7-6-2020

A Study Of The Efficacy Of Machine Learning For Diagnosing Obstructive Coronary Artery Disease In Non-Diabetic Patients

Demond Larae Handley
Illinois State University, dlhandley05@gmail.com

A STUDY OF THE EFFICACY OF MACHINE LEARNING FOR DIAGNOSING OBSTRUCTIVE CORONARY ARTERY DISEASE IN NON-DIABETIC PATIENTS

DEMOND LARAE HANDLEY

32 Pages

According to the Centers for Disease Control and Prevention, about 18.2 million adults age 20 and older have Coronary Artery Disease in the United States. Early diagnosis is therefore of crucial importance to help prevent debilitating consequences, and principally death for many patients. In this study we use data containing gene expression values from peripheral blood samples in 198 non-diabetic patients, with the goal of developing an age and sex gene expression model for diagnosis of Coronary Artery Disease. We employ machine learning methods to obtain a classification based on genetic information, age and sex. Our implementation uses feed forward neural networks, support vector machines and random forest classification. The neural network outperforms not only the other two but also an early Ridge Regression algorithm that used age, sex, and 23 genes clustered in a set of six metagenes. Our analysis provides valuable insight into the increasing effectiveness of machine learning applied to CAD diagnosis.

KEYWORDS: Machine Learning, Coronary Artery Disease, Classification, Neural Networks, Genetics

A STUDY OF THE EFFICACY OF MACHINE LEARNING FOR DIAGNOSING OBSTRUCTIVE CORONARY ARTERY DISEASE IN NON-DIABETIC PATIENTS

DEMOND LARAE HANDLEY

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Mathematics

ILLINOIS STATE UNIVERSITY

2020

A STUDY OF THE EFFICACY OF MACHINE LEARNING FOR DIAGNOSING OBSTRUCTIVE CORONARY ARTERY DISEASE IN NON-DIABETIC PATIENTS

DEMOND LARAE HANDLEY

COMMITTEE MEMBERS:

Olcay Akman, Chair

Epaminondas Rosa

Maochao Xu

CONTENTS

TABLES

FIGURES

CHAPTER I: INTRODUCTION

Obstructive Coronary Artery Disease (CAD) is the most common heart disease in the united states [1]. CAD is the result of major plaque buildup in the arteries, which narrows the blood flow channel. The plaque is typically made up of cholesterol restricting blood flow to the heart, resulting in angina (chest pain) and heart attacks. This process is called atherosclerosis. Such processes build up over years and are the result of obesity, unhealthy eating and using products such as tobacco. Several tests are used to diagnose CAD. These tests range from electrocardiograms (ECG) [2], echocardiograms [3], exercise stress tests [4], coronary angiograms [5], etc. While these test do exist, identifying if a patient has CAD involves a genetic component which shall be discussed.

Genetics play an essential role in why diseases spread among people. The genetics of coronary artery disease are no exception. Genome-wide association studies have been conducted to further understand this relationship. [6] confirmed the role of genetics in diabetic and non-diabetic patients, by identifying three loci associated with CAD. [7] identified an additional six loci associated with CAD. Both studies were conducted using different genomic data sets. Viruses such as HIV are also associated with CAD. [8] showed this by identifying twenty-three SNPs associated with both CAD in HIV infected patients. Machine learning techniques play a role identifying genetic predictors. Such techniques will be discussed next.

Machine learning techniques have been employed to examine the effects of disease on populations. Machine learning is the study of algorithms and statistical models that computational machines use to complete specified tasks without instruction. The Models learn how to complete such tasks over time, often given data to train on. Such techniques include artificial neural networks, support vector machines (SVM), random forest, regression analysis, Bayesian networks and genetic algorithms [9]. Each technique is designed to address a variety of problems in genetics.

Artificial neural networks are computational systems designed and modeled after

mammalian brains [10]. They are used to solve complex problems in real-world situations. Neural networks had been utilized in several genetic studies involving health issues. [11] developed a neural network model to identify genes associated with pancreatic ductal adenocarcinoma. They were successful in identifying over forty genes, where most were associated with cytoskeleton reorganization.

Neural networks have been useful for developing diagnosis models for certain cancers. [12] developed a model to test for colorectal cancer, using gene expression profiles along with artificial intelligence algorithms. Their best model proved efficient with 100% accuracy. [13] developed a model to classify leukemia, using gene expression profiles. They tested their model against SVM, Logistic Regression, Naïve Bayesian Networks, etc. and determined that the neural network performed best. Such studies have proved the utility of artificial neural networks in the medical field.

The focus of this study was the development three machine learning models to classify obstructive CAD in non-diabetic patients. The premises for this work were based on a previous study designed by [14], where a Ridge Regression model was developed test for CAD. Neural network models were developed for both males and females and tested against the Ridge Regression models. For further comparison, SVM and random forest classification models were developed. The development of each model is described below. The performance metrics for each model are given followed by a discussion comparing their efficiencies.

CHAPTER II: RIDGE REGRESSION METHODS

II.1 Data

The data were collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. There were 198 patients reported, where ninety-nine were diagnosed with CAD and ninety-nine healthy. The data consist of gene expressions collected from whole blood samples of patients across multiple testing centers throughout the United States. The blood samples were collected prior to coronary angiography. Patients were classified as having CAD if there was $> 50\%$ stenosis in at least one major coronary vessel and were healthy if there was $\leq 50\%$ stenosis. The data were gathered using the R packages Biobase and GEOquery.

The original twenty-three predictor genes, patient age, sex and disease state were all extracted from the GEO data set. The patient age was given in standard date format and was converted to actual age. Sex and disease state were binary variables and were coded as such. For patient sex, 0 = female and 1 = Male. For the disease state, 0 = Healthy/Control and 1 = Disease/Case. The data were then stratified into male and female sets according to the sex variable. This gave a total of 150 male patients and forty-eight female patients. The metagene expressions were calculated based on the formulas provided by [14]. Once metagene expressions were calculated, both data sets were split into testing and training sets. The training set consisted of seventy percent of the data, while the test set consisted of thirty percent. Because Ridge Regression was used to develop the previous model in [14], it suffices to discuss it.

II.2 Ridge Regression Classification Model

Ridge Regression, a commonly used traditional model, [14], was utilized during a trial to develop a diagnostic test for CAD in non-diabetic patients. Because of the equal number of case/control patients, this was a pairwise design. These patients were originally a part of the Personalized Risk Evaluation and Diagnosis in the Coronary Tree

(PREDICT) cohort, who participated in a multi-center study referring patients for coronary angiography. Whole blood samples were taken from each patient and were examined as microarray data.

The trials were initiated with 45,000 genes. Logistic regression was used to reduced the number of associated genes to 5,935. Reverse transcription polymerase chain reaction (RT-PCR) was then used to highlight genes with higher expression values. This reduced the number of associated genes to 113. Hierarchical clustering analysis was conducted to group the genes into meta-genes based on their association with specific cells. There were eighteen cluster groups, each representing genes associated with different white blood cells. The cell specific genes were grouped into meta-genes.

Meta-genes that were independently associated with CAD were selected via the LASSO method [15]. The final algorithm was fitted using Ridge Regression [16], where the predictors were the metagenes selected from LASSO along with sex and age. Ridge Regression is a technique that is useful in mitigating issues of co-linearity in linear regression problems. Ridge Regression attempts to mitigate such problems by imposing a penalty on the size of model coefficients. The Ridge coefficients minimize a penalized residual sum of squares. Thus, the optimal function (residual sum of squares) is given in equation 1 below.

$$\sum_{i=1}^{M}(y_i - \hat{y_i})^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \times x_{ij})^2 + \lambda \sum_{j=0}^{p} w_j^2 \tag{II.1}$$

y_i represent the predicted values, $\hat{y_i}$ is the observed data, w_j is represent the weight of each input variable, x_{ij} are the input variable and λ is the penalty term. λ regularizes the coefficients such that if the coefficients take large values, the optimization function is penalized.

The response variable was the disease state. In total, the final model consisted of twenty-three genes grouped into six meta-genes. The Ridge Regression model for males and females are given in equations (2.1)-(2.2) and are defined as

$$y = Age \ + \ (N_{up} - N_{down}) \ + \ (SCA_1 - Norm_1) \ + \ (NK_{up} - Tcell) \\ + \ (Bcell - Tcell) \ + \ (AF_2 - Norm_2) \ + \ Tspan, \quad (2.1)$$

$$y = (age - 60) + (N_{up} - N_{down}) + (SCA_1 - Neut) + (NK_{up} - Tcell) + \\ (Bcell - Tcell) + (AF_2 - Norm_2), \quad (2.2)$$

where y is the disease state. The following terms were used in both models:

$$N_{up} = \frac{1}{3}CASP5 + \frac{1}{3}IL18RAP + \frac{1}{3}TNFAIP6 \quad (2.3)$$

$$N_{down} = \frac{1}{4}IL8RB + \frac{1}{4}TNFRSF10C + \frac{1}{4}TLR4 + \frac{1}{4}KCNE3 \quad (2.4)$$

$$SCA_1 = \frac{1}{3}S100A12 + \frac{1}{3}CLEC4E + \frac{1}{3}S100A8 \quad (2.5)$$

$$Norm_1 = RPL28 \quad (2.6)$$

$$Neut = \frac{1}{2}AQP9 + \frac{1}{2}NCF4 \quad (2.7)$$

$$NK_{up} = \frac{1}{2}SLAMF7 + \frac{1}{2}KLRC4 \quad (2.8)$$

$$Tcell = \frac{1}{2}CD3D + \frac{1}{2}TMC8 \tag{2.9}$$

$$Bcell = \frac{2}{3}CD79B + \frac{1}{3}SPIB \tag{2.10}$$

$$AF_2 = AF_2gene \tag{2.11}$$

$$Norm_2 = \frac{1}{2}HNRNPF + \frac{1}{2}TFCP2 \tag{2.12}$$

$$Tspan = \begin{cases} 1, & if\ AF161365 - Norm_2 > 6.27\ or\ AF161365 = NoCall \\ 0, & Otherwise \end{cases} \tag{2.13}$$

The terms of each expression represent a weighted gene. The terms in (2.6) and (2.7) contain male and female specific genes respectively. The first terms in (2.1)-(2.2) contain (2.3)-(2.4). It represent the ratio of neutrophil expressed meta-genes that are up and down regulated with CAD. For males, the second term contains (2.5)-(2.6), and for the female model, the second term contains (2.5) and (2.7). These are sex specific for each model and are thought to perhaps reflect sex-specific differences in the significance of neutrophil counts in CAD and myocardial infarction (MI). The third term in both models contain (2.8)-(2.9). Term three represents two natural killer (NK) cell receptors $SLAMF7$ and $KLRC4$, which are regulated to the T-cell specific genes $CD3D$ and $TMC8$. NK cells tend to be associated with low lymphocyte cell counts during cardiac events.

Term four contains (2.9)-(2.10). This represents a gene expression-based measure of the B-cell T-cell ratio. Up-regulation of B-cells was correlated with CAD. The fifth term contains (2.11)-(2.12) and the sixth term contains (2.13). (2.13) contains genes of unknown function. However, it is seen that (2.13) is male specific and still requires further investigation for full understanding. Now that Ridge Regression had been discussed, it will be sufficient to go over the details of neural networks, support vector machines and random forest classification. These three machine learning methods were used to compare performance efficacy to the Ridge Regression model.

CHAPTER III: MACHINE LEARNING METHODS

III.1 Neural Network Model

Artificial neural networks are computational systems based on the biological neural systems in the brain. They are composed of highly interconnected processing elements, which process information by their dynamic state response to external outputs [10]. These processing elements are called nodes and they are modeled after biological neurons. A neural network can range from three simple nodes to thousands of interconnected nodes. However, this is still pale compared to the billions of neurons in a mammalian brain. Neural networks are usually organized into layers. The simplest neuron structure is composed of three layers: an input layer, hidden layer and output layer [17]. Each layer is made up of interconnected nodes which contain an activation function. The data is fed into the input layer, which communicates to the hidden layers, where the data is processed through weighted connections. The hidden layers then connect to the output layer where the network gives a response. This is usually the response variable when training the network. Figure 1 shows a diagram of this process.

The activation function is a computational way of simulating an action potential within neural networks [18]. It defines the output of a node, given a set of inputs. The input of an activation function of a node in layer l is the sum of outputs from nodes in the previous layer $l-1$. For example, if layer $l-1$ had four nodes and each node gave an output h_j^{l-1}, then the input of a node in layer l can be defined given in equation (3).

$$x = \sum_{j=1}^{4} w_{ij} h_j^{l-1} \tag{3}$$

w_{ij} represents the weight between node i in layer l and node j in layer $l-1$. There are many types of activation functions, but the hyperbolic tangent function was used in this work. The function is defined in equation (4).

$$\tanh(x) \ = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4}$$

The hyperbolic tangent function was best used to activate the output node of the neural network in this paper, given that the outcome is binary and that it provided the best results. The output variable follows a binomial distribution and because the range of the function is $(-1, 1)$, the hyperbolic tangent function was be appropriate for its use modeling the activation of the output node.

The model was trained using stochastic gradient descent via the backpropagation algorithm [19]. The loss function used to minimize the error is the binary cross-entropy function. The cross-entropy function calculates the probability that a given instance in our data will belong to a certain class. As the data is training, it updates regularly to reflect how well the model is progressing in its prediction. It is defined in equation (5) as

$$f\left(\hat{y},\ y,\ W\right) = \ -y\ln\left(\hat{y}\right) \ - (1-y)\ln\left(1-\hat{y}\right) + \alpha||W||_2^2 \ \ , \tag{5}$$

where $\hat{y}$ is the observed data, y is the predicted value and W is the weight matrix for the neural network. α is the learning rate parameter used for training neural networks.

The backpropagation algorithm takes the given output and its error and splits it among the connected weights in the network. Through repetition, the neural network adjust these weights so that the error between the output and true observed data is minimized. See [20] for details on the algorithm.

III.2 Overfitting

There may be cases where overfitting issues arise. Overfitting issues occur when a trained neural network does not perform well with a given testing data set. The network memorized the training data set, but it could not generalize it's training to other data. Using a larger sample almost always helps alleviate this problem. Techniques such as L2

regularization, dropout and batch normalization can also be used to correct such issues [21]. Another method to mitigate overfitting is to change the complexity of the model, by using a lower number of weights and smaller weight values [22]. For this study, it was determined that overfitting would not be an issue using the L2 regularization technique. The loss curve for both males and females are given in figures 2-3 show the model performs well given testing data. Due to the structure of both models and the time of training, the range of the lost curves for each model differ.

Support vector machines (SVM) are supervised machine learning models that were designed for classification analysis. SVM was first introduced by [23]. The objective of SVM is to find a hyperplane in a N-dimensional space that classify each data point. The hyperplanes are decision boundaries that classify the data. For binary variables like disease state, the data may land on one of two sides of the plane. The dimensions of the plane depend on the number of parameters for the model. SVM maximizes the distance between the data and the hyperplane, creating output classes [24].

Figure 4 shows a sketch of the hyperplane. Given a data set with points $(\overrightarrow{x_1}, y_1), ...(\overrightarrow{x_n}, y_n)$, where each y_i are either 1 or -1 indicting the class of each $\overrightarrow{x_i}$ and each $\overrightarrow{x_i}$ is a p dimensional vector. The goal is to find the maximum margin hyperplane that separates the group of $\overrightarrow{x_i}$'s for which $y_i = 1$ from those where $y_i = -1$. Finding the maximum margin hyperplane can be written in form of

$$\min_{\overrightarrow{w},b} \frac{1}{2}||\overrightarrow{w}||_2^2,$$

such that $y_i(\overrightarrow{w}^\top \overrightarrow{x_1} + b) \geq 1$, $i = 1, ..., n$. $\overrightarrow{w}$ is the normal vector to the hyperplane and b is the optimal bias term. $y_i(\overrightarrow{w}^\top \overrightarrow{x_1} + b)$ represent the support vectors.

III.3 Random Forest Classification

Random forest classification is another machine learning technique for classifying data. The algorithm was first introduced by [25]. The technique can also be used for

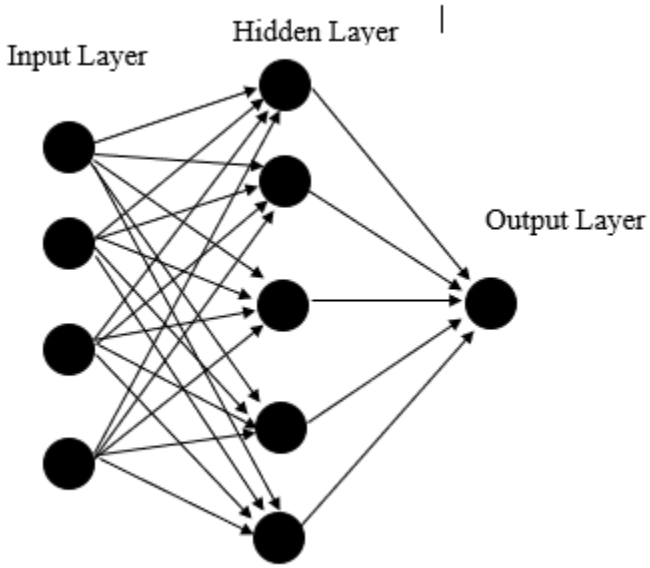

Figure 1: Neural network diagram, which contains the three basic layers: the Input Layer, hidden layer, and the output layer. Between each layer of nodes are connection weights symbolized as lines. The arrows point to the right, indicating that this is a feed-forward neural network.

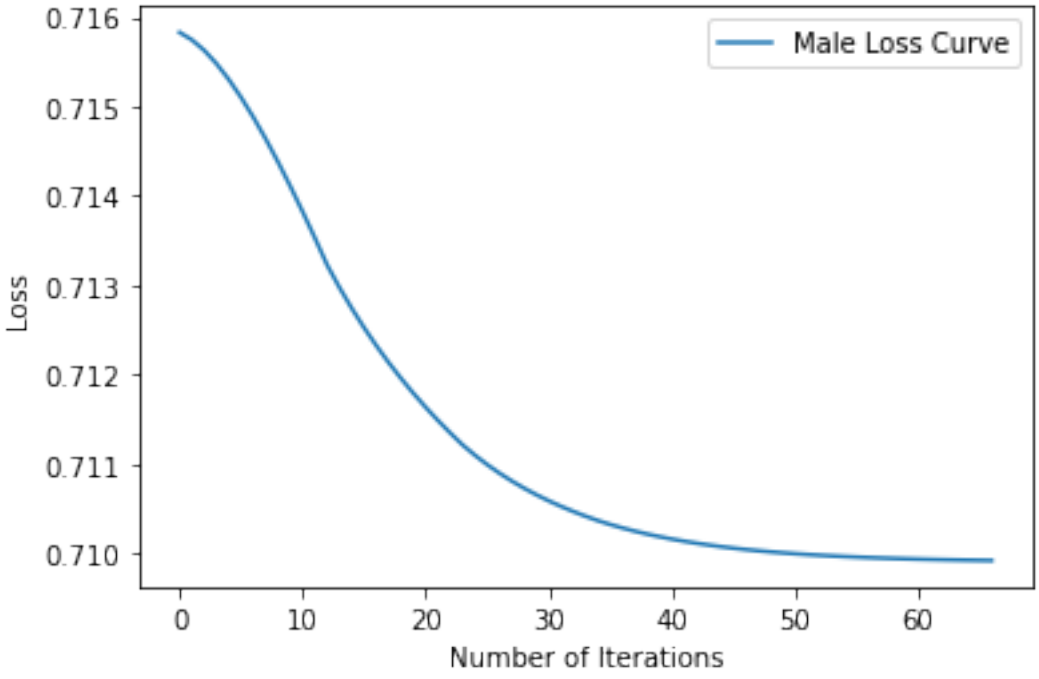

Figure 2: Male loss curve, showing that the male neural network model is improving as it is given validation data sets. As the number of iterations increase, the loss curve should decrease over time.

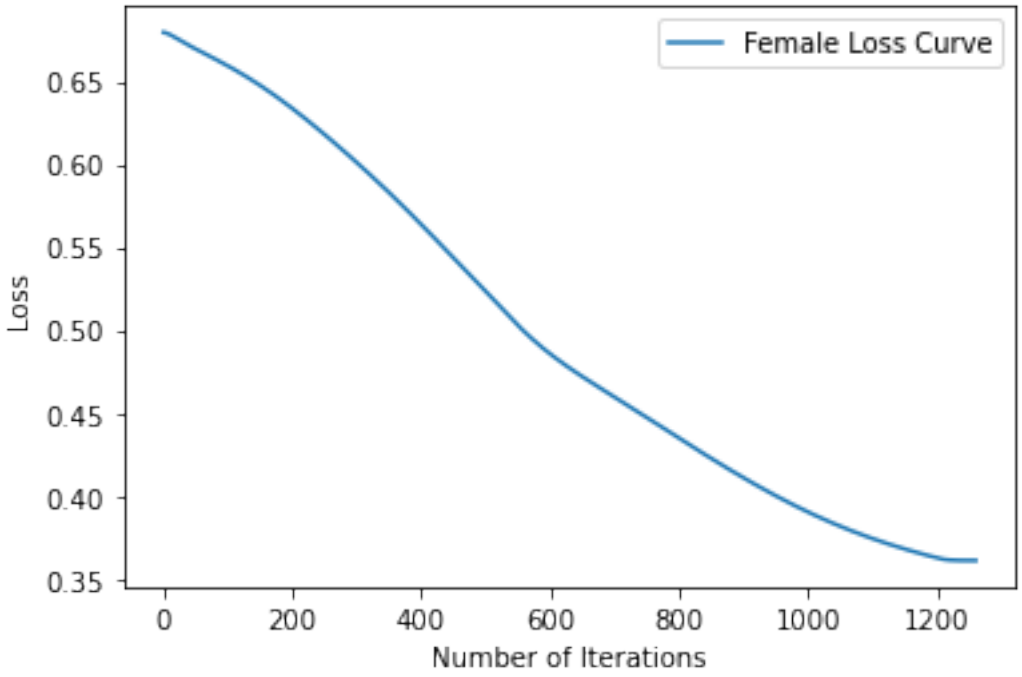

Figure 3: Female loss curve, showing that the male neural network model is improving as it is given validation data sets. As the number of iterations increase, the loss curve should decrease over time.

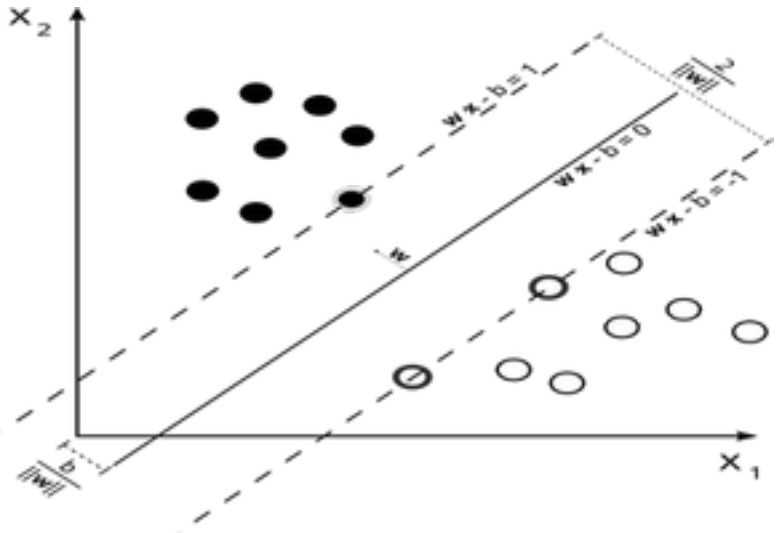

Figure 4: This is a SVM diagram. The hyperplane is the area within the two dashed lines. SVM works to maximize the hyperplane so that the classes are finely separated. The dashed lines are known as the support vectors and and they act as thresholds for class separation. https://commons.wikimedia.org/wiki/File:Svm-max-sep-hyperplane-with-margin.png.

regression and is often regarded as one of the easiest machine learning algorithms to implement. A random forest is comprised of decision trees that work on randomly selected data samples. Each tree gives a prediction and the best solution is selected by a simple majority vote.

Figure 5 below shows a detailed diagram of the random forest algorithm. The algorithm for random forest classification is as follows:

1. For $a = 1$ to A:
   (a) Take sample from of size N from training data.
   (b) Grow random forest tree T_a to the sample, by recursively repeating the following steps for each terminal node of the tree until the minimum node size is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m variables.
      iii. Split node into two daughter nodes.
2. Output the ensemble of trees $(T_a)_1^A$.

To make a prediction for any data point, let $\hat{D}_a(x)$ be the class prediction for the a^{th} random forest tree. Then $\hat{D}_{rf}^A(x) =$ majority vote $\hat{D}_1^A(x)$.

III.4 Performance Metrics

To evaluate each model, performance metrics were implemented and were used to compare the efficacy of each model. These metrics were sensitivity, specificity, mean square error (MSE), receiver operator characteristic (ROC) curve, and area under the curve (AUC). A confusion matrix showing the terms for each of these calculations is given in table 1. Sensitivity, also called the true positive rate (TPR) measures the proportion of actual positives that are correctly identified. It given in equation (6) as

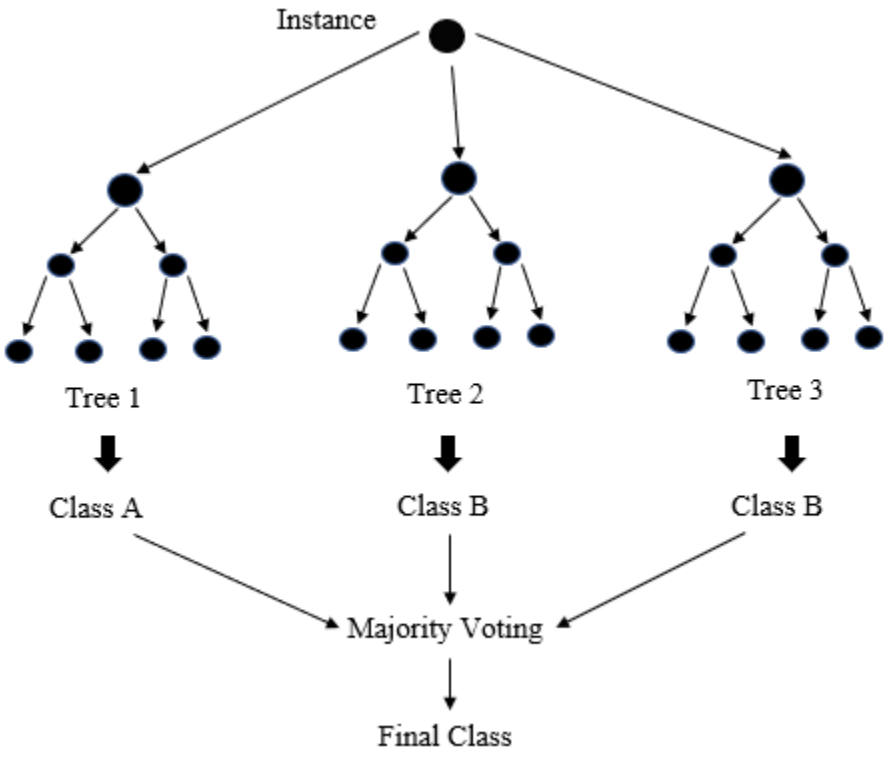

Figure 5: Random forest classification diagram. A random forest classification model using three decision trees. A data point is given. Each decision tree makes its own prediction on what class that data point belongs to. The output from all three trees are tallied, and the class with the most outputs is the new predicted class for the data point.

$$TPR = \frac{TP}{TP + FN}, \tag{6}$$

where TP is the number of actual positive cases and FN is the number of false negative cases. Specificity, also called the true negative rate (TNR) measures the proportion of actual negatives that are correctly identified. It is given by equation (7) as

$$TNR = \frac{TN}{TN + FP}, \tag{7}$$

where TN is the number of actual negative cases and FP is the number of false positive cases. The MSE is the sum of the false positives and false negatives, divided by the total sample size. It is defined in equation (8) as

$$MSE = \frac{FP + FN}{TP + FP + TN + FN}. \tag{9}$$

Table 1: Confusion Matrix. This is a unique contingency table to show the results of a classification model. It also provides a good ideal of how sensitivity, specificity and error are calculated. The actual observed data is matched with the predicted data. For example, the true positive cell represents when the model predicts a that a patient is a case, when they are actually a case, while the false positive cell represents when the model predicted that a patient has a case, when in fact they are a control.

Confusion Matrix		Actual Class	
		Case	Control
Predicted Class	Case	True Positive (TP)	False Positive (FP) (Type I Error)
	Control	False Negative (FN) (Type II Error)	True Negative (TN)

The ROC curve plots true positive rate against the false positive rate at different thresholds. The class prediction for each patience is made based on a random variable X,

which is a score computed for each instance in a data set. Given some threshold T, the patient is classified as a positive (Case) if $X > T$ and negative (control) if $X \leq T$. X follows probability density $f_1(p)$ if the patient is actually a case. X follows probability density $f_0(x)$, otherwise. Therefore the true positive rate and positive rate at different thresholds are defined in equations (9)-(10) as

$$TPR(T) = \int_T^{\infty} f_1(x)dx \tag{9}$$

and

$$FPR(T) = \int_T^{\infty} f_0(x)dx. \tag{10}$$

The ROC is obtained when these two curves are plotted against each other.

The AUC is a measure of how well a model can distinguish between two classes, where $0 \leq AUC \leq 1$. The best case scenario is achieved when $AUC = 1$, where the model a 100% chance of distinguishing between classes. The worst case scenario occurs when $AUC = 0.5$. This indicates that there is a 50% chance that the model can distinguish between classes, rendering it almost ineffective. The AUC is calculated by integrating the ROC curve.

CHAPTER IV: PERFORMANCE COMPARISONS

All machine learning classification analyses were conducted using python's scikit-learn library. The Ridge Regression analysis was conducted using R's GLM package. MSE, sensitivity, specificity, and Area Under the Curve (AUC) are given in table 2. For Ridge Regression classification, the model was run for 1000 iterations to obtain the average output. For males, the average MSE was 0.3807 and the variance of MSE was 0.0038. The average sensitivity was 0.6317, while the average specificity was 0.6214. The AUC for males was 0.6249. For females, the average MSE was 0.3116 and the variance of MSE was 0.0121. The average sensitivity was 0.7245 while the average specificity was 0.69. The AUC for females was 0.7018.

The neural network structure differed for males and females. For the males, three hidden layers were used. Each hidden layer consisted of seven nodes along with the bias +1 node. The model was set to learn for 5000 iterations along with an adaptive learning rate of 0.001. An MSE of 0.2888 was recorded. A sensitivity of 0.6818 was recorded along with a specificity of 0.75. The AUC is given in figure 6. For females, the structure of the neural network consisted of seven hidden layers. Each hidden layer consisted of six nodes along with the +1 bias node. The model was set to learn for 5000 iterations along with an adaptive learning rate of 0.001. The female model recorded an MSE of 0.1333, a sensitivity of 1.0 and a specificity of 0.75. The AUC is given in figure 7.

The SVM model was run for both males and females. For males, the MSE was 0.3777, sensitivity was 0.6190 and specificity was 0.6250. The AUC is given in figure 8. For females an MSE of 0.3333 was record along with a sensitivity of 0.8333 and specificity of 0.5555. the AUC is given in figure 9. The random forest classification was implemented for both males and females. For males, an MSE of 0.31111 was obtained with sensitivity of 0.75 and specificity of 0.6551. For the female model, an MSE of 0.3333 was recorded along with sensitivity of 0.75 and specificity of 0.5714. The AUC for both males and females is given in figures 10-11.

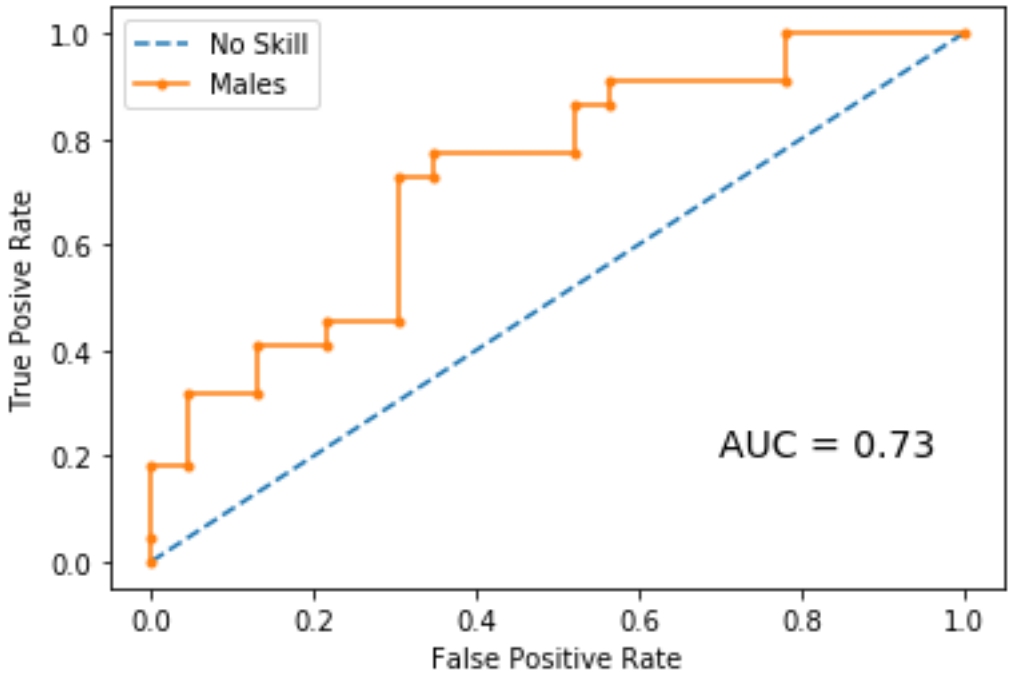

Figure 6: ROC curve and AUC for the male neural network model. AUC is a measure of how well the model is able to distinguish between the disease and control states. The blue dashed line represents an AUC value of 0.5, where the model is not able to distinguish between classes. The orange line is the ROC curve, which is obtained by plotting the sensitivity against the false positive rate at different thresholds.

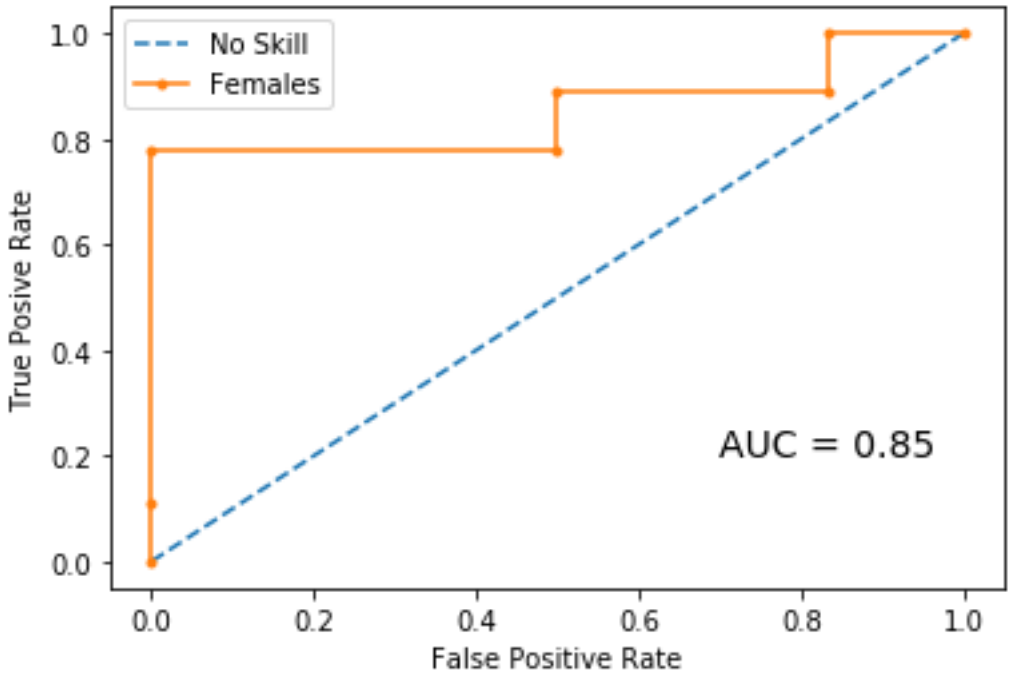

Figure 7: ROC curve and AUC for the female neural network model. AUC is a measure of how well the model is able to distinguish between the disease and control states. The blue dashed line represents an AUC value of 0.5, where the model is not able to distinguish between classes. The orange line is the ROC curve, which is obtained by plotting the sensitivity against the false positive rate at different thresholds.

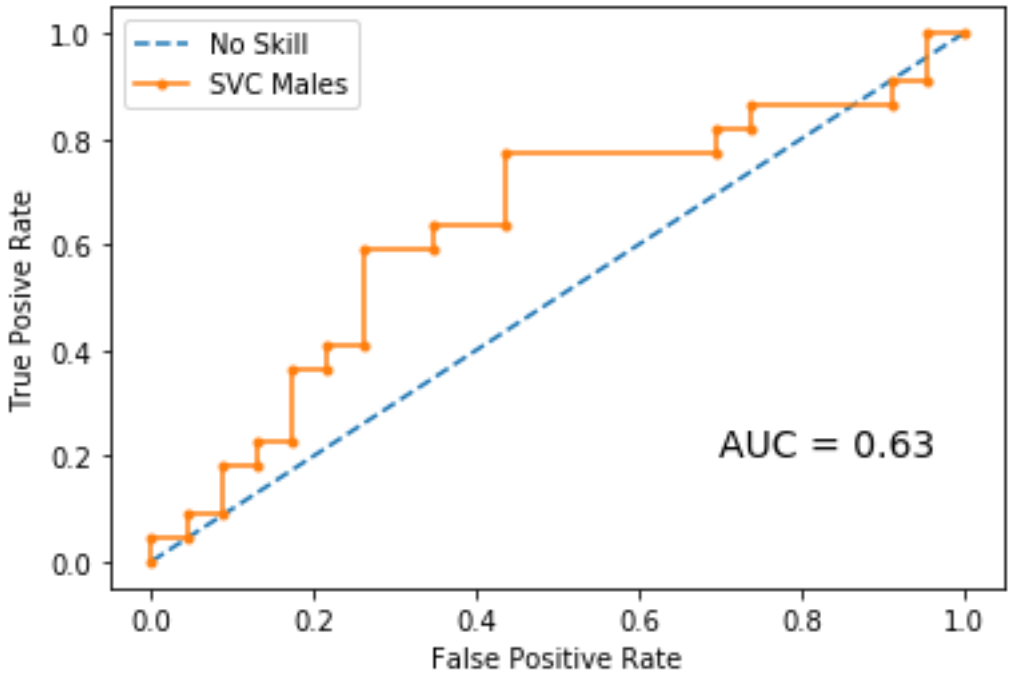

Figure 8: ROC curve and AUC for the male support vector machine model AUC is a measure of how well the model is able to distinguish between the disease and control states. The blue dashed line represents an AUC value of 0.5, where the model is not able to distinguish between classes. The orange line is the ROC curve, which is obtained by plotting the sensitivity against the false positive rate at different thresholds.

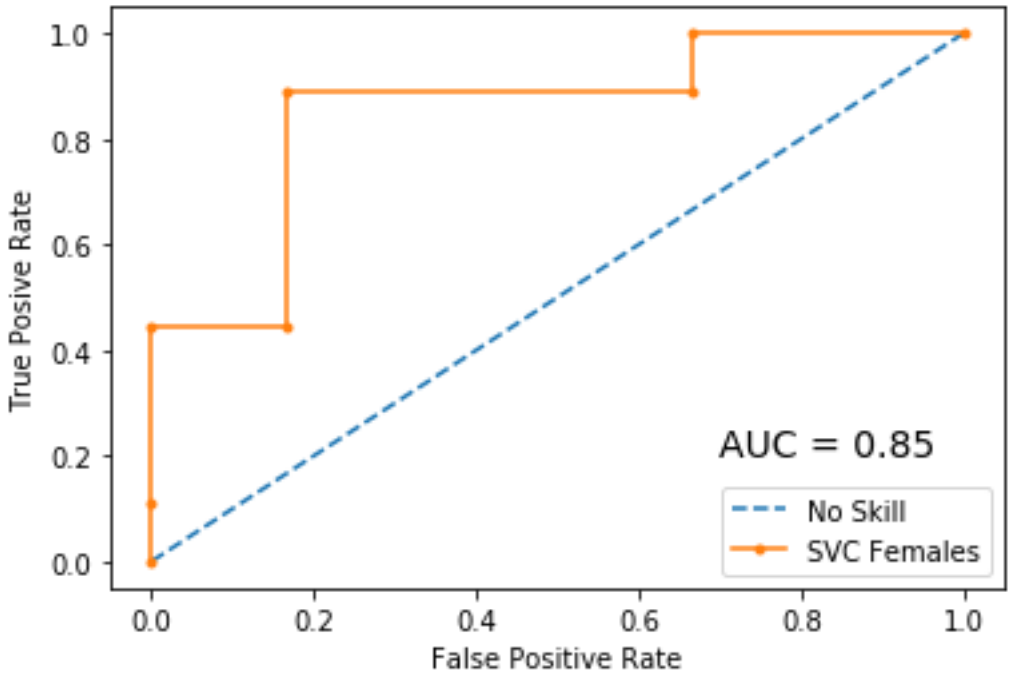

Figure 9: ROC curve and AUC for the female support vector machine model. AUC is a measure of how well the model is able to distinguish between the disease and control states. The blue dashed line represents an AUC value of 0.5, where the model is not able to distinguish between classes. The orange line is the ROC curve, which is obtained by plotting the sensitivity against the false positive rate at different thresholds.

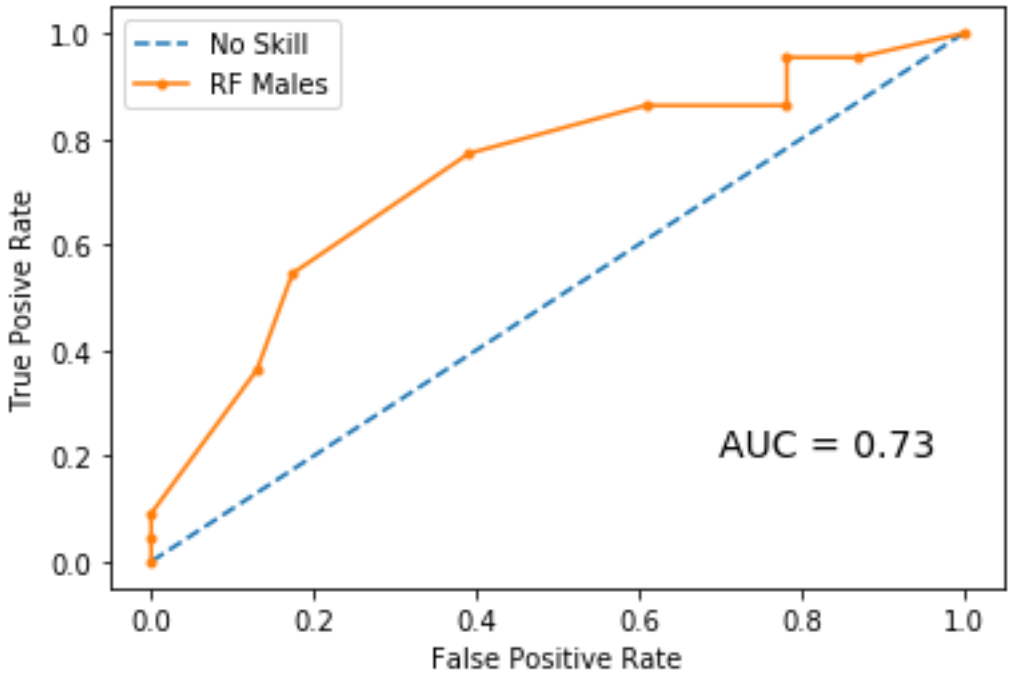

Figure 10: ROC curve and AUC for the male random forest model. AUC is a measure of how well the model is able to distinguish between the disease and control states. The blue dashed line represents an AUC value of 0.5, where the model is not able to distinguish between classes. The orange line is the ROC curve, which is obtained by plotting the sensitivity against the false positive rate at different thresholds.

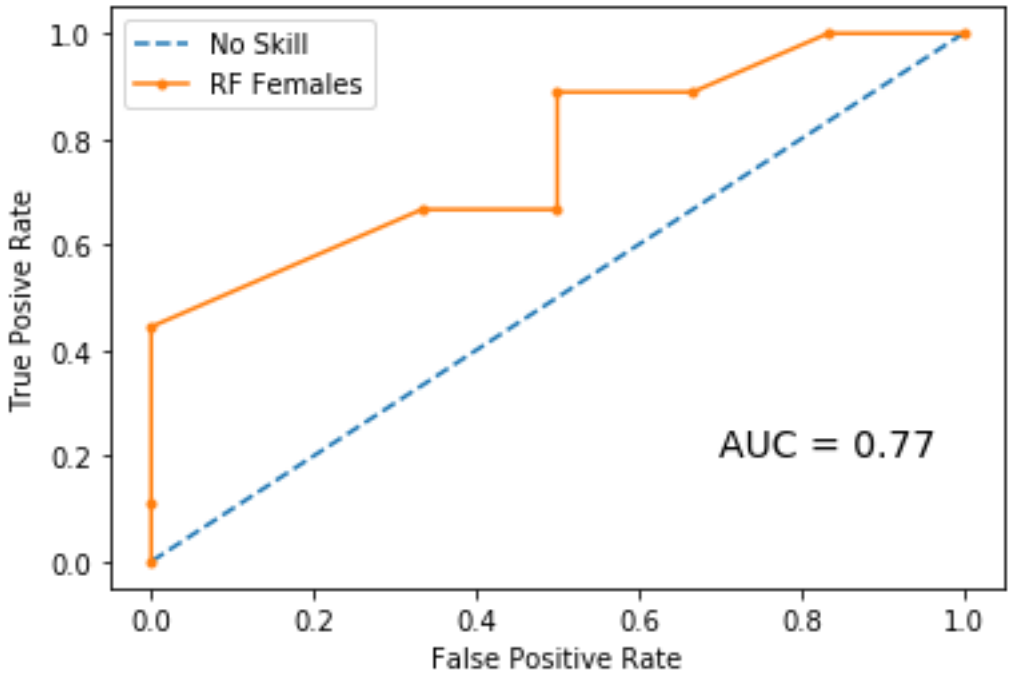

Figure 11: ROC curve and AUC for the female random forest model. AUC is a measure of how well the model is able to distinguish between the disease and control states. The blue dashed line represents an AUC value of 0.5, where the model is not able to distinguish between classes. The orange line is the ROC curve, which is obtained by plotting the sensitivity against the false positive rate at different thesholds.

Table 2: The performance metrics for each model. The metrics are given for both males and females.

Performance-Male				
Method	MSE	Sensitivity	Specificity	AUC
Logistic Regression	0.3807	0.6317	0.6214	0.6249
Neural Network	0.2888	0.6818	0.75	0.73
SVM	0.3777	0.6190	0.6250	0.63
Random Forest	0.3111	0.75	0.6551	0.73
Performance-Female				
Method	MSE	Sensitivity	Specificity	AUC
Logistic Regression	0.3116	0.7245	0.69	0.7018
Neural Network	0.1333	1.0	0.75	0.85
SVM	0.3333	0.8333	0.5555	0.85
Random Forest	0.3333	0.75	0.5714	0.77

CHAPTER V: DISCUSSION

V.1 Neural Networks vs. Ridge Regression

The neural network model was developed for both male and females. After, performance was compared to the results of the previously developed Ridge Regression model. The MSE was given as well as the sensitivity, specificity and AUC. The neural network had the best performance for both males and females. The female neural network model had the lower MSE as well the highest sensitivity, specificity and AUC, compared to its Ridge Regression counterpart. The MSE for the female neural network was 58% lower while the AUC was 21% higher. The male neural network had a lower MSE, higher sensitivity and specificity compared to its Ridge Regression counterpart. The female Ridge Regression model did have a higher sensitivity than the male neural network model, while the male neural network model had a higher specificity and AUC. The MSE for the male neural network was 23% lower and the AUC was 18% higher.

V.2 Neural Networks vs. SVM

Performance of the neural network models were then compared to the SVM models. The MSE, sensitivity, specificity and AUC were compared. The male neural network model outperformed its SVM counterpart. It had a lower MSE, higher sensitivity and specificity as well as a higher AUC. The MSE for the male neural network was 23% lower and the AUC was 15% higher. The female neural network model did outperform the female SVM model, but not in every aspect. The female neural network model had a lower error and a higher sensitivity and specificity. However, both model had roughly the same AUC. The MSE for the female neural network was 60% lower than its SVM counterpart. This result was surprising given that the lower sensitivity and specificity. This could have been the result of the female SVM model reaching a higher sensitivity with lower false positive rate at certain classification thresholds.

V.3 Neural Networks vs. Random Forest

The neural network model performance was then compared to the random forest model performance. Analogous to the comparison with SVM, all metrics are compared. The neural network for males outperformed its random forest counterpart in some areas. It had a lower MSE and higher specificity. However, the male random forest model had a higher sensitivity. The MSE for the male neural neural network was 6% lower while the sensitivity was lower by about 9%. Both models had the roughly the same AUC of 0.73. The female neural network model outperformed its random forest counterpart in every given metric. The MSE for the female neural network was 60% higher while the AUC was 10% higher.

V.4 Ridge Regression vs. SVM

The Ridge Regression models for both males and females were compared to their SVM counterparts. All given metrics are compared. The male Ridge Regression model and the male SVM model were roughly equal in all metrics. The female Ridge Regression model had a lower MSE and higher specificity than its SVM counterpart. However, the SVM model had a higher sensitivity.

V.5 Ridge Regression vs. Random Forest

The Ridge Regression models for both male and females were compared to the male and female random forest models. The male random forest model outperformed the male Ridge Regression model in all metrics. Its MSE, sensitivity and specificity all excelled the Ridge Regression model. The MSE for the male random forest model was 18% lower, while the AUC was 17% higher. The female Ridge Regression model had a lower MSE and higher specificity Than than its random forest counterpart. However, the male random forest model had a lower sensitivity. The MSE for the female random forest model was 6% higher while the area under the curve was 10% higher.

V.6 SVM vs. Random Forest

The SVM models were compared to their random forest counterparts. The male random forest model outperformed the male SVM model in all given metrics. The MSE of the male random forest model was 22% lower while the AUC was 15% higher. The male SVM model had the worst performance of all four male models. The female SVM model outperformed the female random forest model. The MSE were roughly even, while the SVM model had a higher sensitivity and AUC. However, the random forest model had a higher specificity. The AUC of the female SVM model was 10% higher, while the specificity was lower by about 1%.

V.7 General Comparisons

For the males, the neural network model performed in more metrics than it is three counterparts. It had the lowest MSE, higher specificity and its AUC was the highest. The male random forest model had the highest male sensitivity among the four models. For the females, the neural network model performed best in all metrics. It had the lowest MSE, the highest sensitivity and specificity and AUC. Only the SVM model had an AUC roughly equal to it.

The male and female neural network structures differed quite significantly. The female structure contained more hidden layers, while the male structure contained more nodes per hidden layer. The male structure only contained three hidden layers because it was determined that adding extra layers onto the network did not improve its performance. The female model consisted of seven hidden layers, twice as much as the male model. Similar to the male model, it was determined that using fewer or more hidden layers would not improve its performance. For both males and females, each hidden layer contained the same number of nodes as their respective input layers.

The female models performed better than the male models. With exception of the random forest model, the female models had the lower MSE. AUC was higher for all female

models opposed to their male counterparts. This means that the models showed a higher probability of classifying CAD in female patients, compared to male patients. This result is surprising considering the ratio of data for males to female. This ratio is almost 3:1. Models tend to be more accurate, if given as much data as possible. It could also be the case that the predictors for females had a closer association with the disease state, opposed to the males. These relationships require further investigation.

Given the nature of the data collected, it was noted in [10], that the original model was developed to classify Caucasian male patients. This is because majority of patients in the original study were male and Caucasian, while the second largest group were Caucasian females. Therefore, the same assumption must hold for this study. If the neural network model is to ever be used to as a tool to classify CAD, it should primarily be focused on the two before mentioned groups. Using this model to classify CAD in African Americans, Latinos, etc. may not yield proper results because these groups may have different genes or alleles associated with CAD.

CHAPTER VI: CONCLUSION

Proper classification of Coronary Artery Disease is essential for treatment. This study compared the efficacy of three machine learning techniques to Ridge Regression, while given the data of microarray gene expression profiles for male and female patients. Several performance metrics were used to compare the efficiencies of each model. It was shown that artificial neural networks had the best performance, even against other well-known machine learning techniques such as support vector machines. The neural networks outperformed their machine learning counterparts in the majority of metrics mentioned in this study.

For future work, the investigation into classification of CAD in other ethnic groups will be considered. Because of the nature of the data used in this study, it cannot be assumed that these models will do as well in predicting CAD in non-caucasian patients. This will provide opportunities to gain knowledge on the genetic nuances in differenent ethnicies and how such differences effect the development of CAD. Such future studies could lead to advances in Precision Medicine which could lead to the development of more efficient medicines. This could allow ethnicities and/or groups with a specific set of genes to have access to treatments designed for them.

REFERENCES

[1] Prevention CfDCa. Coronary Artery Disease (CAD) Atlanta, Georgia, USA: Centers for Disease Control and Prevention; 2019

[2] Jakl M, Stasek J, Kala P, Rokyta R, Kanovsky J, Ondrus T, et al. Acute myocardial infarction complicated by shock: Outcome analysis based on initial electrocardiogram. Scandinavian cardiovascular journal : SCJ. 2013;48.

[3] Yao S-S, Wever-Pinzon O, Zhang X, Bangalore S, Chaudhry FA. Prognostic value of stress echocardiogram in patients with angiographically significant coronary artery disease. The American journal of cardiology. 2012;109(2):153-8.

[4] Axelsson A, Ruwald MH, Dalsgaard M, Rossing K, Steffensen R, Iversen K. Serial measurements of high-sensitivity cardiac troponin T after exercise stress test in stable coronary artery disease. Biomarkers. 2013;18(4):304-9.

[5] Makaryus AN, Jauhar R, Tortez LM, Pekmezaris R. Comparison of the Diameters of the Major Epicardial Coronary Arteries by Angiogram in Asian-Indians Versus European Americans $<$ 40 Years of Age Undergoing Percutaneous Coronary Artery Intervention. The American Journal of Cardiology. 2017;120(6):924-6.

[6] Fall T, Gustafsson S, Orho-Melander M, Ingelsson E. Genome-wide association study of coronary artery disease among individuals with diabetes: the UK Biobank. Diabetologia. 2018;61(10):2174-9.

[7] Webb TR, Erdmann J, Stirrups KE, Stitziel NO, Masca NGD, Jansen H, et al. Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease. Journal of the American College of Cardiology. 2017;69(7):823-36.

[8] Rotger M, Glass TR, Junier T, Lundgren J, Neaton JD, Poloni ES, et al. Contribution of genetic background, traditional risk factors, and HIV-related factors to coronary artery disease events in HIV-positive persons. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2013;57(1):112-21.

[9] Pendharkar P. A comparison of gradient ascent, gradient descent and genetic-algorithm-based artificial neural networks for the binary classification problem. Expert Systems. 2007;24:65-86.

[10] Kleene SC. Representation of Events in Nerve Nets and Finite Automata. Princeton, New Jersey, USA: Princeton University Press; 1956. p. 3-42.

[11] Almeida PP, Cardoso CP, de Freitas LM. PDAC-ANN: an artificial neural network to predict Pancreatic Ductal Adenocarcinoma based on gene expression. bioRxiv. 2019:698209.

[12] Wang Q, Wei J, Chen Z, Zhang T, Zhong J, Zhong B, et al. Establishment of multiple diagnosis models for colorectal cancer with artificial neural networks. Oncology letters. 2019;17(3):3314-22.

[13] Dwivedi AK. Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications. 2018;29(12):1545-54.

[14] Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. BMC medical genomics. 2011;4:26-.

[15] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

16. Brown PJ, Brown PDSCMPJ. Measurement, Regression, and Calibration: Clarendon Press; 1993.

[16] Brown PJ, Brown PDSCMPJ. Measurement, Regression, and Calibration: Clarendon Press; 1993.

[17] Haykin S. Neural Networks A Comprehensive Foundation. 2nd ed. Patparganji, Delhi, India: Pearson Education, Inc.; 1999.

[18] Günther F, Fritsch S. neuralnet: Training of neural networks. The R journal. 2010;2(1):30-8.

[19] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6.

[20] Hecht-Nielsen R. III.3 - Theory of the Backpropagation Neural Network**Based on "nonindent" by Robert Hecht-Nielsen, which appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593–611, June 1989. © 1989 IEEE. In: Wechsler H, editor. Neural Networks for Perception: Academic Press; 1992. p. 65-93.

[21] Kim Y, Kim HG, Li Z, Choi HJ. Avoiding Overfitting in Deep Neural Networks for Clinical Opinions Generation from General Blood Test Results. Stud Health Technol Inform. 2017;245:1274.

[22] Bishop CM. Neural Networks for Pattern Recognition. 1 ed. USA: Oxford University Press; 1995.

[23] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97.

[24] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory; Pittsburgh, Pennsylvania, USA: Association for Computing Machinery; 1992. p. 144–52.

[25] Ho TK. Random Decision Forest. Montreal, Quebec, Canada: Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995. p. 278-82.