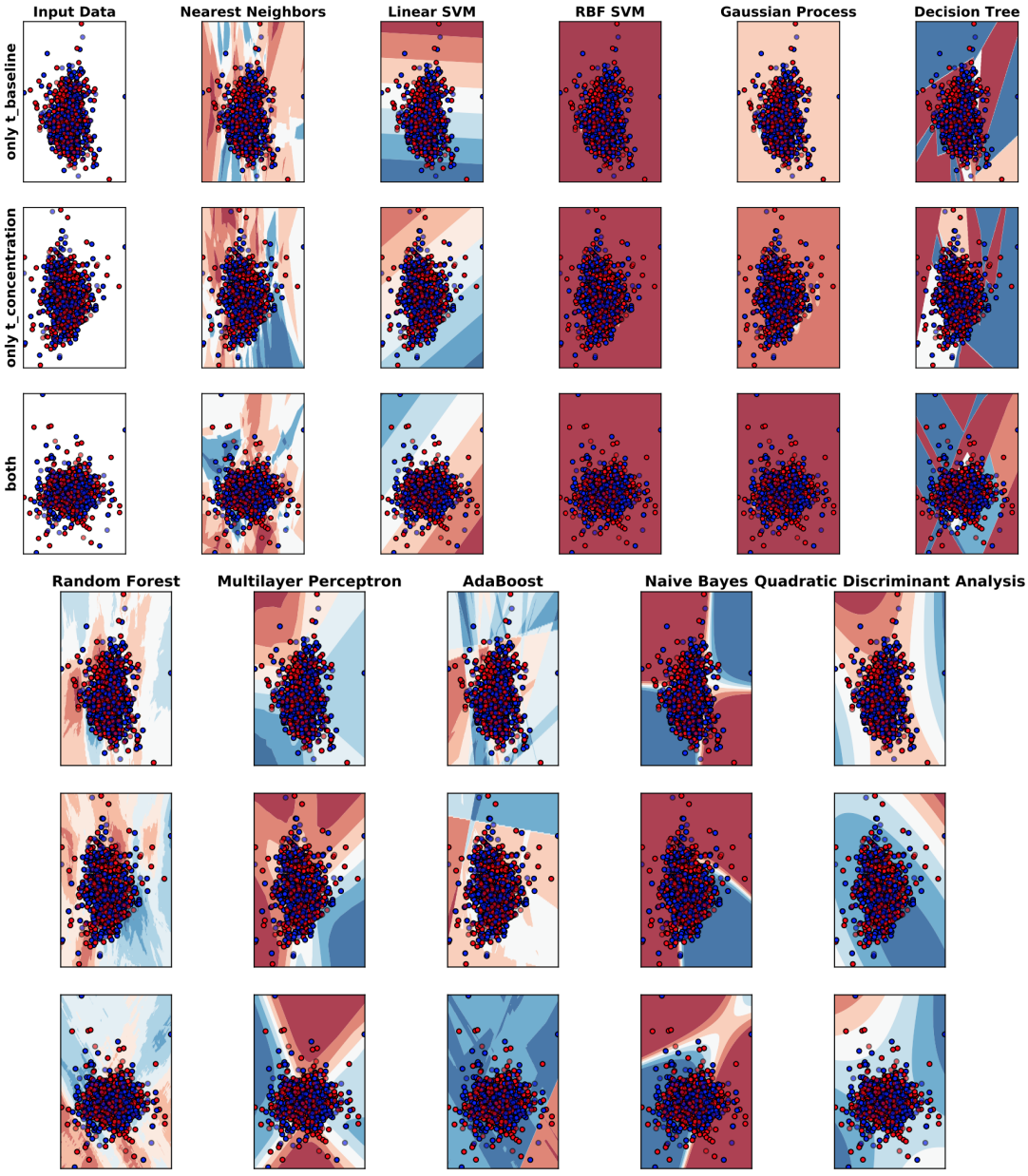

Figure 2: Decision boundary for the classifiers

predictors in all three models because we have no prior information from existing literature on how these covariates relate to response to hollistic treatment of depression. To control for pre-BDI scores, we naturally include the pre-BDI scores per patient as a predictor in

all three of our models.

For our first logistic regression model, we only consider pre-BDI scores and the top eight baseline brain regions from the dimensionality-reduction analyses: left cerebellum 7B, right crus cerebellum 1, left superior temporal pole, right rectus, right cerebellum 9, right superior temporal lobe, vermis 3, and right superior occipital. The model equation 1 is given below, along with the estimates of each coefficient using exponentiated posterior sample means and 95% credible intervals.

	exp(means)	exp(95% credible intervals)
Intercept	0.312	(0.049, 2.021)
Pre-BDI	1.009	(0.990, 1.030)
Left cerebellum 7B	0.980	(0.944, 1.018)
Right crus cerebellum 1	1.030	(0.989, 1.027)
Left superior temporal pole	1.015	(0.949, 1.088)
Right rectus	1.020	(0.980, 1.061)
Right cerebellum 9	0.990	(0.957, 1.024)
Right superior temporal pole	1.029	(0.969, 1.094)
Vermis 3	**0.964**	**(0.936, 0.993)**
Right superior occipital	1.002	(0.960, 1.047)

Table 3: Summary of results from Model 1

The results of our Bayesian logistic regression analysis indicate that the only brain region out of the eight regions tested in our model that is weakly associated with odds of responding to treatment is Vermis 3. We estimate that each additional unit of baseline activation of Vermis 3 is associated with about a 3.6% *decrease* in the odds of response to treatment, after controlling for pre-BDI scores and the baseline activation of the other 7 brain regions. Furthermore, there is a 95% probability that this association of Vermis 3 with odds of response to treatment is between a 0.7% and 6.4% decrease.

In Model 2, we consider pre-BDI scores with some co-diagnoses of interest: anxiety disorder, ADHD, substance abuse disorder, frontal lobe dysfunction, diagnosed brain trauma, and PTSD. The equation for Model 2 is given below, along with results from our Bayesian logistic regression analysis.

	exp(means)	exp(95% credible intervals)
Intercept	1.297	(0.580, 2.471)
Pre-BDI	1.010	(0.992, 1.032)
Anxiety Disorder	**0.613**	**(0.417, 0.957)**
ADHD	0.863	(0.625, 1.195)
Substance Abuse Disorder	1.337	(0.854, 2.195)
Frontal Lobe Dysfunction	0.872	(0.635, 1.250)
Diagnosed Brain Trauma	0.879	(0.631, 1.247)
PTSD	0.964	(0.642, 1.427)

Table 4: Summary of results from Model 2

From these results, the only co-diagnosis out of the six that are tested in Model 2 that is associated with odds of response to treatment is anxiety disorder. After adjusting for the other 5 co-diagnoses, we estimate that being diagnosed with anxiety disorder in addition to depression is associated with about a 39% *decrease* in the odds of response to treatment, compared to not being diagnosed with anxiety. Furthermore, there is a 95% probability that being diagnosed with anxiety is associated with between a 4.3% and 58.3% decrease in odds of response to treatment.

Lastly, we consider a third and final Bayesian logistic regression model in which we include pre-BDI scores, the eight brain regions from Model 1, the six co-diagnoses from Model 2, compliance ratings, age, and gender. The equation of Model 3 is given below, along with the table of results from our Bayesian logistic regression analysis.

Consistent with findings from Models 1 and 2, the results of our Bayesian logistic regression analysis of Model 3 show that Vermis 3 and anxiety disorder diagnosis are associated with the odds of response to treatment. After controlling all of the other covariates in the model, we estimate that each additional unit of the baseline activation of Vermis 3 is associated with about a 4.1% *decrease* in the odds of response to treatment. We also find a negative association between anxiety disorder and odds of response to treatment. Adjusting for all other covariates in the model, we estimate that being diagnosed with anxiety disorder is associated with about a 36.2% decrease in odds of response to treatment.

A new finding from the analysis of Model 3 indicates that holding all other covariates constant, a self-report of being "very compliant" with the treatment plan is associated with about a 113% increase in the odds of response to treatment, compared to not giving a self-report on compliance to treatment. Furthermore, we find that age and gender are not associated with odds of response to treatment.

All of the model results are based on 4 chains of 10,000 samples samples each. The chains were verified for convergence using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997).

	exp(means)	exp(95% credible intervals)
Intercept	0.343	(0.069, 3.057)
Pre-BDI	1.014	(0.992, 1.033)
Left cerebellum 7B	0.980	(0.941, 1.023)
Right crus cerebellum 1	1.038	(0.998, 1.077)
Left superior temporal pole	1.024	(0.956, 1.101)
Right rectus	1.013	(0.973, 1.066)
Right cerebellum 9	0.991	(0.957, 1.024)
Right superior temporal pole	1.027	(0.966, 1.089)
Vermis 3	**0.959**	**(0.935, 0.984)**
Right superior occipital	1.000	(0.960, 1.048)
Anxiety Disorder	**0.638**	**(0.443, 0.922)**
ADHD	0.921	(0.705, 1.277)
Substance Abuse Disorder	1.403	(0.974, 1.980)
Frontal Lobe Dysfunction	0.819	(0.581, 1.169)
Diagnosed Brain Trauma	0.922	(0.665, 1.260)
PTSD	1.052	(0.777, 1.500)
Compliance: Not	0.533	(0.271, 1.109)
Compliance: Somewhat	0.837	(0.499, 1.349)
Compliance: Very	**2.131**	**(1.262, 3.585)**
Age	1.002	(0.990, 1.013)
Gender: Male	1.186	(0.830, 1.692)

Table 5: Summary of results from Model 3

5 Bayes Factors

In addition to estimating the coefficients of our models in a Bayesian framework, we can also quantify the evidence for each of our models by making pairwise comparisons between them using Bayes factors. Bayes factors are a standard Bayesian approach to model selection that implicitly controls for goodness-of-fit and model complexity. It is the degree of change from prior to posterior information given by the data. We compute the natural logarithm of the Bayes factors between each pairwise comparison of our three models. A positive log Bayes factor indicates evidence for the first model in the comparison, and a negative log Bayes factor indicates evidence for the second model. Standard interpretive boundaries at log-odds of 2, 6, and 10 correspond to "moderate", "strong", and "very strong" (Kass & Raftery, 1995). The table below shows the matrix of the natural log Bayes factors resulting from all 9 pairwise comparisons between Models 1, 2, and 3.

	Model 1	Model 2	Model 3
Model 1	0.0	-29.4	55.3
Model 2	29.4	0.0	84.8
Model 3	-55.3	-84.8	0.0

Table 6: Bayes factors between Models 1, 2, and 3

There is overwhelming evidence that Model 2 with just pre-BDI and 6 co-diagnoses is preferred over both Models 1 and 3. In particular, the saturated Model 3 with the most predictors is the worst because it overfits the response data. Model 1 with just pre-BDI and 8 brain regions is preferred over the saturated Model 3, but not preferred when compared to Model 2. These results indicate very strong evidence that baseline activation of the 8 brain regions do not provide more information if pre-BDI and co-diagnoses are already included in the model.

6 ADHD/depression Classification

Along with the Depression patient data set, data from 1449 ADHD patients were also available from the Amen Clinics records. These patients also had their SPECT data recorded, along with their Pre- and Post-BDI scores (even though they were not being treated for depression specifically). Since the primary focus of this study is on the change in BDI for Depression patients, all patients with a Depression diagnosis were included in the Depression patient set; some of the Depression patients have an ADHD diagnosis, but none of the patients in the ADHD set were diagnosed with depression.

In an effort to understand the association of SPECT readings with depression, a random forest model is employed to classify patients by depression diagnosis, using all SPECT baseline and concentration readings, and adjustment variables age and gender. The random forest model is constructed as an aggregation of decision trees, each fit to a randomly chosen subset of the available predictors, and using the Gini impurity to choose the best splits. For each individual tree constructed for the forest, 50 covariates are randomly selected from the 258 total covariates (128 T Baseline, 128 T Concentration, Age, and Gender); 1000 such trees are built, and their individual predictions are aggregated to produce the random forest's prediction. First considering the forest model for diagnostic classification of ADHD from Depression patients using only SPECT, Age, and Gender, the test set ROC, which has a AUC of 0.7153, indicating moderate overall predictive performance, but probably under the level of clinical applicability. The plot of variable importance is given below. The plot makes it clear that age is by far the strongest predictor of depression diagnosis, which is expected from the strong correlation between age and depression status in the sample. The second strongest predictor is the SPECT baseline reading for Vermis 1-2, with a larger importance score than even gender, which is known to be significantly associated with depression diagnosis. The variable importance plot suggests that a small number of

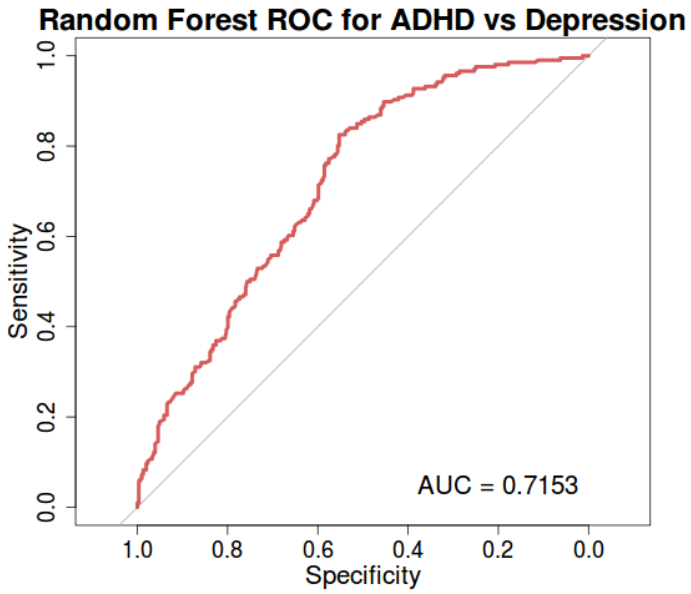

Figure 3: Test set ROC for random forest classification model for depression diagnosis. The predictors in this model are all SPECT baseline and concentration readings, and age and gender adjustment variables.

the SPECT regions are strongly associated with depression diagnosis, with the majority of regions not providing significant information regarding depression status. A sparse signal is expected, and this result confirms the need for dimensionality reduction when modeling with SPECT data.

7 Conclusion

Our models did not predict patient outcome at an accuracy level that we had hoped for. There are several possible explanations. First of all, patients were categorized into two groups in a somewhat arbitrary manner. It is easy to imagine that there are not substantial differences between patients whose BDI changes slightly differ. Yet patients with similar BDI changes could be responders and non-responders. Secondly, we did not expect logistic regression to achieve satisfactory results as the data are not linearly separable. In other words, responders and non-responders have comparable brain region scores overall. Though random forest and support vector machine are non-linear methods that can theoretically capture interactions between variables, the number of variables pose a challenge to create the most predictive features.
Our models did separate depression and ADHD patients and found important brain re-

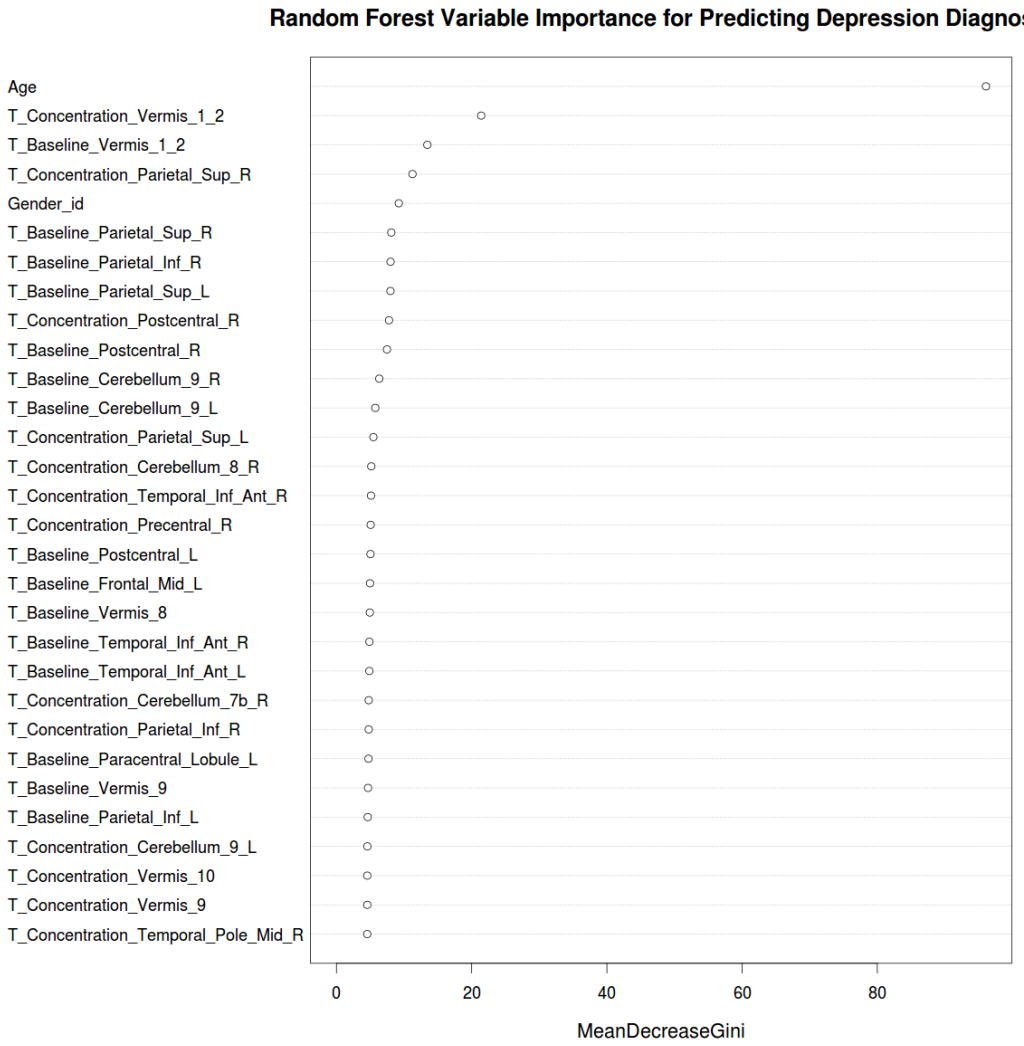

Figure 4: Variable importance plot for classifying depression patients from ADHD. Age the most important predictor by a wide margin. The most important SPECT co- variate is the baseline for Vermis 1-2.

gions adjusting for other variables. It implies that SPECT scans provide useful insights beyond the demographic and diagnostic information. This, without a doubt, could support the effectiveness of using SPECT scan for patient diagnosis as several previous studies. It is worth noting that in the combined depression and ADHD dataset, many patients are diagnosed with both, which increases the difficulty of the classification task.

After all, most patients improved their BDI scores to some degree regardless of their diagnosis and pre-BDI. Our Bayesian regression analysis suggests that unsurprisingly among patients who were very compliant, the odds of responding to treatment are estimated to be from 1.262, to 3.585 times higher compared to those who did not report. Perhaps above anything else ensuring patient compliancy would further improve patient outcome.