Predicting the Presence of Heart Disease with Machine Learning

Brian Chen, Michelle Ikoma, Juan Shi, John Won

Abstract

Heart disease is the leading cause of death in the United States; however, due to its insidious onset, cardiovascular disease can be difficult to detect before a catastrophic event such as a myocardial infarction occurs. Thus, this paper sought to understand whether machine learning can detect heart disease based on commonly available clinical information with an accuracy ≥ 90% and which of 3 machine learning algorithms (K-nearest neighbor [KNN], logistic regression, and random forest) predicts the presence of heart disease with the greatest accuracy. With all of our methods, we did a 5-fold cross-validation, while tuning for parameters within our cross-validation (CV) for KNN and Random Forest. Our cross-validation (CV) error rate with the logistic regression with a threshold of 0.5 was 0.165, while the CV error rate using the optimal threshold based on maximizing Youden's index was 0.132. After tuning the parameters in our 5-fold cross-validation, the CV error rate with KNN was 0.162, and the error rate with random forest was 0.199. Thus, the CV errors were similar across the different algorithms and accuracy greater than 90% was not achieved with any of our models. However, by comparing sensitivity measurements, we found logistic regression with Youden's index to have the highest true positivity rate (0.861). Our research provides a foundational step in incorporating machine learning to predict heart disease based on a collection of features.

Introduction

Heart disease is the leading cause of death in the United States, accounting for more than 650,000 deaths each year (Kochanek et al., 2020). However, heart disease develops gradually over time and often remains asymptomatic for several years. In many cases, catastrophic—and sometimes fatal—events like myocardial infarctions (more commonly known as "heart attacks") are the first indication that individuals have underlying pathology in their cardiovascular system.

Thus, researchers have invested an enormous amount of effort to detect heart disease in earlier stages. Several risk factors for cardiovascular disease have been identified, including hypertension (also known as high blood pressure), diabetes mellitus, high cholesterol, smoking history (current or past), and advanced age. Patients deemed to be at intermediate risk for heart disease based on these and other risk factors are often sent for additional tests such as electrocardiograms (ECGs) and/or exercise stress tests. Those considered to be at high risk for coronary artery disease are sent for invasive testing in which fluoroscopy is used to visualize the coronary arteries to determine if any vessels are significantly occluded.

This data on patients' risk factors and comorbid conditions combined with data from clinicians' histories and diagnostic tests ultimately helps health care providers determine whether an individual has or does not have heart disease. However, given the vast quantity of information clinicians must synthesize and inherent subjectivity in diagnosing chronic, progressive

conditions, many cases of heart disease go undetected while others are erroneously diagnosed with heart disease.

Advanced computational methods like machine learning hold tremendous promise to accurately detect heart disease based on available clinical data. In recent years, the body of research on machine learning applications in the diagnosis and management of cardiovascular disease has grown exponentially. In 2010, a PubMed search for the term "'Machine learning' AND (heart OR cardiovascular OR cardiac) AND (disease)" returned only 9 articles compared to 586 in 2020. Machine learning algorithms have been applied to several areas of cardiology research and practice including developing risk stratification models, predicting disease endpoints (e.g. myocardial infarction, stent restenosis), and cardiac imaging interpretation (Al'Aref et al., 2019; Ambale-Venkatesh et al., 2017; Johnson et al., 2018).

The purpose of this paper was to understand (1) whether machine learning can be used to detect heart disease based on commonly available clinical information with an accuracy ≥ 90%, and (2) which machine learning algorithm predicts the presence of heart disease with the greatest accuracy.

Methods and Materials

For this data analysis, we used the "Heart Disease UCI" dataset available at Kaggle.com. This dataset includes information about 13 clinical parameters used to determine the presence or absence of heart disease in a patient as well as a binary "target" value indicating the presence or absence of heart disease for 303 individuals.

Detailed information about each clinical parameter is summarized in Table 1. In addition, the following adjustments were made to clean the raw dataset prior to analysis based on insights obtained from the discussion forum associated with the dataset: (1) "ca" values of 4 were recoded as NA since these were coded as NA in the original dataset; (2) "thal" values of 0 were recoded as NA since these were coded as NA in the original dataset; (3) "target" values were inverted such that 0 = no heart disease and 1 = heart disase so that prediction algorithms would follow the conventional interpretation of 0 indicating disease absence and 1 indicating disease presence.

In our analysis, we used three different types of algorithms: Logistic Regression, K-nearest neighbor (KNN) and random forest. To select the algorithm that has the best performance, we used a 5-fold cross-validation. All analysis was done in R version 4.03 (R Core Team , 2020), and the primary goal was to select the algorithms that have the smallest cross validation error.

Logistic Regression: Logistic regression is a classification algorithm used to find the probability of event success and event failure. All predictors were included in this model to predict the "target" variable, representing the binary outcome of heart disease (1) or no heart disease (0). Two different thresholds were used to classify heart disease or no heart disease given the probability of the outcome. Our first threshold was a 0.5 threshold, where >0.5 was classified as having heart disease and ≤0.5 was classified as no heart disease. We also ran logistic regression

where we selected the threshold based on Youden's index (the closest point to the top left of the ROC curve) within each cross-validation fold to maximize sensitivity and specificity. The cross validation error and cross validation standard error were then computed.

K-nearest Neighbor (KNN): KNN is a classification algorithm that assumes that similar things exist in close proximity. In our study all the predictors were included in this model to predict the "target" variable, representing the binary outcome of heart disease (1) or no heart disease (0). Within each cross-validation fold, we tuned our algorithm to the training set to find the optimal number of neighbors using a tune length of 20. The cross validation error and cross validation standard error were then computed.

Random Forest: Random forest is one kind of decision trees which make an improvement over bagged trees by way of a small tweak that decorrelates the trees. When building decision trees, each time a split in a tree is considered, a random sample of predictors is chosen as split candidates from the full set of predictors. In our study, all predictors were included in this model to predict the "target" variable, representing the binary outcome of heart disease (1) or no heart disease (0). A 5-fold cross validation was used across a grid of tuning parameters to find the choice resulting in the lowest CV estimate of test error. The parameters tuned were the number of trees (50, 250, and 500) and the number of predictors selected at each split ($p/2 = 6.5$, $\sqrt{p} = 3.606$, $p = 13$).

Results

After correcting the misclassified variables and deleting them from the dataset, there were 296 out of 303 patients left in the dataset. Table 2 provides descriptive statistics of each variable. Stratified by the target variable (heart disease or no heart disease), means and standard deviations are given for continuous variables and proportions are given for categorical variables. Moreover, an analysis of variance (ANOVA) test was done between the two target groups for continuous and a Chi-Squared test was done for categorical variables between the two target groups. For these tests, all of the variables produced significant p-values between the two target groups except for Cholesterol and Fasting Blood Sugar > 120 mg/dL. However, given their known relationship with heart disease, they were kept in the features used in our model. Figure 1 shows the pairwise correlations between continuous variables, but these correlations were not high enough for us to suspect collinearity between our variables. Notably, the correlation between the slope of the peak exercise ST segment and ST depression induced by exercise relative to rest had a correlation of 0.578, but both variables were kept in our analysis.

Logistic Regression

Table 3 shows the CV error of our 5-fold cross-validation for each of our 4 methods. Comparing methodologies between using a common threshold of 0.5 to choosing the threshold that maximized Youden's index, the Youden's index method produced a lower CV error. The 0.5 threshold method produced a CV error of 0.165 and standard error of 0.032 while the Youden's index method had a CV error of 0.132 and standard error of 0.039. However, our thresholds

based on Youden's index were variable, having a mean of 0.475 and standard deviation of 0.207. Additionally, Table 4 shows the sensitivity and specificity for each model. The average sensitivity for the 0.5 threshold method was 0.779 with a standard error 0.069, and a CV specificity of 0.881 and a standard error of 0.056. The CV sensitivity for the Youden's index threshold method was 0.861 with a standard error of 0.111, and a CV specificity of 0.875 with a standard error of 0.062.

KNN

When tuning our KNN, the average number of neighbors was 37 with a standard deviation of 8.124. This produced an CV error of 0.162 with a CV standard error of 0.033. Moreover, KNN had a CV sensitivity of 0.809 with a standard error of 0.096, and a CV specificity of 0.863 and standard error of 0.052.

Random forest

Our random forest model was tuned by the number of trees and the number of predictors selected at each split. The best tuning parameters for each fold is shown in Table 5. The most common number of trees was 250 trees and 6.5 (p/2) predictors. This tuning achieved a CV error of 0.199 and a standard error of 0.047. Additionally, this produced a CV sensitivity of 0.757 with a standard error of 0.094, and a CV specificity of 0.838 and a standard error of 0.068.

Discussion

Heart disease is the leading cause of death in the United States, accounting for about 1 in 4 deaths (Kochanek et al., 2020). Maintaining a healthy lifestyle that consists of healthy foods, regular physical activity, and abstaining from smoking can reduce one's risk of heart disease or heart attack (CDC, 2021). However, according to a recent National Center for Health Statistics (NCHS) report, only 22.9% of American adults (ages 18 to 64) meet CDC recommendations for activity level each week (Blackwell, 2018). In addition, an estimated 14.0% (34.1 million) of U.S. adults were cigarette smokers in 2019 (CDC, 2021). This begins to suggest the vast potential population at risk for heart disease and the heightened relevancy of our research.

The asymptomatic nature of heart disease coupled with its potentially fatal consequences emphasize the importance of early detection through artificial intelligence. With the growing abundance of medical data, there is a clear need for advanced approaches to optimize patient health outcomes, especially with conditions as prevalent as heart disease. Thus, machine learning provides a potential solution that will allow clinical practice to become more efficient, convenient, and personalized (Johnson et al., 2018). Undoubtedly, well-developed algorithms can ultimately supplement cardiologists to provide more accurate diagnoses without conducting expensive and invasive procedures.

When using a 5-fold cross validation approach, Youden's index logistic regression obtained the lowest mean CV error (0.132) and a relatively low CV standard error (0.0338). This model's CV standard error was only slightly higher than the CV standard error of logistic regression using a

0.5 threshold (0.0321). As expected, choosing the threshold based on Youden's index produced a lower CV error than using a fixed threshold of 0.5 since this was an additional means of tuning the logistic regression model. However, the mean of Youden's index was 0.475 with a standard deviation of 0.207 suggesting high variability around the fixed threshold of 0.5 used for the other logistic regression model. As a result, although the use of Youden's index would provide more accurate predictions in theory, we cannot conclusively state that this is the case for our dataset based on CV error alone. However, when we take into account sensitivity and specificity values, logistic regression with Youden's index (sens = 0.861) delivered the highest CV sensitivity. Because of the aforementioned nature of heart disease, we are primarily concerned with sensitivity over specificity; we seek to accurately identify those with the disease. The model with the next highest CV sensitivity was our KNN model at 0.809, followed by logistic regression using 0.5 threshold (0.779) and random forest (0.757).

We can speculate the reasons for the lower performances for these other models. In comparison to our other tested models, logistic regression with Youden's index performed better than both KNN (CV error = 0.155, CV SE = 0.041) and random forest (CV error = 0.199, CV SE = 0.047). Although random forest models have demonstrated strong predictive ability in other research (Couronné, R.2018), our specific RF model may have been prone to overfitting, resulting in a decreased performance in terms of CV error rates. Meanwhile, a KNN model may struggle to predict heart disease if the dataset becomes too large due to the cost of distance computation. Moreover, testing a greater variety of methods, such as SVM, could potentially produce a more powerful model with higher sensitivity values.

Ultimately, none of our models achieved an accuracy of greater than 90%, and from the results above, the three algorithms that we used for this prediction are limited. One possible reason for these lower accuracies relative to our expectation is that both our variables and sample were limited. In future studies, we could increase the dataset in order to more thoroughly test our results. In addition, we could potentially gain more insight into heart disease outcome by stratifying patients based on certain conditions such as age group. Collecting more detailed information such as patients' weights or historical disease records may also allow for more meaningful results. In general, this is a topic that is worth future study because of the prevalence and fatal consequences of heart disease.

References

1. Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., Pandey, M., & Maliakal, G. (2019). *Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging*. *40*(24), 1975–1986.
2. Ambale-Venkatesh, B., Yang, X., Wu, C. O., Liu, K., Hundley, W. G., McClelland, R., Gomes, A. S., Folsom, A. R., Shea, S., Guallar, E., Bluemke, D. A., & Lima, J. A. C. (2017). Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circulation Research*, *121*(9), 1092–1101. https://doi.org/10.1161/CIRCRESAHA.117.311312
3. Blackwell, D. L., & Clarke, T. C. (2018). State Variation in Meeting the 2008 Federal Guidelines for Both Aerobic and Muscle-strengthening Activities Through Leisure-time Physical Activity Among Adults Aged 18–64: United States, 2010–2015. *National Health Statistics Reports*. No. 112. https://www.cdc.gov/nchs/data/nhsr/nhsr112.pdf
4. CDC. (2021). Burden of Cigarette Use in the U.S. *Centers for Disease Control and Prevention*. https://www.cdc.gov/tobacco/campaign/tips/resources/data/cigarette-smoking-in-united-states.html.
5. CDC. (2021). Heart Disease. *Centers for Disease Control and Prevention*. https://www.cdc.gov/heartdisease/.
6. Couronné, R., Probst, P. & Boulesteix, AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 19, 270 (2018). https://doi.org/10.1186/s12859-018-2264-5
7. Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521
8. Kochanek, K. D., Xu, J., & Arias, E. (2020). *Mortality in the United States, 2019* (NCHS Data Brief No. 395). National Center for Health Statistics (NCHS). https://www.cdc.gov/nchs/products/databriefs/db395.htm
9. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Appendix

Table 1

Predictor	**Description**
age	Age in years
sex	Sex (1: male, 0: female)
cp	Chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
trestbps	Resting systolic blood pressure in mmHg
chol	Serum cholesterol in mg/dL
fbs	Presence of fasting blood sugar > 120 mg/dL (1: present, 0: absent)
restecg	Resting electrocardiogram results (0: normal, 1: ST-T wave abnormality present, 2: Probable or definite left ventricular hypertrophy present by Estes' criteria)
thalach	Maximum heart rate achieved during exercise stress test
exang	Presence of exercise induced angina (1: present, 0: absent)
oldpeak	ST segment depression induced by exercise relative to rest in mm
slope	Slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
ca	Number of major coronary vessels (0-3) colored by fluoroscopy
thal	Results of nuclear stress test (3: normal, 6: fixed defect, 7: reversible defect)

Figure 1

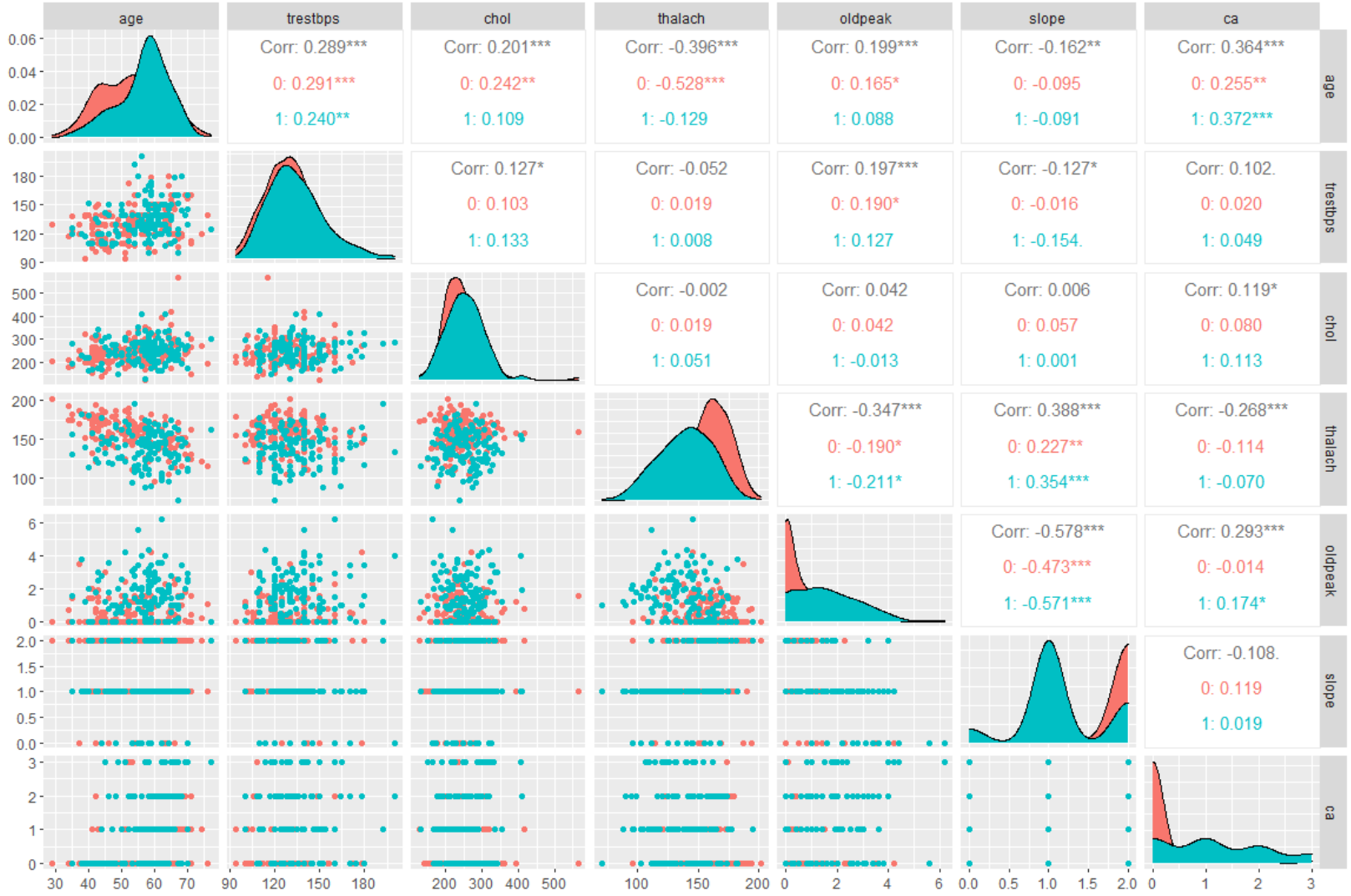
age
trestbps
chol
thalach
oldpeak
slope
ca
Corr: 0.289***
0: 0.291***
1: 0.240**
Corr: 0.201***
0: 0.242**
1: 0.109
Corr: -0.396***
0: -0.528***
1: -0.129
Corr: 0.199***
0: 0.165*
1: 0.088
Corr: -0.162**
0: -0.095
1: -0.091
Corr: 0.364***
0: 0.255**
1: 0.372***
Corr: 0.127*
0: 0.103
1: 0.133
Corr: -0.052
0: 0.019
1: 0.008
Corr: 0.197***
0: 0.190*
1: 0.127
Corr: -0.127*
0: -0.016
1: -0.154.
Corr: 0.102.
0: 0.020
1: 0.049
Corr: -0.002
0: 0.019
1: 0.051
Corr: 0.042
0: 0.042
1: -0.013
Corr: 0.006
0: 0.057
1: 0.001
Corr: 0.119*
0: 0.080
1: 0.113
Corr: -0.347***
0: -0.190*
1: -0.211*
Corr: 0.388***
0: 0.227**
1: 0.354***
Corr: -0.268***
0: -0.114
1: -0.070
Corr: -0.578***
0: -0.473***
1: -0.571***
Corr: 0.293***
0: -0.014
1: 0.174*
Corr: -0.108.
0: 0.119
1: 0.019

Table 2

Characteristic	N	0, N = 160[1]	1, N = 136[1]	p-value[2]
Age	296	53 (10)	57 (8)	<0.001
Sex	296			<0.001
0		71 / 160 (44%)	24 / 136 (18%)	
1		89 / 160 (56%)	112 / 136 (82%)	
Chest pain type	296			<0.001
0		39 / 160 (24%)	102 / 136 (75%)	
1		40 / 160 (25%)	9 / 136 (6.6%)	
2		65 / 160 (41%)	18 / 136 (13%)	
3		16 / 160 (10%)	7 / 136 (5.1%)	
Resting systolic BP	296	129 (16)	134 (19)	0.010
Cholesterol	296	243 (54)	251 (50)	0.2
Fasting blood sugar > 120 mg/dL	296			>0.9
0		137 / 160 (86%)	116 / 136 (85%)	
1		23 / 160 (14%)	20 / 136 (15%)	
Resting ECG results	296			0.006
0		67 / 160 (42%)	78 / 136 (57%)	
1		92 / 160 (57%)	55 / 136 (40%)	
2		1 / 160 (0.6%)	3 / 136 (2.2%)	
Maximum HR	296	159 (19)	139 (23)	<0.001
Exercise induced angina	296			<0.001
0		137 / 160 (86%)	62 / 136 (46%)	
1		23 / 160 (14%)	74 / 136 (54%)	
ST depression induced by exercise	296	0.60 (0.79)	1.60 (1.30)	<0.001
Slope of peak exercise ST segment	296			<0.001
0		9 / 160 (5.6%)	12 / 136 (8.8%)	
1		48 / 160 (30%)	89 / 136 (65%)	

Characteristic	N	0, N = 160[1]	1, N = 136[1]	p-value[2]
2		103 / 160 (64%)	35 / 136 (26%)	
# major vessels seen on fluoroscopy	296			<0.001
0		129 / 160 (81%)	44 / 136 (32%)	
1		21 / 160 (13%)	44 / 136 (32%)	
2		7 / 160 (4.4%)	31 / 136 (23%)	
3		3 / 160 (1.9%)	17 / 136 (12%)	
Stress test results	296			<0.001
1		6 / 160 (3.8%)	12 / 136 (8.8%)	
2		127 / 160 (79%)	36 / 136 (26%)	
3		27 / 160 (17%)	88 / 136 (65%)	

[1]Mean (SD); n / N (%)

[2]One-way ANOVA; Pearson's Chi-squared test; Fisher's exact test

Table 3

Method	CV Error Mean	CV Standard Error
Logistic: 0.5 Threshold	0.165	0.032
Logistic: Youden's Index	0.132	0.039
KNN	0.162	0.033
Random Forest	0.199	0.047

Table 4

Method	CV Sensitivity	CV Sensitivity SE	CV Specificity	CV Specificity SE
Logistic: 0.5 Threshold	0.779	0.069	0.881	0.056
Logistic: Youden's Index	0.861	0.111	0.875	0.062
KNN	0.809	0.096	0.863	0.052
Random Forest	0.757	0.094	0.838	0.068

Table 5

# of Trees	# of Features	OOB Error
500	6.500	0.156
500	3.606	0.152
250	6.500	0.194
250	6.500	0.169
250	13.000	0.190